9e

Statistics for the Behavioral Sciences

Frederick J Gravetter
Larry B. Wallnau

Statistics for the Behavioral Sciences

This page intentionally left blank

Statistics for the Behavioral Sciences

Ninth Edition

Frederick J Gravetter
The College at Brockport, State University of New York

Larry B. Wallnau
The College at Brockport, State University of New York

WADSWORTH
CENGAGE Learning

Australia · Brazil · Japan · Korea · Mexico · Singapore · Spain · United Kingdom · United States

WADSWORTH
CENGAGE Learning

Statistics for the Behavioral Sciences,
Ninth Edition
Frederick J Gravetter and Larry B. Wallnau

Publisher: Jon-David Hague
Psychology Editor: Tim Matray
Developmental Editor: Tangelique Williams
Freelance Developmental Editor: Liana Sarkisian
Assistant Editor: Kelly Miller
Editorial Assistant: Lauren K. Moody
Media Editor: Mary Noel
Marketing Program Manager: Sean Foy
Marketing Communications Manager: Laura Localio
Content Project Manager: Charlene M. Carpentier
Design Director: Rob Hugel
Art Director: Pam Galbreath
Manufacturing Planner: Judy Inouye
Rights Acquisitions Specialist: Tom McDonough
Production Service: Graphic World Inc.
Text Designer: Cheryl Carrington
Text Researcher: Karyn Morrison
Copy Editor: Graphic World Inc.
Illustrator: Graphic World Inc.
Cover Designer: Cheryl Carrington
Cover Image: Edouard Benedictus, Dover Publications
Compositor: Graphic World Inc.

Library of Congress Control Number: 2011934937

Student Edition:
ISBN-13: 978-1-111-83099-1
ISBN-10: 1-111-83099-1

Loose-leaf Edition:
ISBN-13: 978-1-111-83576-7
ISBN-10: 1-111-83576-4

Wadsworth
20 Davis Drive
Belmont, CA 94002-3098
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **www.cengage.com/global.**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **www.cengage.com.**

Purchase any of our products at your local college store or at our preferred online store **www.CengageBrain.com.**

Printed in the United States of America
1 2 3 4 5 6 7 15 14 13 12 11

Contents in Brief

PART V Correlations and Nonparametric Tests

Contents

This page intentionally left blank

Preface

Many students in the behavioral sciences view the required statistics course as an intimidating obstacle that has been placed in the middle of an otherwise interesting curriculum. They want to learn about human behavior—not about math and science. As a result, they see the statistics course as irrelevant to their education and career goals. However, as long as the behavioral sciences are founded in science, knowledge of statistics will be necessary. Statistical procedures provide researchers with objective and systematic methods for describing and interpreting their research results. Scientific research is the system that we use to gather information, and statistics are the tools that we use to distill the information into sensible and justified conclusions. The goal of this book is not only to teach the methods of statistics, but also to convey the basic principles of objectivity and logic that are essential for science and valuable in everyday life.

Those familiar with previous editions of *Statistics for the Behavioral Sciences* will notice that some changes have been made. These changes are summarized in the section titled "To the Instructor." In revising this text, our students have been foremost in our minds. Over the years, they have provided honest and useful feedback. Their hard work and perseverance has made our writing and teaching most rewarding. We sincerely thank them. Students who are using this edition should please read the section of the preface titled "To the Student."

ANCILLARIES

Ancillaries for this edition include the following:

- *Study Guide:* Contains chapter summaries, learning objectives, new terms and concepts with definitions, new formulas, step-by-step procedures for problem solving, study hints and cautions, self-tests, and review. The Study Guide contains answers to the self-test questions.
- *Instructor's Manual with Test Bank:* Contains a detailed table of contents, chapter outlines, annotated learning objectives, lecture suggestions, test items, and solutions to all end-of-chapter problems in the text. Test items are also available as a Microsoft Word® download or for ExamView® computerized test bank software with multiple-choice, true/false, and short-answer questions. An answer key is provided for all questions, and each question is cross-referenced to a page in the textbook.
- *PowerLecture with ExamView:* The fastest, easiest way to build powerful, customized, media-rich lectures, PowerLecture provides a collection of book-specific Microsoft PowerPoint® lecture and class tools to enhance the educational experience. ExamView allows you to create, deliver, and customize tests and study guides (both print and online) in minutes.
- *WebTutor™ on Blackboard and WebCT™:* Jumpstart your course with customizable, text-specific content for use within your course-management system. Whether you want to Web-enable your class or put an entire course online, WebTutor delivers. WebTutor offers a wide array of resources including glossary, flashcards, quizzing, and more.

- *Psychology CourseMate®:* Psychology CourseMate, with Engagement Tracker, a first-of-its-kind tool that monitors student engagement in the course, includes:
  - An interactive eBook
  - Interactive teaching and learning tools including:
    - Quizzes
    - Glossary
    - Flashcards
    - and more

ACKNOWLEDGMENTS

It takes a lot of good, hard-working people to produce a book. Our friends at Wadsworth/Cengage have made enormous contributions to this textbook. We thank: Linda Schreiber-Ganster, Publisher/Executive Editor; Timothy Matray, Acquisitions Editor; Tangelique Williams, Managing Developmental Editor; Kelly Miller, Assistant Editor; Lauren K. Moody, Editorial Assistant/Associate; Charlene M. Carpentier, Content Project Manager; Mary Noel, Media Editor; and Pam Galbreath, Art Director. Special thanks go to Liana Sarkisian, our Development Editor, and to Mike Ederer who led us through production at Graphic World.

Reviewers play a very important role in the development of a manuscript. Accordingly, we offer our appreciation to the following colleagues for their assistance with the ninth edition: Patricia Case, University of Toledo; Kevin David, Northeastern State University; Adia Garrett, University of Maryland, Baltimore County; Carrie E. Hall, Miami University; Deletha Hardin, University of Tampa; Angela Heads, Prairie View A&M University; Roberto Heredia, Texas A&M International University; Alisha Janowski, University of Central Florida; Matthew Mulvaney, The College at Brockport (SUNY); Nicholas Von Glahn, California State Polytechnic University, Pomona; and Ronald Yockey, Fresno State University.

TO THE INSTRUCTOR

Those of you familiar with the previous edition of *Statistics for the Behavioral Sciences* will notice a number of changes in the ninth edition. Throughout the book, research examples have been updated, real-world examples have been added, and the end-of-chapter problems have been extensively revised. The book has been separated into five sections to emphasize the similarities among groups of statistical methods. Each section contains two to four chapters, and begins with an introduction and concludes with a review, including review exercises.

Major revisions for this edition include:

1. The former Chapter 12 on Estimation has been eliminated. In its place, sections on confidence intervals have been added to the three *t*-statistic chapters.
2. The former Chapter 20 covering hypothesis tests for ordinal data has been converted into an appendix.
3. A new final chapter discusses the process of selecting the correct statistics to be used with different categories of data and replaces the Statistics Organizer, which appeared as an appendix in earlier editions.

Other examples of specific and noteworthy revisions include:

Chapter 1 A separate section explains how statistical methods can be classified using the same categories that are used to group data structures and research methods. A new heading clarifies the concept that different scales of measurement require different statistical methods.

Chapter 2 The discussion of histograms has been modified to differentiate discrete and continuous variables. The section on stem-and-leaf displays has been heavily edited to simplify.

Chapter 3 A modified definition of the median acknowledges that this value is not algebraically defined, and that determining the median, especially for discrete variables, can be somewhat subjective. A new Box demonstrates that precisely locating the median for a continuous variable is equivalent to using interpolation to find the 50th percentile (as was demonstrated in Chapter 2).

Chapter 4 Alternative definitions of the range have been added, and discussion of the interquartile range has been deleted. Greater emphasis has been placed on conceptual definitions of standard deviation and the sum of squared deviations *(SS)*. The section on variance and inferential statistics has been simplified and the section comparing measures of variability has been deleted.

Chapter 5 Relatively minor editing for clarity.

Chapter 6 The concepts of *random sample* and *independent random sample* are clarified with separate definitions. A new figure helps demonstrate the process of using the unit normal table to find proportions for negative *z*-scores. The section on the binomial distribution has been shortened and simplified.

Chapter 7 Relatively minor editing for clarity.

Chapter 8 The chapter has been shortened by substantial editing that eliminated several pages of unnecessary text, particularly in the sections on errors (Type I and II) and power. The distinction between reporting one-tailed and two-tailed tests was clarified.

Chapter 9 The section describing how sample size and sample variance influence the outcome of a hypothesis test has been moved so that it appears immediately after the hypothesis test example. A new section introduces confidence intervals in the context of describing effect size, describes how confidence intervals are reported in the literature, and discusses factors affecting the width of a confidence interval.

Chapter 10 An expanded section discusses how sample variance and sample size influence the outcome of an independent-measures hypothesis test and measures of effect size. A new section introduces confidence intervals as an alternative for describing effect size. The relationship between a confidence interval and a hypothesis test is also discussed. We also note that the Mann-Whitney test (presented in Appendix E) can be used as an alternative to the independent-measures *t* test if high variance causes problems or if an assumption is violated.

Chapter 11 The description of repeated-measures and matched-subjects designs was clarified and we increased emphasis on the concept that all calculations for the related-samples test are done with the difference scores. A new section introduces confidence intervals as an alternative for describing effect size and discusses the relationship between a confidence interval and a hypothesis test. We also note that the Wilcoxon test (presented in Appendix E) can be used as an alternative to the repeated-measures *t* test if high variance causes problems or if an assumption is violated.

The former Chapter 12 has been deleted. The content from this chapter covering confidence intervals has been added to Chapters 9, 10, and 11.

Chapter 12 (former Chapter 13, introducing ANOVA) The discussion of testwise alpha levels versus experimentwise alpha levels was moved from a Box into the text, and definitions of the two terms were added. To emphasize the concepts of ANOVA rather than the formulas, $SS_{\text{between treatments}}$ is routinely found by subtraction instead of being computed directly. Two alternative equations for $SS_{\text{between treatments}}$ were moved from the text into a Box. We also note that the Kruskal-Wallis test (presented in Appendix E) can be used as an alternative to the independent-measures ANOVA if high variance causes problems or if an assumption is violated.

Chapter 13 (former Chapter 14, introducing repeated-measures ANOVA) A new section demonstrates the relationship between ANOVA and the *t* test when a repeated-measures study is comparing only two treatments. Major editing shortened the chapter and simplified the presentation. We also note that the Friedman test (presented in Appendix E) can be used as an alternative to the repeated-measures ANOVA if high variance causes problems or if an assumption is violated.

Chapter 14 (formerly Chapter 15, introducing two-factor ANOVA) A new section demonstrates how using a participant characteristic as a second factor can reduce the variance caused by individual differences. Major editing shortened the chapter and simplified the presentation.

Chapter 15 (formerly Chapter 16, introducing correlations) The introduction to partial correlations was simplified and moved from the regression chapter into the section discussing the Pearson correlation.

Chapter 16 (formerly Chapter 17, introducing regression) Major editing shortened and simplified the section on multiple regression. A printout showing the results of multiple regression from SPSS was added as a figure to illustrate the different elements of the process.

Chapter 17 (formerly Chapter 18, introducing chi-square tests) Relatively minor editing to shorten and clarify.

Chapter 18 (formerly Chapter 19, introducing the binomial test) Relatively minor editing for clarity.

Chapter 19 A completely new chapter outlining the process of selecting the correct statistical procedures to use with different sets of data.

The former Chapter 20 covering hypothesis tests for ordinal data has been substantially shortened and converted into an Appendix.

Matching the Text to Your Syllabus The book chapters are organized in the sequence that we use for our own statistics courses. However, different instructors may prefer different organizations and probably will choose to omit or deemphasize specific topics. We have tried to make separate chapters, and even sections of chapters, completely self-contained, so they can be deleted or reorganized to fit the syllabus for nearly any instructor. Some common examples are as follows:

- It is common for instructors to choose between emphasizing analysis of variance (Chapters 12, 13, and 14) or emphasizing correlation/regression (Chapters 15 and 16). It is rare for a one-semester course to complete coverage of both topics.

- Although we choose to complete all the hypothesis tests for means and mean differences before introducing correlation (Chapter 15), many instructors prefer to place correlation much earlier in the sequence of course topics. To accommodate this, sections 15.1, 15.2, and 15.3 present the calculation and interpretation of the Pearson correlation and can be introduced immediately following Chapter 4 (variability). Other sections of Chapter 15 refer to hypothesis testing and should be delayed until the process of hypothesis testing (Chapter 8) has been introduced.
- It is also possible for instructors to present the chi-square tests (Chapter 17) much earlier in the sequence of course topics. Chapter 17, which presents hypothesis tests for proportions, can be presented immediately after Chapter 8, which introduces the process of hypothesis testing. If this is done, we also recommend that the Pearson correlation (Sections 15.1, 15.2, and 15.3) be presented early to provide a foundation for the chi-square test for independence.

TO THE STUDENT

A primary goal of this book is to make the task of learning statistics as easy and painless as possible. Among other things, you will notice that the book provides you with a number of opportunities to practice the techniques you will be learning in the form of Learning Checks, Examples, Demonstrations, and end-of-chapter problems. We encourage you to take advantage of these opportunities. Read the text rather than just memorizing the formulas. We have taken care to present each statistical procedure in a conceptual context that explains why the procedure was developed and when it should be used. If you read this material and gain an understanding of the basic concepts underlying a statistical formula, you will find that learning the formula and how to use it will be much easier. In the "Study Hints," that follow, we provide advice that we give our own students. Ask your instructor for advice as well; we are sure that other instructors will have ideas of their own.

Over the years, the students in our classes and other students using our book have given us valuable feedback. If you have any suggestions or comments about this book, you can write to either Professor Emeritus Frederick Gravetter or Professor Emeritus Larry Wallnau at the Department of Psychology, SUNY College at Brockport, 350 New Campus Drive, Brockport, New York 14420. You can also contact Professor Emeritus Gravetter directly at fgravett@brockport.edu.

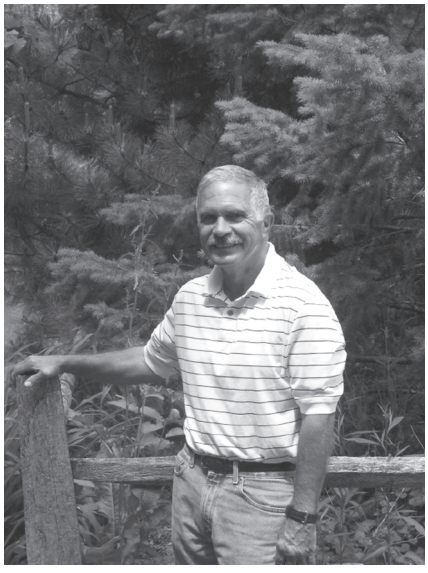
Study Hints You may find some of these tips helpful, as our own students have reported.

- The key to success in a statistics course is to keep up with the material. Each new topic builds on previous topics. If you have learned the previous material, then the new topic is just one small step forward. Without the proper background, however, the new topic can be a complete mystery. If you find that you are falling behind, get help immediately.
- You will learn (and remember) much more if you study for short periods several times per week rather than try to condense all of your studying into one long session. For example, it is far more effective to study half an hour every night than to have a single $3\frac{1}{2}$-hour study session once a week. We cannot even work on *writing* this book without frequent rest breaks.
- Do some work before class. Keep a little ahead of the instructor by reading the appropriate sections before they are presented in class. Although you may not fully understand what you read, you will have a general idea of the topic, which will make the lecture easier to follow. Also, you can identify material that is particularly confusing and then be sure the topic is clarified in class.

- Pay attention and think during class. Although this advice seems obvious, often it is not practiced. Many students spend so much time trying to write down every example presented or every word spoken by the instructor that they do not actually understand and process what is being said. Check with your instructor—there may not be a need to copy every example presented in class, especially if there are many examples like it in the text. Sometimes, we tell our students to put their pens and pencils down for a moment and just listen.
- Test yourself regularly. Do not wait until the end of the chapter or the end of the week to check your knowledge. After each lecture, work some of the end-of-chapter problems and do the Learning Checks. Review the Demonstration Problems, and be sure you can define the Key Terms. If you are having trouble, get your questions answered *immediately*—reread the section, go to your instructor, or ask questions in class. By doing so, you will be able to move ahead to new material.
- Do not kid yourself! Avoid denial. Many students observe their instructor solve problems in class and think to themselves, "This looks easy, I understand it." Do you really understand it? Can you really do the problem on your own without having to leaf through the pages of a chapter? Although there is nothing wrong with using examples in the text as models for solving problems, you should try working a problem with your book closed to test your level of mastery.
- We realize that many students are embarrassed to ask for help. It is our biggest challenge as instructors. You must find a way to overcome this aversion. Perhaps contacting the instructor directly would be a good starting point, if asking questions in class is too anxiety-provoking. You could be pleasantly surprised to find that your instructor does not yell, scold, or bite! Also, your instructor might know of another student who can offer assistance. Peer tutoring can be very helpful.

Frederick J Gravetter
Larry B. Wallnau

About the Authors



Frederick J Gravetter is Professor Emeritus of Psychology at the State University of New York College at Brockport. While teaching at Brockport, Dr. Gravetter specialized in statistics, experimental design, and cognitive psychology. He received his bachelor's degree in mathematics from M.I.T. and his Ph.D. in psychology from Duke University. In addition to publishing this textbook and several research articles, Dr. Gravetter co-authored *Research Methods for the Behavioral Science*s and *Essentials of Statistics for the Behavioral Sciences*.

Fred

Larry B. Wallnau is Professor Emeritus of Psychology at the State University of New York College at Brockport. While teaching at Brockport, he published numerous research articles, primarily on the behavioral effects of psychotropic drugs. With Dr. Gravetter, he co-authored *Essentials of Statistics for the Behavioral Sciences*. He also has provided editorial consulting for numerous publishers and journals. He routinely gives lectures and demonstrations on service dogs for those with disabilities.

Ed Berns

Ben and Larry

This page intentionally left blank

PART

I

Introduction and Descriptive Statistics

We have divided this book into five sections, each covering a general topic area of statistics. The first section, consisting of Chapters 1 to 4, provides a broad overview of statistical methods and a more focused presentation of those methods that are classified as *descriptive statistics*.

By the time you finish the four chapters in this part, you should have a good understanding of the general goals of statistics and you should be familiar with the basic terminology and notation used in statistics. In addition, you should be familiar with the techniques of descriptive statistics that help researchers organize and summarize the results they obtain from their research. Specifically, you should be able to take a set of scores and present them in a table or in a graph that provides an overall picture of the complete set. Also, you should be able to summarize a set of scores by calculating one or two values (such as the average) that describe the entire set.

At the end of this section there is a brief summary and a set of review problems that should help integrate the elements from the separate chapters.

This page intentionally left blank

CHAPTER

1

Introduction to Statistics

Preview

Before we begin our discussion of statistics, we ask you to read the following paragraph taken from the philosophy of Wrong Shui (Candappa, 2000).

> The Journey to Enlightenment
>
> In Wrong Shui, life is seen as a cosmic journey, a struggle to overcome unseen and unexpected obstacles at the end of which the traveler will find illumination and enlightenment. Replicate this quest in your home by moving light switches away from doors and over to the far side of each room.*

Why did we begin a statistics book with a bit of twisted philosophy? Actually, the paragraph is an excellent (and humorous) counterexample for the purpose of this book. Specifically, our goal is to help you avoid stumbling around in the dark by providing lots of easily available light switches and plenty of illumination as you journey through the world of statistics. To accomplish this, we try to present sufficient background and a clear statement of purpose as we introduce each new statistical procedure. Remember that all statistical procedures were developed to serve a purpose. If you understand why a new procedure is needed, you will find it much easier to learn.

The objectives for this first chapter are to provide an introduction to the topic of statistics and to give you some background for the rest of the book. We discuss the role of statistics within the general field of scientific inquiry, and we introduce some of the vocabulary and notation that are necessary for the statistical methods that follow.

As you read through the following chapters, keep in mind that the general topic of statistics follows a well-organized, logically developed progression that leads from basic concepts and definitions to increasingly sophisticated techniques. Thus, the material presented in the early chapters of this book serves as a foundation for the material that follows. The content of the first nine chapters, for example, provides an essential background and context for the statistical methods presented in Chapter 10. If you turn directly to Chapter 10 without reading the first nine chapters, you will find the material confusing and incomprehensible. However, if you learn and use the background material, you will have a good frame of reference for understanding and incorporating new concepts as they are presented.

*Candappa, R. (2000). *The little book of wrong shui.* Kansas City: Andrews McMeel Publishing. Reprinted by permission.

1.1 STATISTICS, SCIENCE, AND OBSERVATIONS

DEFINITIONS OF STATISTICS

By one definition, *statistics* consist of facts and figures such as average income, crime rate, birth rate, baseball batting averages, and so on. These statistics are usually informative and time saving because they condense large quantities of information into a few simple figures. Later in this chapter we return to the notion of calculating statistics (facts and figures) but, for now, we concentrate on a much broader definition of statistics. Specifically, we use the term statistics to refer to a set of mathematical procedures. In this case, we are using the term *statistics* as a shortened version of *statistical procedures*. For example, you are probably using this book for a statistics course in which you will learn about the statistical techniques that are used for research in the behavioral sciences.

Research in psychology (and other fields) involves gathering information. To determine, for example, whether violence on TV has any effect on children's behavior, you would need to gather information about children's behaviors and the TV programs they watch. When researchers finish the task of gathering information, they typically find themselves with pages and pages of measurements such as IQ scores, personality scores, reaction time scores, and so on. In this book, we present the statistics that

researchers use to analyze and interpret the information that they gather. Specifically, statistics serve two general purposes:

1. Statistics are used to organize and summarize the information so that the researcher can see what happened in the research study and can communicate the results to others.
2. Statistics help the researcher to answer the questions that initiated the research by determining exactly what general conclusions are justified based on the specific results that were obtained.

DEFINITION

The term **statistics** refers to a set of mathematical procedures for organizing, summarizing, and interpreting information.

Statistical procedures help to ensure that the information or observations are presented and interpreted in an accurate and informative way. In somewhat grandiose terms, statistics help researchers bring order out of chaos. In addition, statistics provide researchers with a set of standardized techniques that are recognized and understood throughout the scientific community. Thus, the statistical methods used by one researcher are familiar to other researchers, who can accurately interpret the statistical analyses with a full understanding of how the analysis was done and what the results signify.

1.2 POPULATIONS AND SAMPLES

WHAT ARE THEY?

Research in the behavioral sciences typically begins with a general question about a specific group (or groups) of individuals. For example, a researcher may be interested in the effect of divorce on the self-esteem of preteen children. Or a researcher may want to examine the amount of time spent in the bathroom for men compared to women. In the first example, the researcher is interested in the group of *preteen children*. In the second example, the researcher wants to compare the group of *men* with the group of *women*. In statistical terminology, the entire group that a researcher wishes to study is called a *population*.

DEFINITION

A **population** is the set of all the individuals of interest in a particular study.

As you can well imagine, a population can be quite large—for example, the entire set of women on the planet Earth. A researcher might be more specific, limiting the population for study to women who are registered voters in the United States. Perhaps the investigator would like to study the population consisting of women who are heads of state. Populations can obviously vary in size from extremely large to very small, depending on how the researcher defines the population. The population being studied should always be identified by the researcher. In addition, the population need not consist of people—it could be a population of rats, corporations, parts produced in a factory, or anything else a researcher wants to study. In practice, populations are typically very large, such as the population of college sophomores in the United States or the population of small businesses.

Because populations tend to be very large, it usually is impossible for a researcher to examine every individual in the population of interest. Therefore, researchers typically

select a smaller, more manageable group from the population and limit their studies to the individuals in the selected group. In statistical terms, a set of individuals selected from a population is called a *sample*. A sample is intended to be representative of its population, and a sample should always be identified in terms of the population from which it was selected.

DEFINITION

A **sample** is a set of individuals selected from a population, usually intended to represent the population in a research study.

Just as we saw with populations, samples can vary in size. For example, one study might examine a sample of only 10 students in a graduate program, and another study might use a sample of more than 1,000 registered voters representing the population of a major city.

So far we have talked about a sample being selected from a population. However, this is actually only half of the full relationship between a sample and its population. Specifically, when a researcher finishes examining the sample, the goal is to generalize the results back to the entire population. Remember that the research started with a general question about the population. To answer the question, a researcher studies a sample and then generalizes the results from the sample to the population. The full relationship between a sample and a population is shown in Figure 1.1.

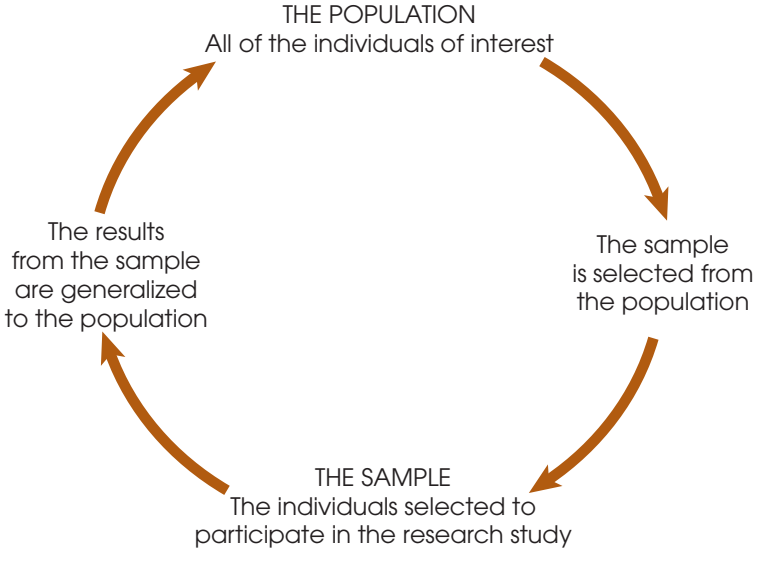
VARIABLES AND DATA

Typically, researchers are interested in specific characteristics of the individuals in the population (or in the sample), or they are interested in outside factors that may influence the individuals. For example, a researcher may be interested in the influence of the weather on people's moods. As the weather changes, do people's moods also change? Something that can change or have different values is called a *variable.*

DEFINITION

A **variable** is a characteristic or condition that changes or has different values for different individuals.

FIGURE 1.1
The relationship between a population and a sample.



Once again, variables can be characteristics that differ from one individual to another, such as height, weight, gender, or personality. Also, variables can be environmental conditions that change, such as temperature, time of day, or the size of the room in which the research is being conducted.

To demonstrate changes in variables, it is necessary to make measurements of the variables being examined. The measurement obtained for each individual is called a *datum* or, more commonly, a *score* or *raw score*. The complete set of scores is called the *data set* or simply the *data*.

DEFINITIONS

Data (plural) are measurements or observations. A **data set** is a collection of measurements or observations. A **datum** (singular) is a single measurement or observation and is commonly called a **score** or **raw score**.

Before we move on, we should make one more point about samples, populations, and data. Earlier, we defined populations and samples in terms of *individuals*. For example, we discussed a population of college sophomores and a sample of preschool children. Be forewarned, however, that we will also refer to populations or samples of *scores*. Because research typically involves measuring each individual to obtain a score, every sample (or population) of individuals produces a corresponding sample (or population) of scores.

PARAMETERS AND STATISTICS

When describing data, it is necessary to distinguish whether the data come from a population or a sample. A characteristic that describes a population—for example, the average score for the population—is called a *parameter*. A characteristic that describes a sample is called a *statistic*. Thus, the average score for a sample is an example of a statistic. Typically, the research process begins with a question about a population parameter. However, the actual data come from a sample and are used to compute sample statistics.

DEFINITIONS

A **parameter** is a value, usually a numerical value, that describes a population. A parameter is usually derived from measurements of the individuals in the population.

A **statistic** is a value, usually a numerical value, that describes a sample. A statistic is usually derived from measurements of the individuals in the sample.

Every population parameter has a corresponding sample statistic, and most research studies involve using statistics from samples as the basis for answering questions about population parameters. As a result, much of this book is concerned with the relationship between sample statistics and the corresponding population parameters. In Chapter 7, for example, we examine the relationship between the mean obtained for a sample and the mean for the population from which the sample was obtained.

DESCRIPTIVE AND INFERENTIAL STATISTICAL METHODS

Although researchers have developed a variety of different statistical procedures to organize and interpret data, these different procedures can be classified into two general categories. The first category, *descriptive statistics*, consists of statistical procedures that are used to simplify and summarize data.

DEFINITION

Descriptive statistics are statistical procedures used to summarize, organize, and simplify data.

Descriptive statistics are techniques that take raw scores and organize or summarize them in a form that is more manageable. Often the scores are organized in a table or a graph so that it is possible to see the entire set of scores. Another common technique is to summarize a set of scores by computing an average. Note that even if the data set has hundreds of scores, the average provides a single descriptive value for the entire set.

The second general category of statistical techniques is called *inferential statistics*. Inferential statistics are methods that use sample data to make general statements about a population.

DEFINITION

Inferential statistics consist of techniques that allow us to study samples and then make generalizations about the populations from which they were selected.

Because populations are typically very large, it usually is not possible to measure everyone in the population. Therefore, a sample is selected to represent the population. By analyzing the results from the sample, we hope to make general statements about the population. Typically, researchers use sample statistics as the basis for drawing conclusions about population parameters.

One problem with using samples, however, is that a sample provides only limited information about the population. Although samples are generally *representative* of their populations, a sample is not expected to give a perfectly accurate picture of the whole population. There usually is some discrepancy between a sample statistic and the corresponding population parameter. This discrepancy is called *sampling error*, and it creates the fundamental problem that inferential statistics must always address (Box 1.1).

DEFINITION

Sampling error is the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter.

The concept of sampling error is illustrated in Figure 1.2. The figure shows a population of 1,000 college students and two samples, each with 5 students, that have been selected from the population. Notice that each sample contains different individuals who have different characteristics. Because the characteristics of each sample depend on the specific people in the sample, statistics vary from one sample to another. For example, the five students in sample 1 have an average age of 19.8 years and the students in sample 2 have an average age of 20.4 years.

BOX 1.1 THE MARGIN OF ERROR BETWEEN STATISTICS AND PARAMETERS

One common example of sampling error is the error associated with a sample proportion. For example, in newspaper articles reporting results from political polls, you frequently find statements such as this:

> Candidate Brown leads the poll with 51% of the vote. Candidate Jones has 42% approval, and the remaining 7% are undecided. This poll was taken from a sample of registered voters and has a margin of error of plus-or-minus 4 percentage points.

The *margin of error* is the sampling error. In this case, the percentages that are reported were obtained from a sample and are being generalized to the whole population. As always, you do not expect the statistics from a sample to be perfect. There is always some margin of error when sample statistics are used to represent population parameters.

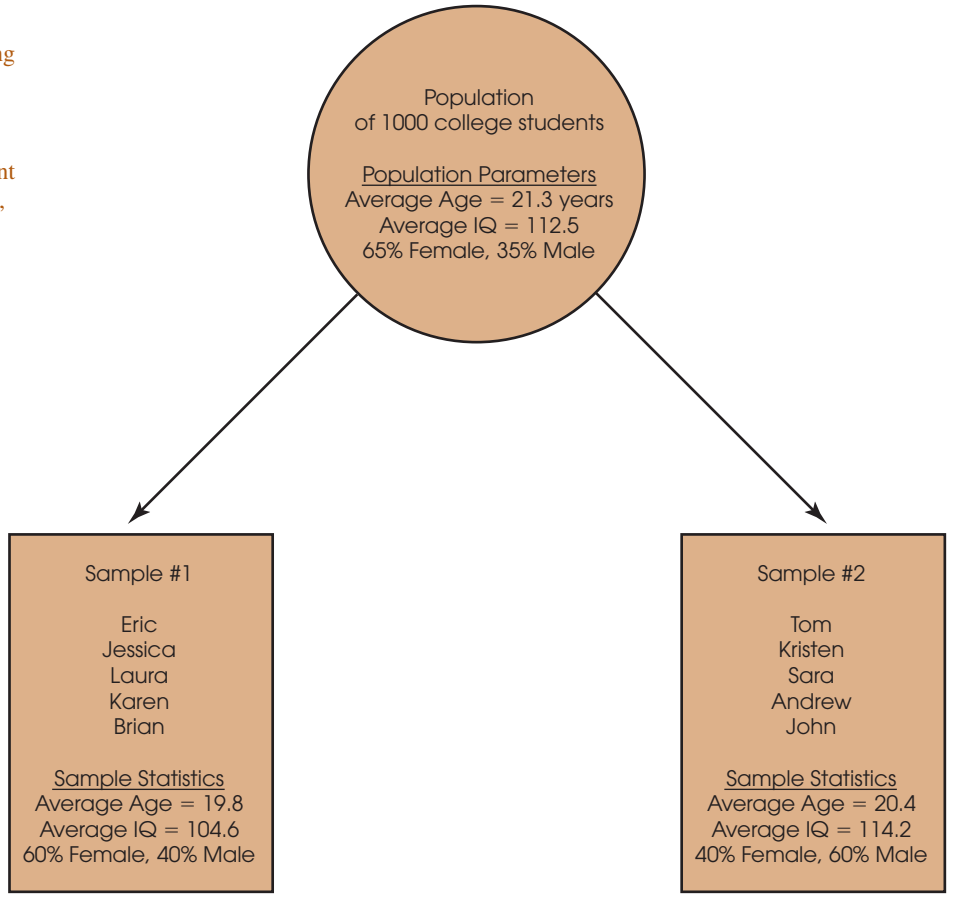


FIGURE 1.2

A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called sampling error.

It is also very unlikely that the statistics obtained for a sample are identical to the parameters for the entire population. In Figure 1.2, for example, neither sample has statistics that are exactly the same as the population parameters. You should also realize that Figure 1.2 shows only two of the hundreds of possible samples. Each sample would contain different individuals and would produce different statistics. This is the basic concept of sampling error: sample statistics vary from one sample to another and typically are different from the corresponding population parameters.

As a further demonstration of sampling error, imagine that your statistics class is separated into two groups by drawing a line from front to back through the middle of the room. Now imagine that you compute the average age (or height, or IQ) for each group. Will the two groups have exactly the same average? Almost certainly they will not. No matter what you chose to measure, you will probably find some difference between the two groups.

However, the difference you obtain does not necessarily mean that there is a systematic difference between the two groups. For example, if the average age for students on the right-hand side of the room is higher than the average for students on the left, it is unlikely that some mysterious force has caused the older people to gravitate to the right side of the room. Instead, the difference is probably the result of random factors such as chance. The unpredictable, unsystematic differences that exist from one sample to another are an example of sampling error.

STATISTICS IN THE CONTEXT OF RESEARCH

The following example shows the general stages of a research study and demonstrates how descriptive statistics and inferential statistics are used to organize and interpret the data. At the end of the example, note how sampling error can affect the interpretation of experimental results, and consider why inferential statistical methods are needed to deal with this problem.

EXAMPLE 1.1

Figure 1.3 shows an overview of a general research situation and demonstrates the roles that descriptive and inferential statistics play. The purpose of the research study

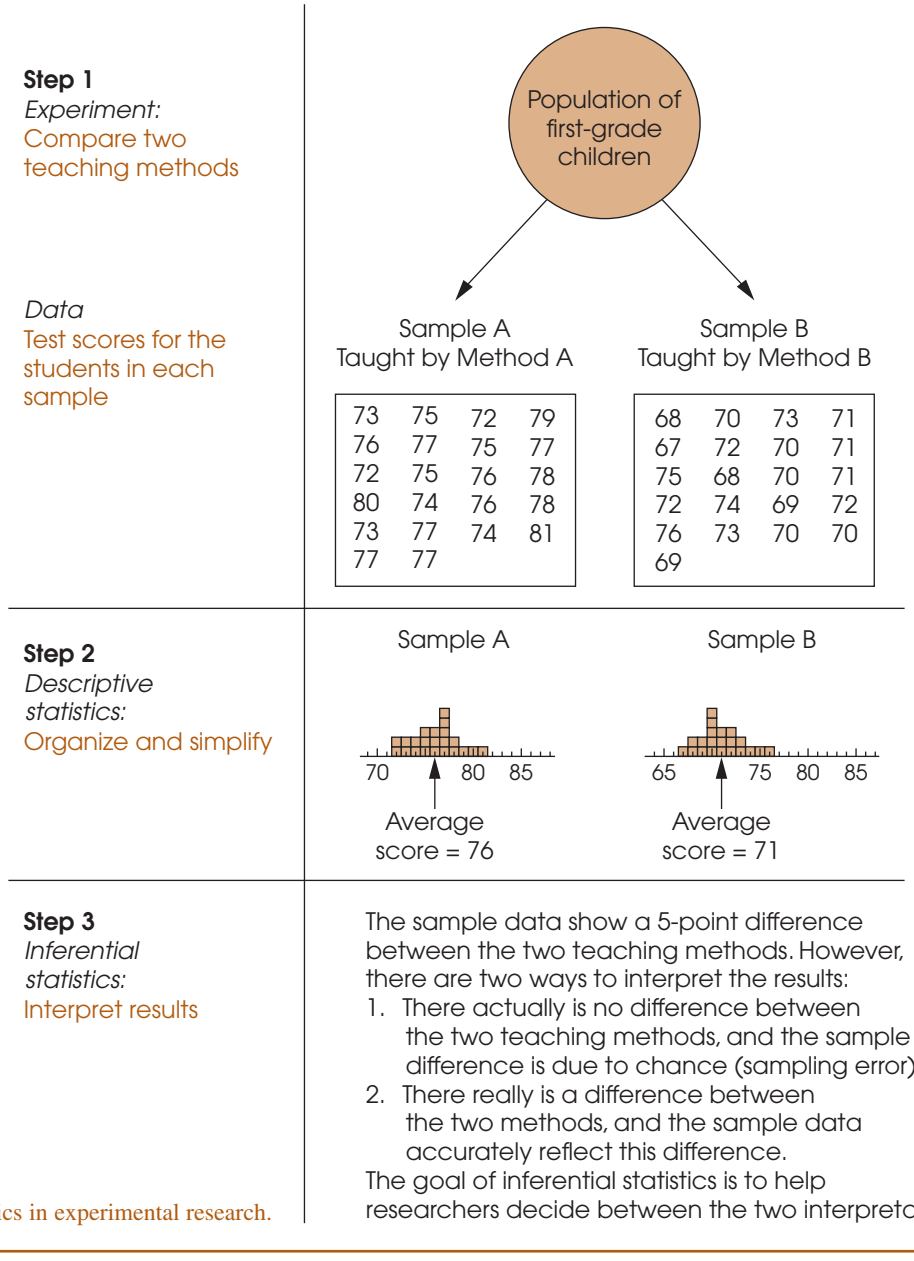


FIGURE 1.3
The role of statistics in experimental research.

is to evaluate the difference between two methods for teaching reading to first-grade children. Two samples are selected from the population of first-grade children. The children in sample A are assigned to teaching method A and the children in sample B are assigned to method B. After 6 months, all of the students are given a standardized reading test. At this point, the researcher has two sets of data: the scores for sample A and the scores for sample B (see Figure 1.3). Now is the time to begin using statistics.

First, descriptive statistics are used to simplify the pages of data. For example, the researcher could draw a graph showing the scores for each sample or compute the average score for each sample. Note that descriptive methods provide a simplified, organized description of the scores. In this example, the students taught by method A averaged 76 on the standardized test, and the students taught by method B averaged only 71.

Once the researcher has described the results, the next step is to interpret the outcome. This is the role of inferential statistics. In this example, the researcher has found a difference of 5 points between the two samples (sample A averaged 76 and sample B averaged 71). The problem for inferential statistics is to differentiate between the following two interpretations:

1. There is no real difference between the two teaching methods, and the 5-point difference between the samples is just an example of sampling error (like the samples in Figure 1.2).
2. There really is a difference between the two teaching methods, and the 5-point difference between the samples was caused by the different methods of teaching.

In simple English, does the 5-point difference between samples provide convincing evidence of a difference between the two teaching methods, or is the 5-point difference just chance? The purpose of inferential statistics is to answer this question.

LEARNING CHECK

1. A researcher is interested in the texting habits of high school students in the United States. If the researcher measures the number of text messages that each individual sends each day and calculates the average number for the entire group of high school students, the average number would be an example of a _____.
2. A researcher is interested in how watching a reality television show featuring fashion models influences the eating behavior of 13-year-old girls.
   a. A group of 30 13-year-old girls is selected to participate in a research study. The group of 30 13-year-old girls is an example of a _____.
   b. In the same study, the amount of food eaten in one day is measured for each girl and the researcher computes the average score for the 30 13-year-old girls. The average score is an example of a _____.
3. Statistical techniques are classified into two general categories. What are the two categories called, and what is the general purpose for the techniques in each category?
4. Briefly define the concept of sampling error.

ANSWERS

1. parameter
2. a. sample
   b. statistic

3. The two categories are descriptive statistics and inferential statistics. Descriptive techniques are intended to organize, simplify, and summarize data. Inferential techniques use sample data to reach general conclusions about populations.
4. Sampling error is the error, or discrepancy, between the value obtained for a sample statistic and the value for the corresponding population parameter.

1.3 DATA STRUCTURES, RESEARCH METHODS, AND STATISTICS

INDIVIDUAL VARIABLES

Some research studies are conducted simply to describe individual variables as they exist naturally. For example, a college official may conduct a survey to describe the eating, sleeping, and study habits of a group of college students. When the results consist of numerical scores, such as the number of hours spent studying each day, they are typically described by the statistical techniques that are presented in Chapters 3 and 4. Non-numerical scores are typically described by computing the proportion or percentage in each category. For example, a recent newspaper article reported that 61% of the adults in the United States currently drink alcohol.

RELATIONSHIPS BETWEEN VARIABLES

Most research, however, is intended to examine relationships between two or more variables. For example, is there a relationship between the amount of violence that children see on television and the amount of aggressive behavior they display? Is there a relationship between the quality of breakfast and level of academic performance for elementary school children? Is there a relationship between the number of hours of sleep and grade point average for college students? To establish the existence of a relationship, researchers must make observations—that is, measurements of the two variables. The resulting measurements can be classified into two distinct data structures that also help to classify different research methods and different statistical techniques. In the following section we identify and discuss these two data structures.

I. Measuring Two Variables for Each Individual: The Correlational Method

One method for examining the relationship between variables is to observe the two variables as they exist naturally for a set of individuals. That is, simply measure the two variables for each individual. For example, research has demonstrated a relationship between sleep habits, especially wake-up time, and academic performance for college students (Trockel, Barnes, and Egget, 2000). The researchers used a survey to measure wake-up time and school records to measure academic performance for each student. Figure 1.4 shows an example of the kind of data obtained in the study. The researchers then look for consistent patterns in the data to provide evidence for a relationship between variables. For example, as wake-up time changes from one student to another, is there also a tendency for academic performance to change?

Consistent patterns in the data are often easier to see if the scores are presented in a graph. Figure 1.4 also shows the scores for the eight students in a graph called a scatter plot. In the scatter plot, each individual is represented by a point so that the horizontal position corresponds to the student's wake-up time and the vertical position corresponds to the student's academic performance score. The scatter plot shows a clear relationship between wake-up time and academic performance: as wake-up time increases, academic performance decreases.

(a)

Student	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0

(b)

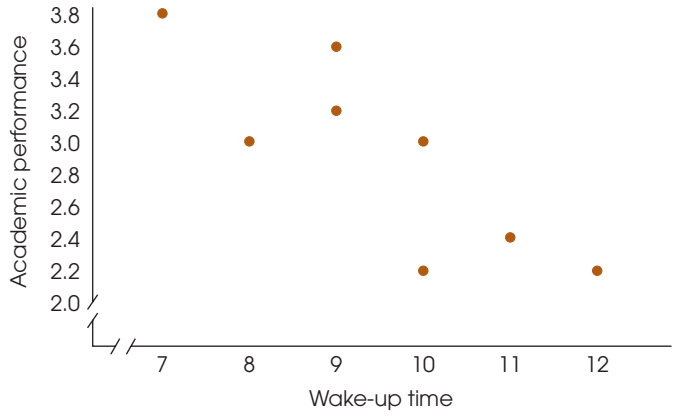


FIGURE 1.4

One of two data structures for studies evaluating the relationship between variables. Note that there are two separate measurements for each individual (wake-up time and academic performance). The same scores are shown in a table (a) and in a graph (b).

A research study that simply measures two different variables for each individual and produces the kind of data shown in Figure 1.4 is an example of the *correlational method*, or the *correlational research strategy*.

DEFINITION

In the **correlational method**, two different variables are observed to determine whether there is a relationship between them.

Limitations of the Correlational Method The results from a correlational study can demonstrate the existence of a relationship between two variables, but they do not provide an explanation for the relationship. In particular, a correlational study cannot demonstrate a cause-and-effect relationship. For example, the data in Figure 1.4 show a systematic relationship between wake-up time and academic performance for a group of college students; those who sleep late tend to have lower performance scores than those who wake early. However, there are many possible explanations for the relationship and we do not know exactly what factor (or factors) is responsible for late sleepers having lower grades. In particular, we cannot conclude that waking students up earlier would cause their academic performance to improve, or that studying more would cause students to wake up earlier. To demonstrate a cause-and-effect relationship between two variables, researchers must use the experimental method, which is discussed next.

II. Comparing Two (or More) Groups of Scores: Experimental and Nonexperimental Methods The second method for examining the relationship between two variables involves the comparison of two or more groups of scores. In this situation, the relationship between variables is examined by using one of the variables to define the groups, and then measuring the second variable to obtain scores for each group. For example, one group of elementary school children is shown a 30-minute action/adventure television program involving numerous instances of violence, and a second group is shown a 30-minute comedy that includes no violence. Both groups are

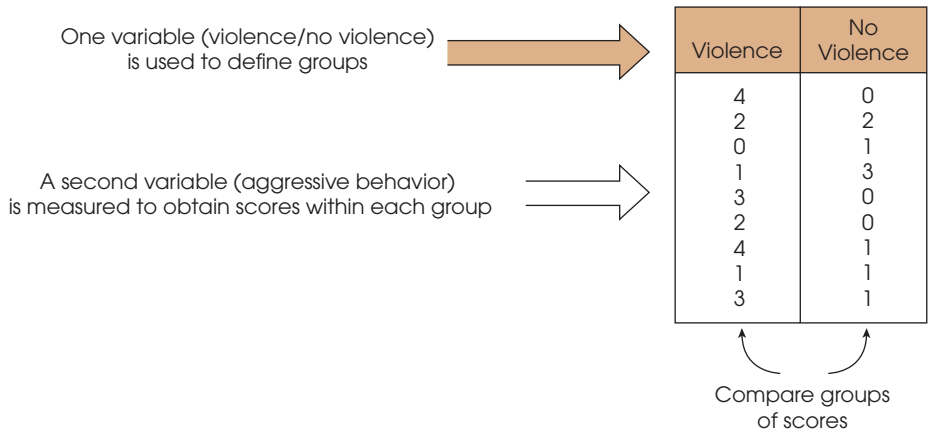


FIGURE 1.5
The second data structure for studies evaluating the relationship between variables. Note that one variable is used to define the groups and the second variable is measured to obtain scores within each group.

then observed on the playground and a researcher records the number of aggressive acts committed by each child. An example of the resulting data is shown in Figure 1.5. The researcher compares the scores for the violence group with the scores for the no-violence group. A systematic difference between the two groups provides evidence for a relationship between viewing television violence and aggressive behavior for elementary school children.

THE EXPERIMENTAL METHOD

One specific research method that involves comparing groups of scores is known as the *experimental method* or the *experimental research strategy*. The goal of an experimental study is to demonstrate a cause-and-effect relationship between two variables. Specifically, an experiment attempts to show that changing the value of one variable causes changes to occur in the second variable. To accomplish this goal, the experimental method has two characteristics that differentiate experiments from other types of research studies:

1. **Manipulation** The researcher manipulates one variable by changing its value from one level to another. A second variable is observed (measured) to determine whether the manipulation causes changes to occur.
2. **Control** The researcher must exercise control over the research situation to ensure that other, extraneous variables do not influence the relationship being examined.

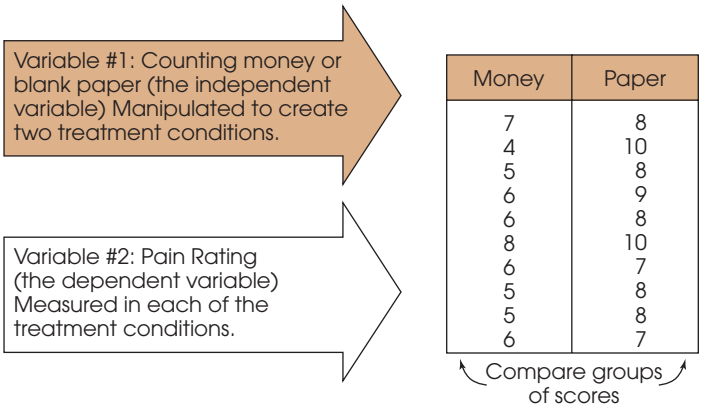
In more complex experiments, a researcher may systematically manipulate more than one variable and may observe more than one variable. Here we are considering the simplest case, in which only one variable is manipulated and only one variable is observed.

To demonstrate these two characteristics, consider an experiment in which researchers demonstrate the pain-killing effects of handling money (Zhou & Vohs, 2009). In the experiment, a group of college students was told that they were participating in a manual dexterity study. The researcher then manipulated the treatment conditions by giving half of the students a stack of money to count and the other half a stack of blank pieces of paper. After the counting task, the participants were asked to dip their hands into bowls of painfully hot water (122° F) and rate how uncomfortable it was. Participants who had counted money rated the pain significantly lower than those who had counted paper. The structure of the experiment is shown in Figure 1.6.

To be able to say that the difference in pain is caused by the money, the researcher must rule out any other possible explanation for the difference. That is, any other

FIGURE 1.6

The structure of an experiment. Participants are randomly assigned to one of two treatment conditions: counting money or counting blank pieces of paper. Later, each participant is tested by placing one hand in a bowl of hot (122° F) water and rating the level of pain. A difference between the ratings for the two groups is attributed to the treatment (paper versus money).



variables that might affect pain tolerance must be controlled. There are two general categories of variables that researchers must consider:

1. **Participant Variables** These are characteristics such as age, gender, and intelligence that vary from one individual to another. Whenever an experiment compares different groups of participants (one group in treatment A and a different group in treatment B), researchers must ensure that participant variables do not differ from one group to another. For the experiment shown in Figure 1.6, for example, the researchers would like to conclude that handling money instead of plain paper causes a change in the participants' perceptions of pain. Suppose, however, that the participants in the money condition were primarily females and those in the paper condition were primarily males. In this case, there is an alternative explanation for any difference in the pain ratings that exists between the two groups. Specifically, it is possible that the difference in pain was caused by the money, but it also is possible that the difference was caused by the participants' gender (females can tolerate more pain than males can). Whenever a research study allows more than one explanation for the results, the study is said to be *confounded* because it is impossible to reach an unambiguous conclusion.
2. **Environmental Variables** These are characteristics of the environment such as lighting, time of day, and weather conditions. A researcher must ensure that the individuals in treatment A are tested in the same environment as the individuals in treatment B. Using the money-counting experiment (see Figure 1.6) as an example, suppose that the individuals in the money condition were all tested in the morning and the individuals in the paper condition were all tested in the evening. Again, this would produce a confounded experiment because the researcher could not determine whether the differences in the pain ratings were caused by the money or caused by the time of day.

Researchers typically use three basic techniques to control other variables. First, the researcher could use *random assignment*, which means that each participant has an equal chance of being assigned to each of the treatment conditions. The goal of random assignment is to distribute the participant characteristics evenly between the two groups so that neither group is noticeably smarter (or older, or faster) than the other. Random

assignment can also be used to control environmental variables. For example, participants could be assigned randomly for testing either in the morning or in the afternoon. Second, the researcher can use *matching* to ensure equivalent groups or equivalent environments. For example, the researcher could match groups by ensuring that every group has exactly 60% females and 40% males. Finally, the researcher can control variables by *holding them constant.* For example, if an experiment uses only 10-year-old children as participants (holding age constant), then the researcher can be certain that one group is not noticeably older than another.

DEFINITION

In the **experimental method**, one variable is manipulated while another variable is observed and measured. To establish a cause-and-effect relationship between the two variables, an experiment attempts to control all other variables to prevent them from influencing the results.

Terminology in the Experimental Method Specific names are used for the two variables that are studied by the experimental method. The variable that is manipulated by the experimenter is called the *independent variable*. It can be identified as the treatment conditions to which participants are assigned. For the example in Figure 1.6, money versus paper is the independent variable. The variable that is observed and measured to obtain scores within each condition is the *dependent variable*. For the example in Figure 1.6, the level of pain is the dependent variable.

DEFINITIONS

The **independent variable** is the variable that is manipulated by the researcher. In behavioral research, the independent variable usually consists of the two (or more) treatment conditions to which subjects are exposed. The independent variable consists of the *antecedent* conditions that were manipulated *prior* to observing the dependent variable.

The **dependent variable** is the variable that is observed to assess the effect of the treatment.

Control conditions in an experiment An experimental study evaluates the relationship between two variables by manipulating one variable (the independent variable) and measuring one variable (the dependent variable). Note that in an experiment only one variable is actually measured. You should realize that this is different from a correlational study, in which both variables are measured and the data consist of two separate scores for each individual.

Often an experiment will include a condition in which the participants do not receive any treatment. The scores from these individuals are then compared with scores from participants who do receive the treatment. The goal of this type of study is to demonstrate that the treatment has an effect by showing that the scores in the treatment condition are substantially different from the scores in the no-treatment condition. In this kind of research, the no-treatment condition is called the *control condition*, and the treatment condition is called the *experimental condition*.

DEFINITIONS

Individuals in a **control condition** do not receive the experimental treatment. Instead, they either receive no treatment or they receive a neutral, placebo treatment. The purpose of a control condition is to provide a baseline for comparison with the experimental condition.

Individuals in the **experimental condition** do receive the experimental treatment.

Note that the independent variable always consists of at least two values. (Something must have at least two different values before you can say that it is "variable.") For the money-counting experiment (see Figure 1.6), the independent variable is money versus plain paper. For an experiment with an experimental group and a control group, the independent variable is treatment versus no treatment.

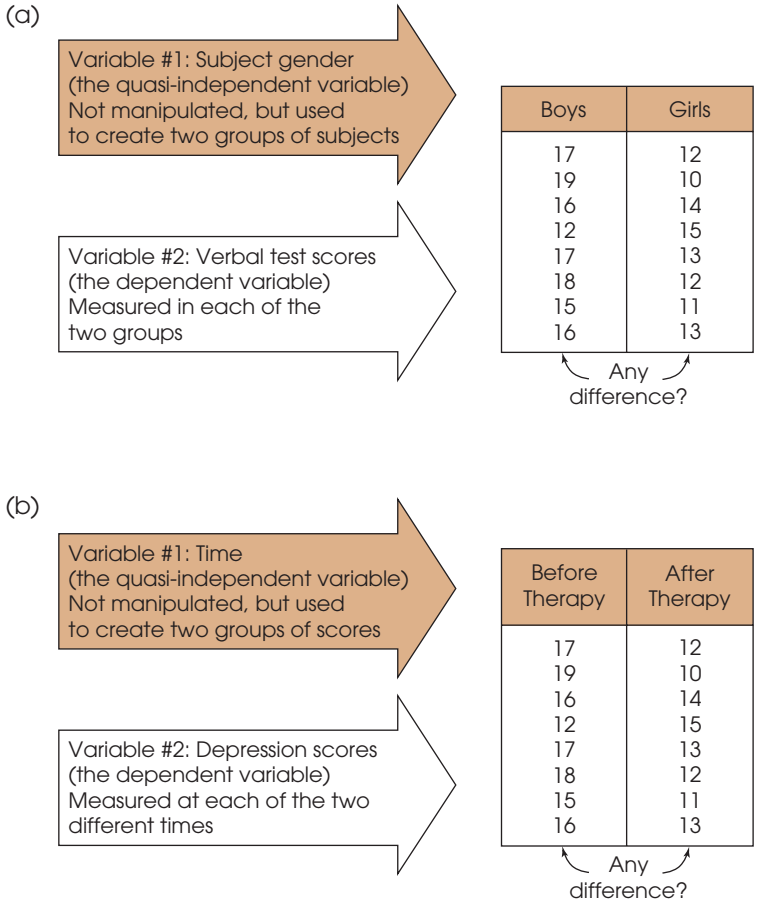
NONEXPERIMENTAL METHODS: NONEQUIVALENT GROUPS AND PRE–POST STUDIES

In informal conversation, there is a tendency for people to use the term *experiment* to refer to any kind of research study. You should realize, however, that the term only applies to studies that satisfy the specific requirements outlined earlier. In particular, a real experiment must include manipulation of an independent variable and rigorous control of other, extraneous variables. As a result, there are a number of other research designs that are not true experiments but still examine the relationship between variables by comparing groups of scores. Two examples are shown in Figure 1.7 and are discussed in the following paragraphs. This type of research study is classified as nonexperimental.

The top part of Figure 1.7 shows an example of a *nonequivalent groups* study comparing boys and girls. Notice that this study involves comparing two groups of scores (like an experiment). However, the researcher has no ability to control which

FIGURE 1.7

Two examples of nonexperimental studies that involve comparing two groups of scores. In (a) the study uses two preexisting groups (boys/girls) and measures a dependent variable (verbal scores) in each group. In (b) the study uses time (before/after) to define the two groups and measures a dependent variable (depression) in each group.



participants go into which group—all the males must be in the boy group and all the females must be in the girl group. Because this type of research compares preexisting groups, the researcher cannot control the assignment of participants to groups and cannot ensure equivalent groups. Other examples of nonequivalent group studies include comparing 8-year-old children and 10-year-old children, people with an eating disorder and those with no disorder, and comparing children from a single-parent home and those from a two-parent home. Because it is impossible to use techniques like random assignment to control participant variables and ensure equivalent groups, this type of research is not a true experiment.

Correlational studies are also examples of nonexperimental research. In this section, however, we are discussing non-experimental studies that compare two or more groups of scores.

The bottom part of Figure 1.7 shows an example of a *pre–post* study comparing depression scores before therapy and after therapy. The two groups of scores are obtained by measuring the same variable (depression) twice for each participant; once before therapy and again after therapy. In a pre–post study, however, the researcher has no control over the passage of time. The "before" scores are always measured earlier than the "after" scores. Although a difference between the two groups of scores may be caused by the treatment, it is always possible that the scores simply change as time goes by. For example, the depression scores may decrease over time in the same way that the symptoms of a cold disappear over time. In a pre–post study, the researcher also has no control over other variables that change with time. For example, the weather could change from dark and gloomy before therapy to bright and sunny after therapy. In this case, the depression scores could improve because of the weather and not because of the therapy. Because the researcher cannot control the passage of time or other variables related to time, this study is not a true experiment.

Terminology in nonexperimental research Although the two research studies shown in Figure 1.7 are not true experiments, you should notice that they produce the same kind of data that are found in an experiment (see Figure 1.6). In each case, one variable is used to create groups, and a second variable is measured to obtain scores within each group. In an experiment, the groups are created by manipulation of the independent variable, and the participants' scores are the dependent variable. The same terminology is often used to identify the two variables in nonexperimental studies. That is, the variable that is used to create groups is the independent variable and the scores are the dependent variable. For example, the top part of Figure 1.7, gender (boy/girl), is the independent variable and the verbal test scores are the dependent variable. However, you should realize that gender (boy/girl) is not a true independent variable because it is not manipulated. For this reason, the "independent variable" in a nonexperimental study is often called a *quasi-independent variable.*

DEFINITION

In a nonexperimental study, the "independent variable" that is used to create the different groups of scores is often called the **quasi-independent variable**.

DATA STRUCTURES AND STATISTICAL METHODS

The two general data structures that we used to classify research methods can also be used to classify statistical methods.

I. One Group with Two Variables Measured for each Individual Recall that the data from a correlational study consist of two scores, representing two different variables, for each individual. The scores can be listed in a table or displayed in a scatter plot as in Figure 1.5. The relationship between the two variables is usually measured and described using a statistic called a *correlation*. Correlations and the correlational method are discussed in detail in Chapters15 and 16.

Occasionally, the measurement process used for a correlational study simply classifies individuals into categories that do not correspond to numerical values. For example, a researcher could classify a group of college students by gender (male or female) and by cell-phone preference (talk or text). Note that the researcher has two scores for each individual but neither of the scores is a numerical value. This type of data is typically summarized in a table showing how many individuals are classified into each of the possible categories. Table 1.1 shows an example of this kind of summary table. The table shows, for example, that 30 of the males in the sample preferred texting to talking. This type of data can be coded with numbers (for example, male = 0 and female = 1) so that it is possible to compute a correlation. However, the relationship between variables for non-numerical data, such as the data in Table 1.1, is usually evaluated using a statistical technique known as a *chi-square test.* Chi-square tests are presented in Chapter 17.

II. Comparing Two or More Groups of Scores Most of the statistical procedures presented in this book are designed for research studies that compare groups of scores, like the experimental study in Figure 1.6 and the nonexperimental studies in Figure 1.7. Specifically, we examine descriptive statistics that summarize and describe the scores in each group, and we examine inferential statistics that allow us to use the groups, or samples, to generalize to the entire population.

When the measurement procedure produces numerical scores, the statistical evaluation typically involves computing the average score for each group and then comparing the averages. The process of computing averages is presented in Chapter 3, and a variety of statistical techniques for comparing averages are presented in Chapters 8–14. If the measurement process simply classifies individuals into non-numerical categories, the statistical evaluation usually consists of computing proportions for each group and then comparing proportions. In Table 1.1 we present an example of non-numerical data examining the relationship between gender and cell-phone preference. The same data can be used to compare the proportions for males with the proportions for females. For example, using text is preferred by 60% of the males compared to 50% of the females. As mentioned before, these data are evaluated using a chi-square test, which is presented in Chapter 17.

TABLE 1.1

Correlational data consisting of non-numerical scores. Note that there are two measurements for each individual: gender and cell phone preference. The numbers indicate how many people are in each category. For example, out of the 50 males, 30 prefer text over talk.

	Cell Phone Preference		
	Text	Talk	
Males	30	20	50
Females	25	25	50

LEARNING CHECK

1. Researchers have observed that high school students who watched educational television programs as young children tend to have higher grades than their peers who did not watch educational television. Is this study an example of an experiment? Explain why or why not.
2. What two elements are necessary for a research study to be an experiment?
3. Loftus and Palmer (1974) conducted an experiment in which participants were shown a video of an automobile accident. After the video, some participants were

asked to estimate the speed of the cars when they "smashed into" each other. Others were asked to estimate the speed when the cars "hit" each other. The "smashed into" group produced significantly higher estimates than the "hit" group. Identify the independent and dependent variables for this study.

ANSWERS

1. This study could be correlational or nonexperimental, but it is definitely not an example of a true experiment. The researcher is simply observing, not manipulating, the amount of educational television.
2. First, the researcher must manipulate one of the two variables being studied. Second, all other variables that might influence the results must be controlled.
3. The independent variable is the phrasing of the question and the dependent variable is the speed estimated by each participant.

1.4 VARIABLES AND MEASUREMENT

The scores that make up the data from a research study are the result of observing and measuring variables. For example, a researcher may finish a study with a set of IQ scores, personality scores, or reaction-time scores. In this section, we take a closer look at the variables that are being measured and the process of measurement.

CONSTRUCTS AND OPERATIONAL DEFINITIONS

Some variables, such as height, weight, and eye color are well-defined, concrete entities that can be observed and measured directly. On the other hand, many variables studied by behavioral scientists are internal characteristics that people use to help describe and explain behavior. For example, we say that a student does well in school because he or she is *intelligent*. Or we say that someone is *anxious* in social situations, or that someone seems to be *hungry*. Variables like intelligence, anxiety, and hunger are called *constructs*, and because they are intangible and cannot be directly observed, they are often called *hypothetical constructs*.

Although constructs such as intelligence are internal characteristics that cannot be directly observed, it is possible to observe and measure behaviors that are representative of the construct. For example, we cannot "see" intelligence but we can see examples of intelligent behavior. The external behaviors can then be used to create an operational definition for the construct. An *operational definition* defines a construct in terms of external behaviors that can be observed and measured. For example, your intelligence is measured and defined by your performance on an IQ test, or hunger can be measured and defined by the number of hours since last eating.

DEFINITIONS

Constructs are internal attributes or characteristics that cannot be directly observed but are useful for describing and explaining behavior.

An **operational definition** identifies a measurement procedure (a set of operations) for measuring an external behavior and uses the resulting measurements as a definition and a measurement of a hypothetical construct. Note that an operational definition has two components: First, it describes a set of operations for measuring a construct. Second, it defines the construct in terms of the resulting measurements.

DISCRETE AND CONTINUOUS VARIABLES

The variables in a study can be characterized by the type of values that can be assigned to them. A *discrete variable* consists of separate, indivisible categories. For this type of variable, there are no intermediate values between two adjacent categories. Consider the values displayed when dice are rolled. Between neighboring values—for example, five dots and six dots—no other values can ever be observed.

DEFINITION

A **discrete variable** consists of separate, indivisible categories. No values can exist between two neighboring categories.

Discrete variables are commonly restricted to whole, countable numbers—for example, the number of children in a family or the number of students attending class. If you observe class attendance from day to day, you may count 18 students one day and 19 students the next day. However, it is impossible ever to observe a value between 18 and 19. A discrete variable may also consist of observations that differ qualitatively. For example, people can be classified by gender (male or female), by occupation (nurse, teacher, lawyer, etc.), and college students can be classified by academic major (art, biology, chemistry, etc.). In each case, the variable is discrete because it consists of separate, indivisible categories.

On the other hand, many variables are not discrete. Variables such as time, height, and weight are not limited to a fixed set of separate, indivisible categories. You can measure time, for example, in hours, minutes, seconds, or fractions of seconds. These variables are called *continuous* because they can be divided into an infinite number of fractional parts.

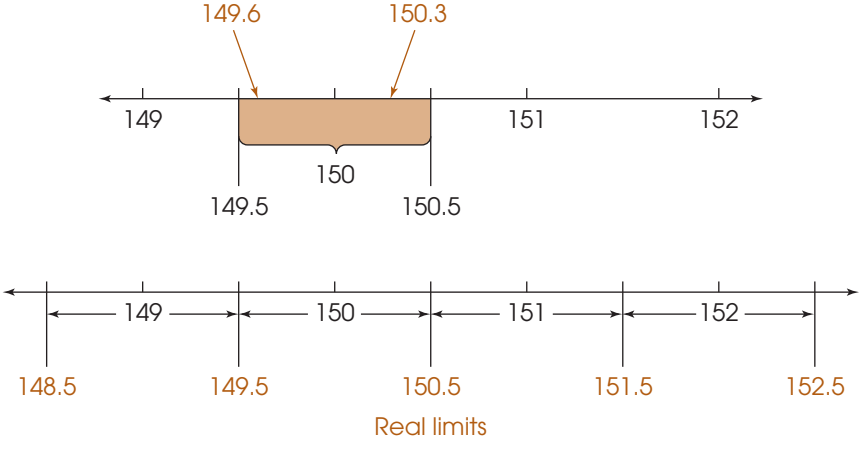
DEFINITION

For a **continuous variable**, there are an infinite number of possible values that fall between any two observed values. A continuous variable is divisible into an infinite number of fractional parts.

Suppose, for example, that a researcher is measuring weights for a group of individuals participating in a diet study. Because weight is a continuous variable, it can be pictured as a continuous line (Figure 1.8). Note that there are an infinite number of possible points on the line without any gaps or separations between neighboring points. For

FIGURE 1.8

When measuring weight to the nearest whole pound, 149.6 and 150.3 are assigned the value of 150 (top). Any value in the interval between 149.5 and 150.5 is given the value of 150.



any two different points on the line, it is always possible to find a third value that is between the two points.

Two other factors apply to continuous variables:

1. When measuring a continuous variable, it should be very rare to obtain identical measurements for two different individuals. Because a continuous variable has an infinite number of possible values, it should be almost impossible for two people to have exactly the same score. If the data show a substantial number of tied scores, then you should suspect that the measurement procedure is very crude or that the variable is not really continuous.
2. When measuring a continuous variable, each measurement category is actually an *interval* that must be defined by boundaries. For example, two people who both claim to weigh 150 pounds are probably not *exactly* the same weight. However, they are both around 150 pounds. One person may actually weigh 149.6 and the other 150.3. Thus, a score of 150 is not a specific point on the scale but instead is an interval (see Figure 1.8). To differentiate a score of 150 from a score of 149 or 151, we must set up boundaries on the scale of measurement. These boundaries are called *real limits* and are positioned exactly halfway between adjacent scores. Thus, a score of $X = 150$ pounds is actually an interval bounded by a *lower real limit* of 149.5 at the bottom and an *upper real limit* of 150.5 at the top. Any individual whose weight falls between these real limits will be assigned a score of $X = 150$.

DEFINITIONS

Real limits are the boundaries of intervals for scores that are represented on a continuous number line. The real limit separating two adjacent scores is located exactly halfway between the scores. Each score has two real limits. The **upper real limit** is at the top of the interval, and the **lower real limit** is at the bottom.

Technical Note: Students often ask whether a value of exactly 150.5 should be assigned to the $X = 150$ interval or the $X = 151$ interval. The answer is that 150.5 is the *boundary* between the two intervals and is not necessarily in one or the other. Instead, the placement of 150.5 depends on the rule that you are using for rounding numbers. If you are rounding up, then 150.5 goes in the higher interval ($X = 151$) but if you are rounding down, then it goes in the lower interval ($X = 150$).

The concept of real limits applies to any measurement of a continuous variable, even when the score categories are not whole numbers. For example, if you were measuring time to the nearest tenth of a second, the measurement categories would be 31.0, 31.1, 31.2, and so on. Each of these categories represents an interval on the scale that is bounded by real limits. For example, a score of $X = 31.1$ seconds indicates that the actual measurement is in an interval bounded by a lower real limit of 31.05 and an upper real limit of 31.15. Remember that the real limits are always halfway between adjacent categories.

Later in this book, real limits are used for constructing graphs and for various calculations with continuous scales. For now, however, you should realize that real limits are a necessity whenever you make measurements of a continuous variable.

Finally, we should warn you that the terms *continuous* and *discrete* apply to the variables that are being measured and not to the scores that are obtained from the measurement. For example, measuring people's heights to the nearest inch produces scores of 60, 61, 62, and so on. Although the scores may appear to be discrete numbers, the underlying variable is continuous. One key to determining whether a variable is continuous or discrete is that a continuous variable can be divided into any number of fractional parts. Height can be measured to the nearest inch, the nearest 0.5 inch, or the nearest 0.1 inch. Similarly, a professor evaluating students' knowledge could use a pass/fail system that classifies students into two broad categories. However, the professor could choose to use a 10-point quiz that divides student knowledge into 11 categories corresponding to quiz scores from 0 to 10. Or the professor could use a 100-point exam that potentially divides student knowledge into 101 categories from 0 to 100. Whenever you are free to choose the degree of precision or the number of categories for measuring a variable, the variable must be continuous.

SCALES OF MEASUREMENT

It should be obvious by now that data collection requires that we make measurements of our observations. Measurement involves assigning individuals or events to categories. The categories can simply be names such as male/female or employed/unemployed, or they can be numerical values such as 68 inches or 175 pounds. The categories used to measure a variable make up a *scale of measurement*, and the relationships between the categories determine different types of scales. The distinctions among the scales are important because they identify the limitations of certain types of measurements and because certain statistical procedures are appropriate for scores that have been measured on some scales but not on others. If you were interested in people's heights, for example, you could measure a group of individuals by simply classifying them into three categories: tall, medium, and short. However, this simple classification would not tell you much about the actual heights of the individuals, and these measurements would not give you enough information to calculate an average height for the group. Although the simple classification would be adequate for some purposes, you would need more sophisticated measurements before you could answer more detailed questions. In this section, we examine four different scales of measurement, beginning with the simplest and moving to the most sophisticated.

THE NOMINAL SCALE

The word *nominal* means "having to do with names." Measurement on a nominal scale involves classifying individuals into categories that have different names but are not related to each other in any systematic way. For example, if you were measuring the academic majors for a group of college students, the categories would be art, biology, business, chemistry, and so on. Each student would be classified in one category according to his or her major. The measurements from a nominal scale allow us to determine whether two individuals are different, but they do not identify either the direction or the size of the difference. If one student is an art major and another is a biology major we can say that they are different, but we cannot say that art is "more than" or "less than" biology and we cannot specify how much difference there is between art and biology. Other examples of nominal scales include classifying people by race, gender, or occupation.

DEFINITION

A **nominal scale** consists of a set of categories that have different names. Measurements on a nominal scale label and categorize observations, but do not make any quantitative distinctions between observations.

Although the categories on a nominal scale are not quantitative values, they are occasionally represented by numbers. For example, the rooms or offices in a building may be identified by numbers. You should realize that the room numbers are simply names and do not reflect any quantitative information. Room 109 is not necessarily bigger than Room 100 and certainly not 9 points bigger. It also is fairly common to use numerical values as a code for nominal categories when data are entered into computer programs. For example, the data from a survey may code males with a 0 and females with a 1. Again, the numerical values are simply names and do not represent any quantitative difference. The scales that follow do reflect an attempt to make quantitative distinctions.

THE ORDINAL SCALE

The categories that make up an *ordinal scale* not only have different names (as in a nominal scale) but also are organized in a fixed order corresponding to differences of magnitude.

DEFINITION

An **ordinal scale** consists of a set of categories that are organized in an ordered sequence. Measurements on an ordinal scale rank observations in terms of size or magnitude.

Often, an ordinal scale consists of a series of ranks (first, second, third, and so on) like the order of finish in a horse race. Occasionally, the categories are identified by verbal labels like small, medium, and large drink sizes at a fast-food restaurant. In either case, the fact that the categories form an ordered sequence means that there is a directional relationship between categories. With measurements from an ordinal scale, you can determine whether two individuals are different and you can determine the direction of difference. However, ordinal measurements do not allow you to determine the size of the difference between two individuals. For example, if Billy is placed in the low-reading group and Tim is placed in the high-reading group, you know that Tim is a better reader, but you do not know how much better. Other examples of ordinal scales include socioeconomic class (upper, middle, lower) and T-shirt sizes (small, medium, large). In addition, ordinal scales are often used to measure variables for which it is difficult to assign numerical scores. For example, people can rank their food preferences but might have trouble explaining "how much" they prefer chocolate ice cream to steak.

THE INTERVAL AND RATIO SCALES

Both an *interval scale* and a *ratio scale* consist of a series of ordered categories (like an ordinal scale) with the additional requirement that the categories form a series of intervals that are all exactly the same size. Thus, the scale of measurement consists of a series of equal intervals, such as inches on a ruler. Other examples of interval and ratio scales are the measurement of time in seconds, weight in pounds, and temperature in degrees Fahrenheit. Note that, in each case, one interval (1 inch, 1 second, 1 pound, 1 degree) is the same size, no matter where it is located on the scale. The fact that the intervals are all the same size makes it possible to determine both the size and the direction of the difference between two measurements. For example, you know that a measurement of 80° Fahrenheit is higher than a measure of 60°, and you know that it is exactly 20° higher.

The factor that differentiates an interval scale from a ratio scale is the nature of the zero point. An interval scale has an arbitrary zero point. That is, the value 0 is assigned to a particular location on the scale simply as a matter of convenience or reference. In particular, a value of zero does not indicate a total absence of the variable being measured. For example a temperature of 0° Fahrenheit does not mean that there is no temperature, and it does not prohibit the temperature from going even lower. Interval scales with an arbitrary zero point are relatively rare. The two most common examples are the Fahrenheit and Celsius temperature scales. Other examples include golf scores (above and below par) and relative measures such as above and below average rainfall.

A ratio scale is anchored by a zero point that is not arbitrary but rather is a meaningful value representing none (a complete absence) of the variable being measured. The existence of an absolute, nonarbitrary zero point means that we can measure the absolute amount of the variable; that is, we can measure the distance from 0. This makes it possible to compare measurements in terms of ratios. For example, an individual who requires 10 seconds to solve a problem (10 more than 0) has taken twice as much time as an individual who finishes in only 5 seconds (5 more than 0). With a ratio scale, we can measure the direction and the size of the difference between two measurements and we can describe the difference in terms of a ratio. Ratio scales are quite common and include physical measures such as height and weight, as well as variables such as reaction time or the number of errors on a test. The distinction between an interval scale and a ratio scale is demonstrated in Example 1.2.

DEFINITIONS

An **interval scale** consists of ordered categories that are all intervals of exactly the same size. Equal differences between numbers on a scale reflect equal differences in magnitude. However, the zero point on an interval scale is arbitrary and does not indicate a zero amount of the variable being measured.

A **ratio scale** is an interval scale with the additional feature of an absolute zero point. With a ratio scale, ratios of numbers do reflect ratios of magnitude.

EXAMPLE 1.2

A researcher obtains measurements of height for a group of 8-year-old boys. Initially, the researcher simply records each child's height in inches, obtaining values such as 44, 51, 49, and so on. These initial measurements constitute a ratio scale. A value of zero represents no height (absolute zero). Also, it is possible to use these measurements to form ratios. For example, a child who is 60 inches tall is one-and-a-half times taller than a child who is 40 inches tall.

Now suppose that the researcher converts the initial measurement into a new scale by calculating the difference between each child's actual height and the average height for this age group. A child who is 1 inch taller than average now gets a score of +1; a child 4 inches taller than average gets a score of +4. Similarly, a child who is 2 inches shorter than average gets a score of −2. On this scale, a score of zero corresponds to average height. Because zero no longer indicates a complete absence of height, the new scores constitute an interval scale of measurement.

Notice that original scores and the converted scores both involve measurement in inches, and you can compute differences, or distances, on either scale. For example, there is a 6-inch difference in height between two boys who measure 57 and 51 inches tall on the first scale. Likewise, there is a 6-inch difference between two boys who measure +9 and +3 on the second scale. However, you should also notice that ratio comparisons are not possible on the second scale. For example, a boy who measures +9 is not three times taller than a boy who measures +3.

STATISTICS AND SCALES OF MEASUREMENT

For our purposes, scales of measurement are important because they influence the kind of statistics that can and cannot be used. For example, if you measure IQ scores for a group of students, it is possible to add the scores together and calculate a mean score for the group. On the other hand, if you measure the academic major for each student, you cannot compute the mean. (What is the mean of three psychology majors, an English major, and two chemistry majors?) The vast majority of the statistical techniques presented in this book are designed for numerical scores from an interval or a ratio scale. For most statistical applications, the distinction between an interval scale and a ratio scale is not important because both scales produce numerical values that permit us to compute differences between scores, to add scores, and to calculate mean scores. On the other hand, measurements from nominal or ordinal scales are typically not numerical values and are not compatible with many basic arithmetic operations. Therefore, alternative statistical techniques are necessary for data from nominal or ordinal scales of measurement (for example, the median and the mode in Chapter 3, the Spearman correlation in Chapter 15, and the chi-square tests in Chapter 17). Additional statistical methods for measurements from ordinal scales are presented in Appendix E.

LEARNING CHECK

1. A survey asks people to identify their age, annual income, and marital status (single, married, divorced, etc.). For each of these three variables, identify the scale of measurement that probably is used and identify whether the variable is continuous or discrete.
2. An English professor uses letter grades (A, B, C, D, and F) to evaluate a set of student essays. What kind of scale is being used to measure the quality of the essays?
3. The teacher in a communications class asks students to identify their favorite reality television show. The different television shows make up a _____ scale of measurement.
4. A researcher studies the factors that determine the number of children that couples decide to have. The variable, number of children, is a _____ (discrete/continuous) variable.
5. **a.** When measuring height to the nearest inch, what are the real limits for a score of 68 inches?
   b. When measuring height to the nearest half inch, what are the real limits for a score of 68 inches?

ANSWERS

1. Age and annual income are measured on ratio scales and are both continuous variables. Marital status is measured on a nominal scale and is a discrete variable.
2. ordinal
3. nominal
4. discrete
5. **a.** 67.5 and 68.5
   b. 67.75 and 68.25

1.5 STATISTICAL NOTATION

The measurements obtained in research studies provide the data for statistical analysis. Most of the statistical analyses use the same general mathematical operations, notation, and basic arithmetic that you have learned during previous years of school. In case you are unsure of your mathematical skills, there is a mathematics review section in Appendix A at the back of this book. The appendix also includes a skills-assessment exam (p. 678) to help you determine whether you need the basic mathematics review. In this section, we introduce some of the specialized notation that is used for statistical calculations. In later chapters, additional statistical notation is introduced as it is needed.

Measuring a variable in a research study typically yields a value or a score for each individual. Raw scores are the original, unchanged scores obtained in the study. Scores for a particular variable are represented by the letter X. For example, if performance in your statistics course is measured by tests and you obtain a 35 on the first test, then we could state that $X = 35$. A set of scores can be presented in a column that is headed by X. For example, a list of quiz scores from your class might be presented as shown in the margin (the single column on the left).

X
37
35
35
30
25
17
16

Scores	
X	Y
72	165
68	151
67	160
67	160
68	146
70	160
66	133

When observations are made for two variables, there will be two scores for each individual. The data can be presented as two lists labeled X and Y for the two variables. For example, observations for people's height in inches (variable X) and weight in pounds (variable Y) can be presented as shown in the double column in the margin. Each pair X, Y represents the observations made of a single participant.

The letter N is used to specify how many scores are in a set. An uppercase letter N identifies the number of scores in a population and a lowercase letter n identifies the number of scores in a sample. Throughout the remainder of the book you will notice that we often use notational differences to distinguish between samples and populations. For the height and weight data in the preceding table, $n = 7$ for both variables. Note that by using a lowercase letter n, we are implying that these data are a sample.

SUMMATION NOTATION

Many of the computations required in statistics involve adding a set of scores. Because this procedure is used so frequently, a special notation is used to refer to the sum of a set of scores. The Greek letter *sigma*, or Σ, is used to stand for summation. The expression ΣX means to add all the scores for variable X. The summation sign, Σ, can be read as "the sum of." Thus, ΣX is read "the sum of the scores." For the following set of quiz scores,

$$10, \quad 6, \quad 7, \quad 4$$

$\Sigma X = 27$ and $N = 4$.

To use summation notation correctly, keep in mind the following two points:

1. The summation sign, Σ, is always followed by a symbol or mathematical expression. The symbol or expression identifies exactly which values are to be added. To compute ΣX, for example, the symbol following the summation sign is X, and the task is to find the sum of the X values. On the other hand, to compute $\Sigma(X - 1)^2$, the summation sign is followed by a relatively complex mathematical expression, so your first task is to calculate all of the $(X - 1)^2$ values and then add the results.
2. The summation process is often included with several other mathematical operations, such as multiplication or squaring. To obtain the correct answer, it is essential that the different operations be done in the correct sequence. Following is a list showing the correct *order of operations* for performing mathematical operations. Most of this list should be familiar, but you should note that we have inserted the summation process as the fourth operation in the list.

Order of Mathematical Operations

More information on the order of operations for mathematics is available in the Math Review appendix, page 679.

1. Any calculation contained within parentheses is done first.
2. Squaring (or raising to other exponents) is done second.
3. Multiplying and/or dividing is done third. A series of multiplication and/or division operations should be done in order from left to right.
4. Summation using the Σ notation is done next.
5. Finally, any other addition and/or subtraction is done.

The following examples demonstrate how summation notation is used in most of the calculations and formulas we present in this book.

EXAMPLE 1.3

A set of four scores consists of values 3, 1, 7, and 4. We will compute ΣX, ΣX^2, and $(\Sigma X)^2$ for these scores. To help demonstrate the calculations, we will use a computational table showing the original scores (the X values) in the first column. Additional columns can then be added to show additional steps in the series of operations. You should notice that the first three operations in the list (parentheses, squaring, and multiplying) all create a new column of values. The last two operations, however, produce a single value corresponding to the sum.

X	X^2
3	9
1	1
7	49
4	16

The table to the left shows the original scores (the X values) and the squared scores (the X^2 values) that are needed to compute ΣX^2.

The first calculation, ΣX, does not include any parentheses, squaring, or multiplication, so we go directly to the summation operation. The X values are listed in the first column of the table, and we simply add the values in this column:

$$\Sigma X = 3 + 1 + 7 + 4 = 15$$

To compute ΣX^2, the correct order of operations is to square each score and then find the sum of the squared values. The computational table shows the original scores and the results obtained from squaring (the first step in the calculation). The second step is to find the sum of the squared values, so we simply add the numbers in the X^2 column.

$$\Sigma X^2 = 9 + 1 + 49 + 16 = 75$$

The final calculation, $(\Sigma X)^2$, includes parentheses, so the first step is to perform the calculation inside the parentheses. Thus, we first find ΣX and then square this sum. Earlier, we computed $\Sigma X = 15$, so

$$(\Sigma X)^2 = (15)^2 = 225$$

EXAMPLE 1.4

Use the same set of four scores from Example 1.3 and compute $\Sigma(X - 1)$ and $\Sigma(X - 1)^2$. The following computational table will help demonstrate the calculations.

X	$(X - 1)$	$(X - 1)^2$
3	2	4
1	0	0
7	6	36
4	3	9

The first column lists the original scores. A second column lists the (X − 1) values, and a third column shows the $(X - 1)^2$ values.

To compute $\Sigma(X - 1)$, the first step is to perform the operation inside the parentheses. Thus, we begin by subtracting one point from each of the X values. The resulting values are listed in the middle column of the table. The next step is to add the $(X - 1)$ values.

$$\Sigma(X - 1) = 2 + 0 + 6 + 3 + = 11$$

The calculation of $\Sigma(X + 1)^2$ requires three steps. The first step (inside parentheses) is to subtract 1 point from each X value. The results from this step are shown in the middle column of the computational table. The second step is to square each of the

$(X - 1)$ values. The results from this step are shown in the third column of the table. The final step is to add the $(X - 1)^2$ values to obtain

$$\Sigma(X - 1)^2 = 4 + 0 + 36 + 9 = 49$$

Notice that this calculation requires squaring before adding. A common mistake is to add the $(X - 1)$ values and then square the total. Be careful!

EXAMPLE 1.5

In both of the preceding examples, and in many other situations, the summation operation is the last step in the calculation. According to the order of operations, parentheses, exponents, and multiplication all come before summation. However, there are situations in which extra addition and subtraction are completed after the summation. For this example, use the same scores that appeared in the previous two examples, and compute $\Sigma X - 1$.

With no parentheses, exponents, or multiplication, the first step is the summation. Thus, we begin by computing ΣX. Earlier we found $\Sigma X = 15$. The next step is to subtract one point from the total. For these data,

$$\Sigma X - 1 = 15 - 1 = 14$$

EXAMPLE 1.6

For this example, each individual has two scores. The first score is identified as X, and the second score is Y. With the help of the following computational table, compute ΣX, ΣY, and ΣXY.

Person	X	Y	XY
A	3	5	15
B	1	3	3
C	7	4	28
D	4	2	8

To find ΣX, simply add the values in the X column.

$$\Sigma X = 3 + 1 + 7 + 4 = 15$$

Similarly, ΣY is the sum of the Y values.

$$\Sigma Y = 5 + 3 + 4 + 2 = 14$$

To compute ΣXY, the first step is to multiply X times Y for each individual. The resulting products (XY values) are listed in the third column of the table. Finally, we add the products to obtain

$$\Sigma XY = 15 + 3 + 28 + 8 = 54$$

LEARNING CHECK

1. Calculate each value requested for the following scores: 6, 2, 4, 2.
   - **a.** ΣX
   - **b.** ΣX^2
   - **c.** $(\Sigma X)^2$
   - **d.** $\Sigma(X - 2)$
   - **e.** $\Sigma(X - 2)^2$
2. Identify the first step in each of the following calculations.
   a. ΣX^2 **b.** $(\Sigma X)^2$ **c.** $\Sigma(X - 2)^2$
3. Use summation notation to express each of the following.
   - **a.** Add 4 points to each score and then add the resulting values.
   - **b.** Add the scores and then square the total.
   - **c.** Square each score, then add the squared values.

ANSWERS

1. **a.** 14
   b. 60
   c. 196
   d. 6
   e. 20
2. **a.** Square each score.
   b. Add the scores.
   c. Subtract 2 points from each score.
3. **a.** $\Sigma(X + 4)$
   b. $(\Sigma X)^2$
   c. ΣX^2

SUMMARY

1. The term *statistics* is used to refer to methods for organizing, summarizing, and interpreting data.
2. Scientific questions usually concern a population, which is the entire set of individuals one wishes to study. Usually, populations are so large that it is impossible to examine every individual, so most research is conducted with samples. A sample is a group selected from a population, usually for purposes of a research study.
3. A characteristic that describes a sample is called a statistic, and a characteristic that describes a population is called a parameter. Although sample statistics are usually representative of corresponding population parameters, there is typically some discrepancy between a statistic and a parameter. The naturally occurring difference between a statistic and a parameter is called sampling error.
4. Statistical methods can be classified into two broad categories: descriptive statistics, which organize and summarize data, and inferential statistics, which use sample data to draw inferences about populations.
5. The correlational method examines relationships between variables by measuring two different variables for each individual. This method allows researchers to measure and describe relationships, but cannot produce a cause-and-effect explanation for the relationship.
6. The experimental method examines relationships between variables by manipulating an independent variable to create different treatment conditions and then measuring a dependent variable to obtain a group of scores in each condition. The groups of scores are then compared. A systematic difference between groups provides evidence that changing the independent variable from one condition to another

also caused a change in the dependent variable. All other variables are controlled to prevent them from influencing the relationship. The intent of the experimental method is to demonstrate a cause-and-effect relationship between variables.

7. Nonexperimental studies also examine relationships between variables by comparing groups of scores, but they do not have the rigor of true experiments and cannot produce cause-and-effect explanations. Instead of manipulating a variable to create different groups, a nonexperimental study uses a preexisting participant characteristic (such as male/female) or the passage of time (before/after) to create the groups being compared.

8. A measurement scale consists of a set of categories that are used to classify individuals. A nominal scale consists of categories that differ only in name and are not differentiated in terms of magnitude or direction. In an ordinal scale, the categories are differentiated in terms of direction, forming an ordered series. An interval scale consists of an ordered series of categories that are all equal-sized intervals. With an interval scale, it is possible to differentiate direction and magnitude (or distance) between categories. Finally, a ratio scale is an interval scale for which the zero point indicates none of the variable being measured. With a ratio scale, ratios of measurements reflect ratios of magnitude.

9. A discrete variable consists of indivisible categories, often whole numbers that vary in countable steps. A continuous variable consists of categories that are infinitely divisible and each score corresponds to an interval on the scale. The boundaries that separate intervals are called real limits and are located exactly halfway between adjacent scores.

10. The letter X is used to represent scores for a variable. If a second variable is used, Y represents its scores. The letter N is used as the symbol for the number of scores in a population; n is the symbol for a number of scores in a sample.

11. The Greek letter sigma (Σ) is used to stand for summation. Therefore, the expression ΣX is read "the sum of the scores." Summation is a mathematical operation (like addition or multiplication) and must be performed in its proper place in the order of operations; summation occurs after parentheses, exponents, and multiplying/dividing have been completed.

KEY TERMS

statistics (5)
population (5)
sample (6)
variable (6)
data (7)
data set (7)
datum (7)
raw score (7)
parameter (7)
statistic (7)
descriptive statistics (7)
inferential statistics (8)
sampling error (8)
correlational method (13)
experimental method (14)
independent variable (16)
dependent variable (16)
control condition (16)
experimental condition (16)
nonequivalent groups study (17)
pre–post study (18)
quasi-independent variable (18)
construct (20)
operational definition (20)
discrete variable (21)
continuous variable (21)
real limits (22)
upper real limit (22)
lower real limit (22)
nominal scale (23)
ordinal scale (23)
interval scale (25)
ratio scale (25)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find practice quizzes and other learning aids for every chapter in this book on the book companion website, as well as a series of workshops and other resources corresponding to the main topic areas. In the left-hand column are a variety of learning exercises for Chapter 1, including a tutorial quiz. Also in the left-hand column, under

Book Resources, is a link to the workshops. For Chapter 1, there is a workshop that reviews the scales of measurement. To get there, click on the *Workshop* link, then click on *Scales of Measurement*. To find materials for other chapters, begin by selecting the desired chapter at the top of the page. Note that the workshops were not developed specifically for this book but are used by several different books written by different authors. As a result, you may find that some of the notation or terminology is different from that which you learned in this text.

At the end of each chapter we remind you about the Web resources. Again, there is a tutorial quiz for every chapter, and we notify you whenever there is a workshop that is related to the chapter content.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

The Statistical Package for the Social Sciences, known as SPSS, is a computer program that performs most of the statistical calculations that are presented in this book, and is commonly available on college and university computer systems. Appendix D contains a general introduction to SPSS. In the Resource section at the end of each chapter for which SPSS is applicable, there are step-by-step instructions for using SPSS to perform the statistical operations presented in the chapter.

FOCUS ON PROBLEM SOLVING

It may help to simplify summation notation if you observe that the summation sign is always followed by a symbol or symbolic expression—for example, ΣX or $\Sigma(X + 3)$. This symbol specifies which values you are to add. If you use the symbol as a column heading and list all the appropriate values in the column, your task is simply to add up the numbers in the column. To find $\Sigma(X + 3)$ for example, start a column headed with $(X + 3)$ next to the column of Xs. List all the $(X + 3)$ values; then find the total for the column.

Often, summation notation is part of a relatively complex mathematical expression that requires several steps of calculation. The series of steps must be performed according to the order of mathematical operations (see page 27). The best procedure is to use a computational table that begins with the original X values listed in the first column. Except for summation, each step in the calculation creates a new column of values. For example, computing $\Sigma(X + 1)^2$ involves three steps and produces a computational table with three columns. The final step is to add the values in the third column (see Example 1.4).

DEMONSTRATION 1.1

SUMMATION NOTATION

A set of scores consists of the following values:

7 3 9 5 4

For these scores, compute each of the following:

ΣX
$(\Sigma X)^2$
ΣX^2
$\Sigma X + 5$
$\Sigma(X - 2)$

Compute ΣX To compute ΣX, we simply add all of the scores in the group.

$$\Sigma X = 7 + 3 + 9 + 5 + 4 = 28$$

Compute $(\Sigma X)^2$ The first step, inside the parentheses, is to compute ΣX. The second step is to square the value for ΣX.

$$\Sigma X = 28 \text{ and } (\Sigma X)^2 = (28)^2 = 784$$

X	X^2
7	49
3	9
9	81
5	25
4	16

Compute ΣX^2 The first step is to square each score. The second step is to add the squared scores. The computational table shows the scores and squared scores. To compute ΣX^2 we add the values in the X^2 column.

$$\Sigma X^2 = 49 + 9 + 81 + 25 + 16 = 180$$

Compute $\Sigma X + 5$ The first step is to compute ΣX. The second step is to add 5 points to the total.

$$\Sigma X = 28 \text{ and } \Sigma X + 5 = 28 + 5 = 33$$

X	$X - 2$
7	5
3	1
9	7
5	3
4	2

Compute $\Sigma(X - 2)$ The first step, inside parentheses, is to subtract 2 points from each score. The second step is to add the resulting values. The computational table shows the scores and the $(X - 2)$ values. To compute $\Sigma(X - 2)$, add the values in the $(X - 2)$ column

$$\Sigma(X - 2) = 5 + 1 + 7 + 3 + 2 = 18$$

PROBLEMS

*1. A researcher is investigating the effectiveness of a treatment for adolescent boys who are taking medication for depression. A group of 30 boys is selected and half receive the new treatment in addition to their medication and the other half continue to take their medication without any treatment. For this study,
   a. Identify the population.
   b. Identify the sample.

2. Define the terms parameter and statistic. Be sure that the concepts of population and sample are included in your definitions.

3. Statistical methods are classified into two major categories: descriptive and inferential. Describe the general purpose for the statistical methods in each category.

4. A researcher plans to compare two treatment conditions by measuring one sample in treatment 1 and a second sample in treatment 2. The researcher then compares the scores for the two treatments and finds a difference between the two groups.
   a. Briefly explain how the difference may have been caused by the treatments.
   b. Briefly explain how the difference simply may be sampling error.

5. Describe the data for a correlational research study. Explain how these data are different from the data obtained in experimental and nonexperimental studies, which also evaluate relationships between two variables.

6. Describe how the goal of an experimental research study is different from the goal for nonexperimental or correlational research. Identify the two elements that are necessary for an experiment to achieve its goal.

7. Strack, Martin, and Stepper (1988) found that people rated cartoons as funnier when holding a pen in their teeth (which forced them to smile) than when holding a pen in their lips (which forced them to frown). For this study, identify the independent variable and the dependent variable.

8. Judge and Cable (2010) found that thin women had higher incomes than heavier women. Is this an example of an experimental or a nonexperimental study?

9. Two researchers are both interested in the relationship between caffeine consumption and activity level for elementary school children. Each obtains a sample of $n = 20$ children.
   a. The first researcher interviews each child to determine the level of caffeine consumption. The researcher then records the level of activity for each child during a 30-minute session on the playground. Is this an experimental or a nonexperimental study? Explain your answer.
   b. The second researcher separates the children into two roughly equivalent groups. The children in one group are given a drink containing 300 mg of caffeine and the other group gets a drink with no caffeine. The researcher then records the level of activity for each child during a 30-minute session on the playground. Is this an experimental or a nonexperimental study? Explain your answer.

10. A researcher would like to evaluate the claim that large doses of vitamin C can help prevent the common cold. One group of participants is given a large dose of the vitamin (500 mg per day), and a second group is given a placebo (sugar pill). The researcher records the number of colds each individual experiences during the 3-month winter season.
   a. Identify the dependent variable for this study.
   b. Is the dependent variable discrete or continuous?
   c. What scale of measurement (nominal, ordinal, interval, or ratio) is used to measure the dependent variable?

11. A research study comparing alcohol use for college students in the United States and Canada reports that more Canadian students drink but American students drink more (Kuo, Adlaf, Lee, Gliksman, Demers, and Wechsler, 2002). Is this study an example of an experiment? Explain why or why not.

12. Oxytocin is a naturally occurring brain chemical that is nicknamed the "love hormone" because it seems to play a role in the formation of social relationships such as mating pairs and parent-child bonding. A recent study demonstrated that oxytocin appears to increase people's tendency to trust others (Kosfeld, Heinrichs, Zak, Fischbacher, and Fehr, 2005). Using an investment game, the study demonstrated that people who inhaled oxytocin were more likely to give their money to a trustee compared to people who inhaled an inactive placebo. For this experimental study, identify the independent variable and the dependent variable.

13. For each of the following, determine whether the variable being measured is discrete or continuous and explain your answer.
   a. Social networking (number of daily minutes on Facebook)
   b. Family size (number of siblings)

*Solutions for odd-numbered problems are provided in Appendix C.

c. Preference between digital or analog watch
d. Number of correct answers on a statistics quiz

14. Four scales of measurement were introduced in this chapter: nominal, ordinal, interval, and ratio.
a. What additional information is obtained from measurements on an ordinal scale compared to measurements on a nominal scale?
b. What additional information is obtained from measurements on an interval scale compared to measurements on an ordinal scale?
c. What additional information is obtained from measurements on a ratio scale compared to measurements on an interval scale?

15. In an experiment examining the effects of humor on memory, Schmidt (1994) showed participants a list of sentences, half of which were humorous and half were nonhumorous. The participants consistently recalled more of the humorous sentences than the nonhumorous sentences.
a. Identify the independent variable for this study.
b. What scale of measurement is used for the independent variable?
c. Identify the dependent variable for this study.
d. What scale of measurement is used for the dependent variable?

16. Explain why *shyness* is a hypothetical construct instead of a concrete variable. Describe how shyness might be measured and defined using an operational definition.

17. Ford and Torok (2008) found that motivational signs were effective in increasing physical activity on a college campus. Signs such as "Step up to a healthier lifestyle" and "An average person burns 10 calories a minute walking up the stairs" were posted by the elevators and stairs in a college building. Students and faculty increased their use of the stairs during times that the signs were posted compared to times when there were no signs.
a. Identify the independent and dependent variables for this study.
b. What scale of measurement is used for the independent variable?

18. For the following scores, find the value of each expression:
a. ΣX
b. ΣX^2
c. $(\Sigma X)^2$
d. $\Sigma(X - 1)$

X
4
2
1
5

19. For the following set of scores, find the value of each expression:
a. ΣX
b. ΣX^2
c. $\Sigma(X + 1)$
d. $\Sigma(X + 1)^2$

X
4
6
0
3
2

20. For the following set of scores, find the value of each expression:
a. ΣX
b. ΣX^2
c. $\Sigma(X + 4)$

X
−4
−2
0
−1
−1

21. Two scores, X and Y, are recorded for each of $n = 4$ subjects. For these scores, find the value of each expression.
a. ΣX
b. ΣY
c. ΣXY

Subject	X	Y
A	6	4
B	0	10
C	3	8
D	2	3

22. Use summation notation to express each of the following calculations:
a. Add 1 point to each score, then add the resulting values.
b. Add 1 point to each score and square the result, then add the squared values.
c. Add the scores and square the sum, then subtract 3 points from the squared value.

23. For the following set of scores, find the value of each expression:
a. ΣX^2
b. $(\Sigma X)^2$
c. $\Sigma(X - 2)$
d. $\Sigma(X - 2)^2$

X
1
0
5
2

CHAPTER

2

Frequency Distributions

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Proportions (math review, Appendix A)
  - Fractions
  - Decimals
  - Percentages
- Scales of measurement (Chapter 1): Nominal, ordinal, interval, and ratio
- Continuous and discrete variables (Chapter 1)
- Real limits (Chapter 1)

Preview

If at first you don't succeed, you are probably not related to the boss.

Did we make you chuckle or, at least, smile a little? The use of humor is a common technique to capture attention and to communicate ideas. Advertisers, for example, often try to make a commercial funny so that people notice it and, perhaps, remember the product. After-dinner speakers always put a few jokes into the speech in an effort to maintain the audience's interest. Although humor seems to capture our attention, does it actually affect our memory?

In an attempt to answer this question, Stephen Schmidt (1994) conducted a series of experiments examining the effects of humor on memory for sentences. Humorous sentences were collected from a variety of sources and then a nonhumorous version was constructed for each sentence. For example, the nonhumorous version of our opening sentence was:

People who are related to the boss often succeed the very first time.

Participants were then presented with a list containing half humorous and half nonhumorous sentences. Later, each person was asked to recall as many sentences as possible. The researcher measured the number of humorous sentences and the number of nonhumorous sentences recalled by each participant. Data similar to the results obtained by Schmidt are shown in Table 2.1.

TABLE 2.1

Memory scores for a sample of 16 participants. The scores represent the number of sentences recalled from each category.

Humorous Sentences				Nonhumorous Sentences			
4	5	2	4	5	2	4	2
6	7	6	6	2	3	1	6
2	5	4	3	3	2	3	3
1	3	5	5	4	1	5	3

The Problem: It is difficult to see any clear pattern simply by looking at the list of numbers. Can you tell whether the memory scores for one type of sentence are generally higher than those for the other type?

The Solution: A frequency distribution provides an overview of the entire group of scores making it easy to see the general level of performance for each type of sentence. For example, the same memory scores that are shown in Table 2.1 have been organized in a frequency distribution graph in Figure 2.1. In the figure, each individual is represented by a block that is placed above that individual's score. The resulting pile of blocks shows a picture of how the individual scores are distributed. For this example, it is now easy to see that the scores for the humorous sentences are generally higher than the scores for the nonhumorous sentences; on average, participants recalled around 5 humorous sentences but only about 3 of the nonhumorous sentences.

In this chapter we present techniques for organizing data into tables and graphs so that an entire set of scores can be presented in a relatively simple display or illustration.

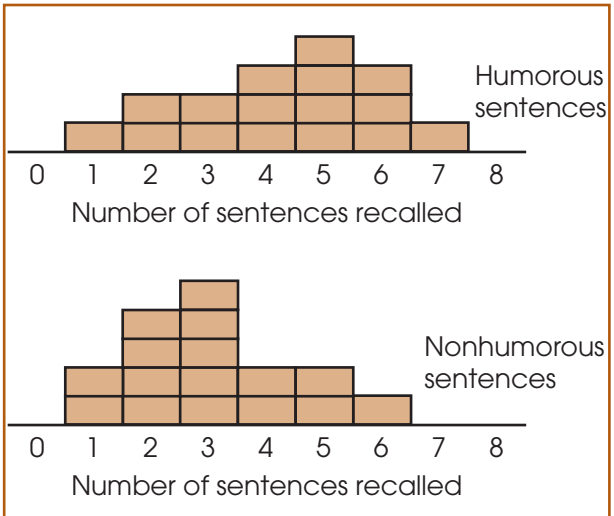


FIGURE 2.1

Hypothetical data showing the number of humorous sentences and the number of nonhumorous sentences recalled by participants in a memory experiment.

2.1 INTRODUCTION TO FREQUENCY DISTRIBUTIONS

The results from a research study usually consist of pages of numbers corresponding to the measurements, or scores, collected during the study. The immediate problem for the researcher is to organize the scores into some comprehensible form so that any patterns in the data can be seen easily and communicated to others. This is the job of descriptive statistics: to simplify the organization and presentation of data. One of the most common procedures for organizing a set of data is to place the scores in a *frequency distribution.*

DEFINITION

A **frequency distribution** is an organized tabulation of the number of individuals located in each category on the scale of measurement.

A frequency distribution takes a disorganized set of scores and places them in order from highest to lowest, grouping together individuals who all have the same score. If the highest score is $X = 10$, for example, the frequency distribution groups together all the 10s, then all the 9s, then the 8s, and so on. Thus, a frequency distribution allows the researcher to see "at a glance" the entire set of scores. It shows whether the scores are generally high or low, whether they are concentrated in one area or spread out across the entire scale, and generally provides an organized picture of the data. In addition to providing a picture of the entire set of scores, a frequency distribution allows you to see the location of any individual score relative to all of the other scores in the set.

A frequency distribution can be structured either as a table or as a graph, but in either case, the distribution presents the same two elements:

1. The set of categories that make up the original measurement scale.
2. A record of the frequency, or number of individuals in each category.

Thus, a frequency distribution presents a picture of how the individual scores are distributed on the measurement scale—hence the name *frequency distribution*.

2.2 FREQUENCY DISTRIBUTION TABLES

It is customary to list categories from highest to lowest, but this is an arbitrary arrangement. Many computer programs list categories from lowest to highest.

The simplest frequency distribution table presents the measurement scale by listing the different measurement categories (X values) in a column from highest to lowest. Beside each X value, we indicate the frequency, or the number of times that particular measurement occurred in the data. It is customary to use an X as the column heading for the scores and an f as the column heading for the frequencies. An example of a frequency distribution table follows.

EXAMPLE 2.1

The following set of $N = 20$ scores was obtained from a 10-point statistics quiz. We organize these scores by constructing a frequency distribution table. Scores:

8, 9, 8, 7, 10, 9, 6, 4, 9, 8,
7, 8, 10, 9, 8, 6, 9, 7, 8, 8

1. The highest score is $X = 10$, and the lowest score is $X = 4$. Therefore, the first column of the table lists the categories that make up the scale of measurement

X	f
10	2
9	5
8	7
7	3
6	2
5	0
4	1

(X values) from 10 down to 4. Notice that all of the possible values are listed in the table. For example, no one had a score of $X = 5$, but this value is included. With an ordinal, interval, or ratio scale, the categories are listed in order (usually highest to lowest). For a nominal scale, the categories can be listed in any order.

2. The frequency associated with each score is recorded in the second column. For example, two people had scores of $X = 10$, so there is a 2 in the f column beside $X = 10$.

Because the table organizes the scores, it is possible to see the general quiz results very quickly. For example, there were only two perfect scores, but most of the class had high grades (8s and 9s). With one exception (the score of $X = 4$), it appears that the class has learned the material fairly well.

Notice that the X values in a frequency distribution table represent the scale of measurement, *not* the actual set of scores. For example, the X column lists the value 10 only one time, but the frequency column indicates that there are actually two values of $X = 10$. Also, the X column lists a value of $X = 5$, but the frequency column indicates that no one actually had a score of $X = 5$.

You also should notice that the frequencies can be used to find the total number of scores in the distribution. By adding up the frequencies, you obtain the total number of individuals:

$$\Sigma f = N$$

OBTAINING ΣX FROM A FREQUENCY DISTRIBUTION TABLE

There may be times when you need to compute the sum of the scores, ΣX, or perform other computations for a set of scores that has been organized into a frequency distribution table. To complete these calculations correctly, you must use all the information presented in the table. That is, it is essential to use the information in the f column as well as the X column to obtain the full set of scores.

When it is necessary to perform calculations for scores that have been organized into a frequency distribution table, the safest procedure is to take the individual scores out of the table before you begin any computations. This process is demonstrated in the following example.

EXAMPLE 2.2

X	f
5	1
4	2
3	3
2	3
1	1

Consider the frequency distribution table shown in the margin. The table shows that the distribution has one 5, two 4s, three 3s, three 2s, and one 1, for a total of 10 scores. If you simply list all 10 scores, you can safely proceed with calculations such as finding ΣX or ΣX^2. For example, to compute ΣX you must add all 10 scores:

$$\Sigma X = 5 + 4 + 4 + 3 + 3 + 3 + 2 + 2 + 2 + 1$$

For the distribution in this table, you should obtain $\Sigma X = 29$. Try it yourself. Similarly, to compute ΣX^2 you square each of the 10 scores and then add the squared values.

$$\Sigma X^2 = 5^2 + 4^2 + 4^2 + 3^2 + 3^2 + 3^2 + 2^2 + 2^2 + 2^2 + 1^2$$

This time you should obtain $\Sigma X^2 = 97$.

An alternative way to get ΣX from a frequency distribution table is to multiply each X value by its frequency and then add these products. This sum may be

expressed in symbols as ΣfX. The computation is summarized as follows for the data in Example 2.2:

Caution: Doing calculations within the table works well for ΣX but can lead to errors for more complex formulas.

X	f	fX	
5	1	5	(the one 5 totals 5)
4	2	8	(the two 4s total 8)
3	3	9	(the three 3s total 9)
2	3	6	(the three 2s total 6)
1	1	1	(the one 1 totals 1)
		$\Sigma X = 29$	

No matter which method you use to find ΣX, the important point is that you must use the information given in the frequency column in addition to the information in the X column.

PROPORTIONS AND PERCENTAGES

In addition to the two basic columns of a frequency distribution, there are other measures that describe the distribution of scores and can be incorporated into the table. The two most common are proportion and percentage.

Proportion measures the fraction of the total group that is associated with each score. In Example 2.2, there were two individuals with $X = 4$. Thus, 2 out of 10 people had $X = 4$, so the proportion would be $\frac{2}{10} = 0.20$. In general, the proportion associated with each score is

$$\text{proportion} = p = \frac{f}{N}$$

Because proportions describe the frequency (f) in relation to the total number (N), they often are called *relative frequencies*. Although proportions can be expressed as fractions (for example, $\frac{2}{10}$), they more commonly appear as decimals. A column of proportions, headed with a p, can be added to the basic frequency distribution table (see Example 2.3).

In addition to using frequencies (f) and proportions (p), researchers often describe a distribution of scores with percentages. For example, an instructor might describe the results of an exam by saying that 15% of the class earned *A*s, 23% earned *B*s, and so on. To compute the percentage associated with each score, you first find the proportion (p) and then multiply by 100:

$$\text{percentage} = p(100) = \frac{f}{N}(100)$$

Percentages can be included in a frequency distribution table by adding a column headed with % (see Example 2.3).

EXAMPLE 2.3

The frequency distribution table from Example 2.2 is repeated here. This time we have added columns showing the proportion (p) and the percentage (%) associated with each score.

X	f	$p = f/N$	% = p(100)
5	1	1/10 = 0.10	10%
4	2	2/10 = 0.20	20%
3	3	3/10 = 0.30	30%
2	3	3/10 = 0.30	30%
1	1	1/10 = 0.10	10%

LEARNING CHECK

1. Construct a frequency distribution table for the following set of scores.

Scores: 3, 2, 3, 2, 4, 1, 3, 3, 5

2. Find each of the following values for the sample in the following frequency distribution table.

a. n
b. ΣX
c. ΣX^2

X	f
5	1
4	2
3	2
2	4
1	1

ANSWERS

1.

X	f
5	1
4	1
3	4
2	2
1	1

2. a. $n = 10$ **b.** $\Sigma X = 28$ **c.** $\Sigma X^2 = 92$ (square then add all 10 scores)

GROUPED FREQUENCY DISTRIBUTION TABLES

When the scores are whole numbers, the total number of rows for a regular table can be obtained by finding the difference between the highest and the lowest scores and adding 1:

rows = highest – lowest + 1

When a set of data covers a wide range of values, it is unreasonable to list all the individual scores in a frequency distribution table. Consider, for example, a set of exam scores that range from a low of $X = 41$ to a high of $X = 96$. These scores cover a *range* of more than 50 points.

If we were to list all of the individual scores from $X = 96$ down to $X = 41$, it would take 56 rows to complete the frequency distribution table. Although this would organize the data, the table would be long and cumbersome. Remember: The purpose for constructing a table is to obtain a relatively simple, organized picture of the data. This can be accomplished by grouping the scores into intervals and then listing the intervals in the table instead of listing each individual score. For example, we could construct a table showing the number of students who had scores in the 90s, the number with scores in the 80s, and so on. The result is called a *grouped frequency distribution table* because we are presenting groups of scores rather than individual values. The groups, or intervals, are called *class intervals*.

There are several guidelines that help guide you in the construction of a grouped frequency distribution table. Note that these are simply guidelines, rather than absolute requirements, but they do help produce a simple, well-organized, and easily understood table.

GUIDELINE 1

The grouped frequency distribution table should have about 10 class intervals. If a table has many more than 10 intervals, it becomes cumbersome and defeats the purpose of a frequency distribution table. On the other hand, if you have too few intervals, you begin to lose information about the distribution of the scores. At the extreme, with only one interval, the table would not tell you anything about how the scores are distributed. Remember that the purpose of a frequency distribution is to help a researcher see the data. With too few or too many intervals, the table will not provide a clear picture. You

should note that 10 intervals is a general guide. If you are constructing a table on a blackboard, for example, you probably want only 5 or 6 intervals. If the table is to be printed in a scientific report, you may want 12 or 15 intervals. In each case, your goal is to present a table that is relatively easy to see and understand.

GUIDELINE 2 The width of each interval should be a relatively simple number. For example, 2, 5, 10, or 20 would be a good choice for the interval width. Notice that it is easy to count by 5s or 10s. These numbers are easy to understand and make it possible for someone to see quickly how you have divided the range of scores.

GUIDELINE 3 The bottom score in each class interval should be a multiple of the width. If you are using a width of 10 points, for example, the intervals should start with 10, 20, 30, 40, and so on. Again, this makes it easier for someone to understand how the table has been constructed.

GUIDELINE 4 All intervals should be the same width. They should cover the range of scores completely with no gaps and no overlaps, so that any particular score belongs in exactly one interval.

The application of these rules is demonstrated in Example 2.4.

EXAMPLE 2.4

An instructor has obtained the set of $N = 25$ exam scores shown here. To help organize these scores, we will place them in a frequency distribution table. The scores are:

82, 75, 88, 93, 53, 84, 87, 58, 72, 94, 69, 84, 61,
91, 64, 87, 84, 70, 76, 89, 75, 80, 73, 78, 60

Remember, when the scores are whole numbers, the number of rows is determined by

highest – lowest + 1

The first step is to determine the range of scores. For these data, the smallest score is $X = 53$ and the largest score is $X = 94$, so a total of 42 rows would be needed for a table that lists each individual score. Because 42 rows would not provide a simple table, we have to group the scores into class intervals.

The best method for finding a good interval width is a systematic trial-and-error approach that uses guidelines 1 and 2 simultaneously. Specifically, we want about 10 intervals and we want the interval width to be a simple number. For this example, the scores cover a range of 42 points, so we will try several different interval widths to see how many intervals are needed to cover this range. For example, if each interval is 2 points wide, it would take 21 intervals to cover a range of 42 points. This is too many, so we move on to an interval width of 5 or 10 points. The following table shows how many intervals would be needed for these possible widths:

Because the bottom interval usually extends below the lowest score and the top interval extends beyond the highest score, you often need slightly more than the computed number of intervals.

Width	Number of Intervals Needed to Cover a Range of 42 Points	
2	21	(too many)
5	9	(OK)
10	5	(too few)

Notice that an interval width of 5 will result in about 10 intervals, which is exactly what we want.

The next step is to actually identify the intervals. The lowest score for these data is $X = 53$, so the lowest interval should contain this value. Because the interval should have a multiple of 5 as its bottom score, the interval should begin at 50. The

interval has a width of 5, so it should contain 5 values: 50, 51, 52, 53, and 54. Thus, the bottom interval is 50–54. The next interval would start at 55 and go to 59. Note that this interval also has a bottom score that is a multiple of 5, and contains exactly 5 scores (55, 56, 57, 58, and 59). The complete frequency distribution table showing all of the class intervals is presented in Table 2.2.

Once the class intervals are listed, you complete the table by adding a column of frequencies. The values in the frequency column indicate the number of individuals who have scores located in that class interval. For this example, there were three students with scores in the 60–64 interval, so the frequency for this class interval is $f = 3$ (see Table 2.2). The basic table can be extended by adding columns showing the proportion and percentage associated with each class interval.

Finally, you should note that after the scores have been placed in a grouped table, you lose information about the specific value for any individual score. For example, Table 2.2 shows that one person had a score between 65 and 69, but the table does not identify the exact value for the score. In general, the wider the class intervals are, the more information is lost. In Table 2.2 the interval width is 5 points, and the table shows that there are three people with scores in the lower 60s and one person with a score in the upper 60s. This information would be lost if the interval width were increased to 10 points. With an interval width of 10, all of the 60s would be grouped together into one interval labeled 60–69. The table would show a frequency of four people in the 60–69 interval, but it would not tell whether the scores were in the upper 60s or the lower 60s.

REAL LIMITS AND FREQUENCY DISTRIBUTIONS

Recall from Chapter 1 that a continuous variable has an infinite number of possible values and can be represented by a number line that is continuous and contains an infinite number of points. However, when a continuous variable is measured, the resulting measurements correspond to *intervals* on the number line rather than single points. If you are measuring time in seconds, for example, a score of $X = 8$ seconds actually represents an interval bounded by the real limits 7.5 seconds and 8.5 seconds. Thus, a frequency distribution table showing a frequency of $f = 3$ individuals all assigned a score of $X = 8$ does not mean that all three individuals had exactly the same measurement. Instead, you should realize that the three measurements are simply located in the same interval between 7.5 and 8.5.

The concept of real limits also applies to the class intervals of a grouped frequency distribution table. For example, a class interval of 40–49 contains scores from $X = 40$ to $X = 49$. These values are called the *apparent limits* of the interval because it appears

TABLE 2.2

This grouped frequency distribution table shows the data from Example 2.4. The original scores range from a high of $X = 94$ to a low of $X = 53$. This range has been divided into 9 intervals with each interval exactly 5 points wide. The frequency column (f) lists the number of individuals with scores in each of the class intervals.

X	f
90–94	3
85–89	4
80–84	5
75–79	4
70–74	3
65–69	1
60–64	3
55–59	1
50–54	1

that they form the upper and lower boundaries for the class interval. If you are measuring a continuous variable, however, a score of $X = 40$ is actually an interval from 39.5 to 40.5. Similarly, $X = 49$ is an interval from 48.5 to 49.5. Therefore, the real limits of the interval are 39.5 (the lower real limit) and 49.5 (the upper real limit). Notice that the next higher class interval is 50–59, which has a lower real limit of 49.5. Thus, the two intervals meet at the real limit 49.5, so there are no gaps in the scale. You also should notice that the width of each class interval becomes easier to understand when you consider the real limits of an interval. For example, the interval 50–59 has real limits of 49.5 and 59.5. The distance between these two real limits (10 points) is the width of the interval.

LEARNING CHECK

1. For each of the following situations, determine what interval width is most appropriate for a grouped frequency distribution and identify the apparent limits of the bottom interval.
   - **a.** Scores range from $X = 7$ to $X = 21$.
   - **b.** Scores range from $X = 52$ to $X = 98$.
   - **c.** Scores range from $X = 16$ to $X = 93$.
2. Using only the frequency distribution table presented in Table 2.2, how many individuals had a score of $X = 73$?

ANSWERS

1. **a.** A width of 2 points would require 8 intervals. Bottom interval is 6−7.
   b. A width of 5 points would require 10 intervals. Bottom interval is 50−54.
   c. A width of 10 points would require 9 intervals. Bottom interval is 10−19.
3. After a set of scores has been summarized in a grouped table, you cannot determine the frequency for any specific score. There is no way to determine how many individuals had $X = 73$ from the table alone. (You can say that *at most* three people had $X = 73$.)

2.3 FREQUENCY DISTRIBUTION GRAPHS

A frequency distribution graph is basically a picture of the information available in a frequency distribution table. We consider several different types of graphs, but all start with two perpendicular lines called *axes*. The horizontal line is the *X*-axis, or the abscissa (ab-SIS-uh). The vertical line is the *Y*-axis, or the ordinate. The measurement scale (set of *X* values) is listed along the *X*-axis with values increasing from left to right. The frequencies are listed on the *Y*-axis with values increasing from bottom to top. As a general rule, the point where the two axes intersect should have a value of zero for both the scores and the frequencies. A final general rule is that the graph should be constructed so that its height (*Y*-axis) is approximately two-thirds to three-quarters of its length (*X*-axis). Violating these guidelines can result in graphs that give a misleading picture of the data (see Box 2.1).

GRAPHS FOR INTERVAL OR RATIO DATA

When the data consist of numerical scores that have been measured on an interval or ratio scale, there are two options for constructing a frequency distribution graph. The two types of graphs are called *histograms* and *polygons.*

Histograms To construct a histogram, you first list the numerical scores (the categories of measurement) along the *X*-axis. Then you draw a bar above each *X* value so that

a. The height of the bar corresponds to the frequency for that category.

b. For continuous variables, the width of the bar extends to the real limits of the category. For discrete variables, each bar extends exactly half the distance to the adjacent category on each side.

For both continuous and discrete variables, each bar in a histogram extends to the midpoint between adjacent categories. As a result, adjacent bars touch and there are no spaces or gaps between bars. An example of a histogram is shown in Figure 2.2.

When data have been grouped into class intervals, you can construct a frequency distribution histogram by drawing a bar above each interval so that the width of the bar extends exactly half the distance to the adjacent category on each side. This process is demonstrated in Figure 2.3.

For the two histograms shown in Figures 2.2 and 2.3, notice that the values on both the vertical and horizontal axes are clearly marked and that both axes are labeled. Also note that, whenever possible, the units of measurement are specified; for example, Figure 2.3 shows a distribution of heights measured in inches. Finally, notice that the horizontal axis in Figure 2.3 does not list all of the possible heights starting from zero and going up to 48 inches. Instead, the graph clearly shows a break between zero and 30, indicating that some scores have been omitted.

A modified histogram A slight modification to the traditional histogram produces a very easy to draw and simple to understand sketch of a frequency distribution. Instead

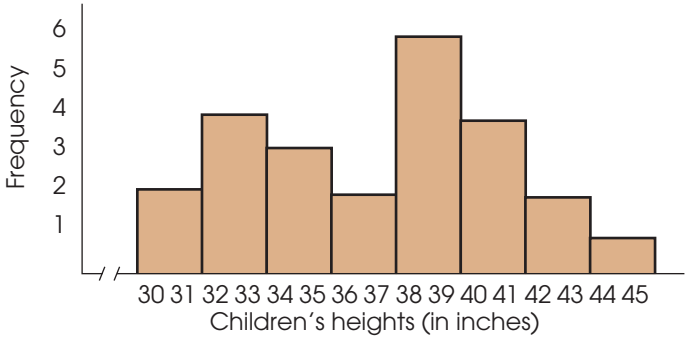
FIGURE 2.2

An example of a frequency distribution histogram. The same set of quiz scores is presented in a frequency distribution table and in a histogram.

X	*f*
5	2
4	3
3	4
2	2
1	1

FIGURE 2.3

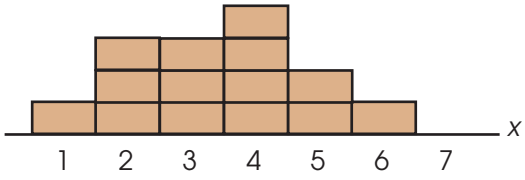
An example of a frequency distribution histogram for grouped data. The same set of children's heights is presented in a frequency distribution table and in a histogram.



X	*f*
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

FIGURE 2.4

A frequency distribution in which each individual is represented by a block placed directly above the individual's score. For example, three people had scores of $X = 2$.



of drawing a bar above each score, the modification consists of drawing a stack of blocks. Each block represents one individual, so the number of blocks above each score corresponds to the frequency for that score. An example is shown in Figure 2.4.

Note that the number of blocks in each stack makes it very easy to see the absolute frequency for each category. In addition, it is easy to see the exact difference in frequency from one category to another. In Figure 2.4, for example, there are exactly two more people with scores of $X = 2$ than with scores of $X = 1$. Because the frequencies are clearly displayed by the number of blocks, this type of display eliminates the need for a vertical line (the *Y*-axis) showing frequencies. In general, this kind of graph provides a simple and concrete picture of the distribution for a sample of scores. Note that we often use this kind of graph to show sample data throughout the rest of the book. You should also note, however, that this kind of display simply provides a sketch of the distribution and is not a substitute for an accurately drawn histogram with two labeled axes.

Polygons The second option for graphing a distribution of numerical scores from an interval or ratio scale of measurement is called a polygon. To construct a polygon, you begin by listing the numerical scores (the categories of measurement) along the *X*-axis. Then,

a. A dot is centered above each score so that the vertical position of the dot corresponds to the frequency for the category.

b. A continuous line is drawn from dot to dot to connect the series of dots.

c. The graph is completed by drawing a line down to the *X*-axis (zero frequency) at each end of the range of scores. The final lines are usually drawn so that they reach the *X*-axis at a point that is one category below the lowest score on the left side and one category above the highest score on the right side. An example of a polygon is shown in Figure 2.5.

A polygon also can be used with data that have been grouped into class intervals. For a grouped distribution, you position each dot directly above the midpoint of the class interval. The midpoint can be found by averaging the highest and the lowest scores in the interval. For example, a class interval that is listed as 20–29 would have a midpoint of 24.5.

$$\text{midpoint} = \frac{20+29}{2} = \frac{49}{2} = 24.5$$

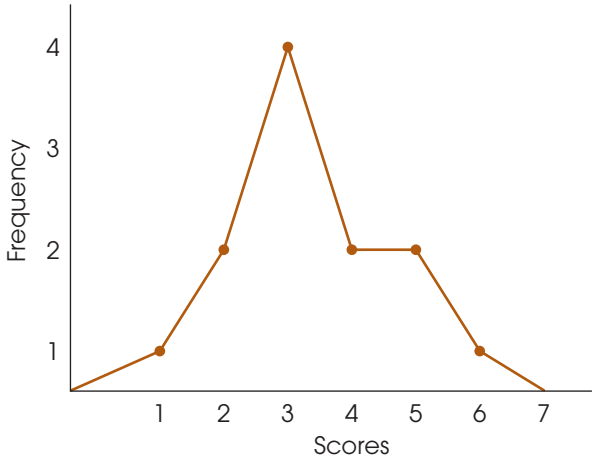
An example of a frequency distribution polygon with grouped data is shown in Figure 2.6.

GRAPHS FOR NOMINAL OR ORDINAL DATA

When the scores are measured on a nominal or ordinal scale (usually non-numerical values), the frequency distribution can be displayed in a *bar graph.*

FIGURE 2.5

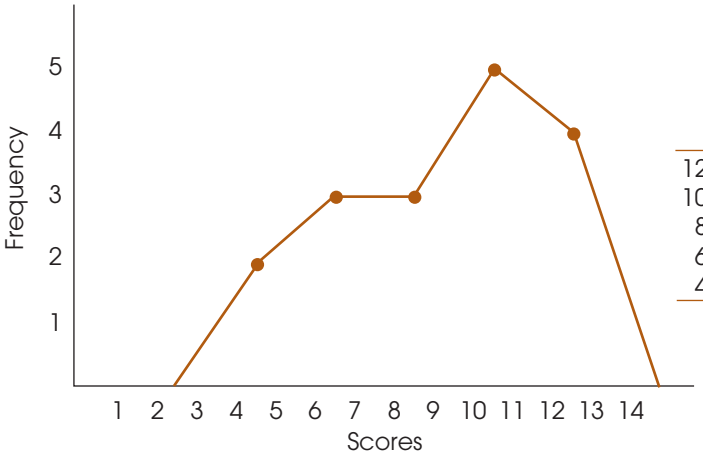
An example of a frequency distribution polygon. The same set of data is presented in a frequency distribution table and in a polygon.



X	*f*
6	1
5	2
4	2
3	4
2	2
1	1

FIGURE 2.6

An example of a frequency distribution polygon for grouped data. The same set of data is presented in a grouped frequency distribution table and in a polygon.



X	*f*
12–13	4
10–11	5
8–9	3
6–7	3
4–5	2

Bar graphs A bar graph is essentially the same as a histogram, except that spaces are left between adjacent bars. For a nominal scale, the space between bars emphasizes that the scale consists of separate, distinct categories. For ordinal scales, separate bars are used because you cannot assume that the categories are all the same size.

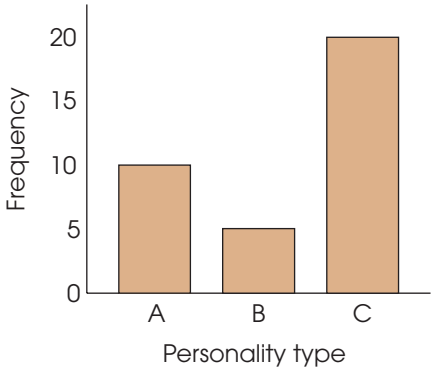
To construct a bar graph, list the categories of measurement along the *X*-axis and then draw a bar above each category so that the height of the bar equals the frequency for the category. An example of a bar graph is shown in Figure 2.7.

GRAPHS FOR POPULATION DISTRIBUTIONS

When you can obtain an exact frequency for each score in a population, you can construct frequency distribution graphs that are exactly the same as the histograms, polygons, and bar graphs that are typically used for samples. For example, if a population is defined as a specific group of $N = 50$ people, we could easily determine how many have IQs of $X = 110$. However, if we are interested in the entire population of adults

FIGURE 2.7

A bar graph showing the distribution of personality types in a sample of college students. Because personality type is a discrete variable measured on a nominal scale, the graph is drawn with space between the bars.



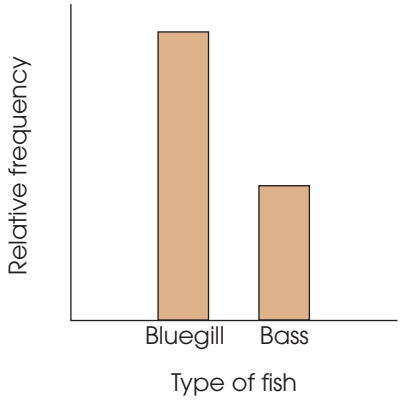
in the United States, it would be impossible to obtain an exact count of the number of people with an IQ of 110. Although it is still possible to construct graphs showing frequency distributions for extremely large populations, the graphs usually involve two special features: relative frequencies and smooth curves.

Relative frequencies Although you usually cannot find the absolute frequency for each score in a population, you very often can obtain *relative frequencies.* For example, you may not know exactly how many fish are in the lake, but after years of fishing you do know that there are twice as many bluegill as there are bass. You can represent these relative frequencies in a bar graph by making the bar above bluegill two times taller than the bar above bass (Figure 2.8). Notice that the graph does not show the absolute number of fish. Instead, it shows the relative number of bluegill and bass.

Smooth curves When a population consists of numerical scores from an interval or a ratio scale, it is customary to draw the distribution with a smooth curve instead of the jagged, step-wise shapes that occur with histograms and polygons. The smooth curve indicates that you are not connecting a series of dots (real frequencies) but instead are showing the relative changes that occur from one score to the next. One commonly occurring population distribution is the normal curve. The word *normal* refers to a specific shape that can be precisely defined by an equation. Less precisely, we can describe

FIGURE 2.8

A frequency distribution showing the relative frequency for two types of fish. Notice that the exact number of fish is not reported; the graph simply says that there are twice as many bluegill as there are bass.



a normal distribution as being symmetrical, with the greatest frequency in the middle and relatively smaller frequencies as you move toward either extreme. A good example of a normal distribution is the population distribution for IQ scores shown in Figure 2.9. Because normal-shaped distributions occur commonly and because this shape is mathematically guaranteed in certain situations, we give it extensive attention throughout this book.

In the future, we will be referring to *distributions of scores*. Whenever the term *distribution* appears, you should conjure up an image of a frequency distribution graph. The graph provides a picture showing exactly where the individual scores are located. To make this concept more concrete, you might find it useful to think of the graph as showing a pile of individuals just like we showed a pile of blocks in Figure 2.4. For the population of IQ scores shown in Figure 2.9, the pile is highest at an IQ score around 100 because most people have average IQs. There are only a few individuals piled up at an IQ of 130; it must be lonely at the top.

2.4 THE SHAPE OF A FREQUENCY DISTRIBUTION

Rather than drawing a complete frequency distribution graph, researchers often simply describe a distribution by listing its characteristics. There are three characteristics that completely describe any distribution: shape, central tendency, and variability. In simple terms, central tendency measures where the center of the distribution is located. Variability tells whether the scores are spread over a wide range or are clustered together. Central tendency and variability will be covered in detail in Chapters 3 and 4. Technically, the shape of a distribution is defined by an equation that prescribes the exact relationship between each X and Y value on the graph. However, we rely on a few less-precise terms that serve to describe the shape of most distributions.

Nearly all distributions can be classified as being either *symmetrical* or *skewed*.

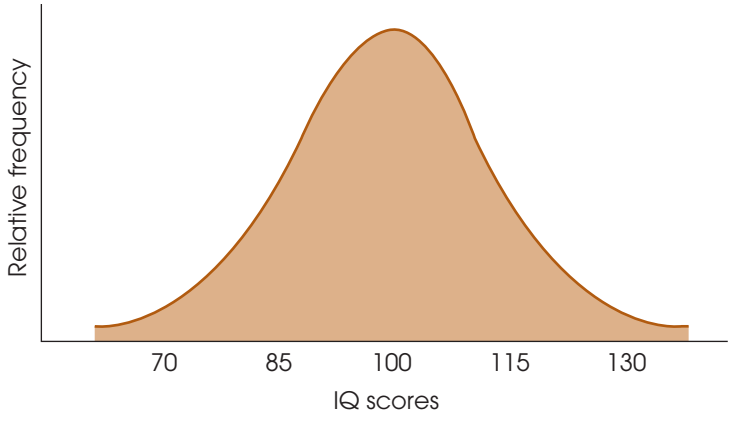
DEFINITIONS

In a **symmetrical distribution**, it is possible to draw a vertical line through the middle so that one side of the distribution is a mirror image of the other (Figure 2.11).

In a **skewed distribution**, the scores tend to pile up toward one end of the scale and taper off gradually at the other end (see Figure 2.11).

FIGURE 2.9

The population distribution of IQ scores: an example of a normal distribution.



BOX 2.1 THE USE AND MISUSE OF GRAPHS

Although graphs are intended to provide an accurate picture of a set of data, they can be used to exaggerate or misrepresent a set of scores. These misrepresentations generally result from failing to follow the basic rules for graph construction. The following example demonstrates how the same set of data can be presented in two entirely different ways by manipulating the structure of a graph.

For the past several years, the city has kept records of the number of homicides. The data are summarized as follows:

Year	Number of Homicides
2007	42
2008	44
2009	47
2010	49

These data are shown in two different graphs in Figure 2.10. In the first graph, we have exaggerated the height and started numbering the *Y*-axis at 40 rather than at zero. As a result, the graph seems to indicate a rapid rise in the number of homicides over the 4-year period. In the second graph, we have stretched out the *X*-axis and used zero as the starting point for the *Y*-axis. The result is a graph that shows little change in the homicide rate over the 4-year period.

Which graph is correct? The answer is that neither one is very good. Remember that the purpose of a graph is to provide an accurate display of the data. The first graph in Figure 2.10 exaggerates the differences between years, and the second graph conceals the differences. Some compromise is needed. Also note that in some cases a graph may not be the best way to display information. For these data, for example, showing the numbers in a table would be better than either graph.

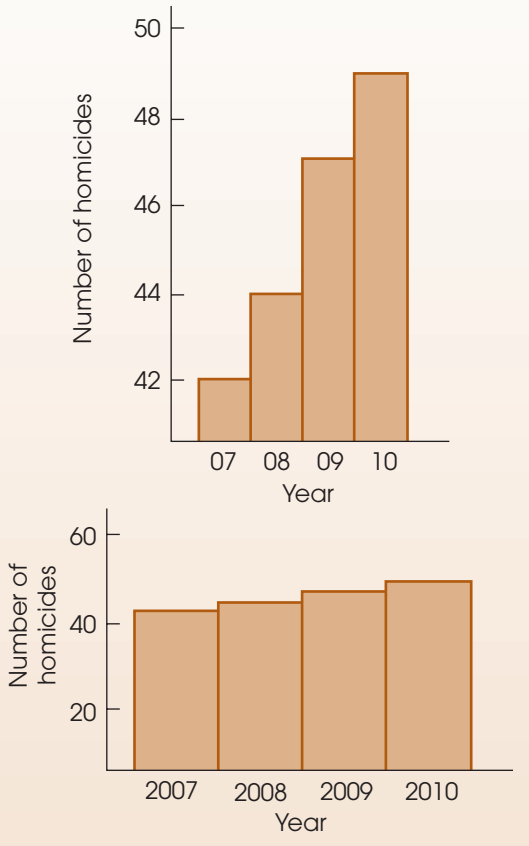


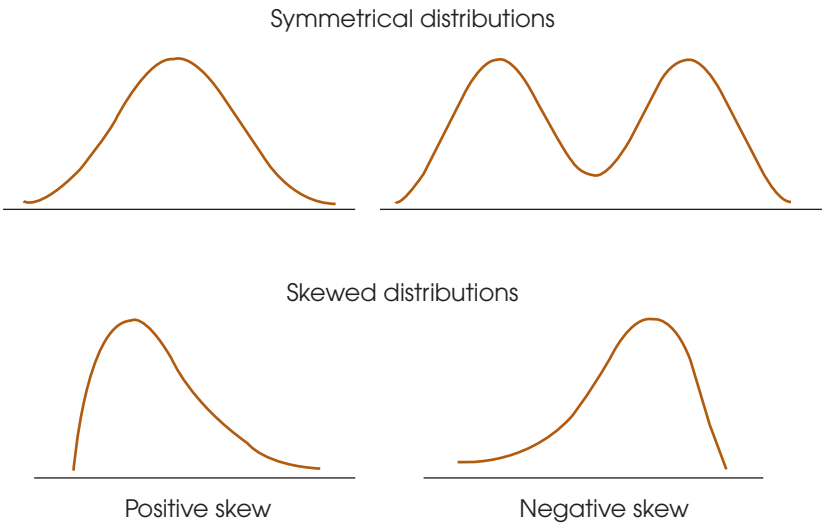
FIGURE 2.10

Two graphs showing the number of homicides in a city over a 4-year period. Both graphs show exactly the same data. However, the first graph gives the appearance that the homicide rate is high and rising rapidly. The second graph gives the impression that homicides rate is low and has not changed over the 4-year period.

The section where the scores taper off toward one end of a distribution is called the **tail** of the distribution.

A skewed distribution with the tail on the right-hand side is **positively skewed** because the tail points toward the positive (above-zero) end of the *X*-axis. If the tail points to the left, the distribution is **negatively skewed** (see Figure 2.11).

FIGURE 2.11
Examples of different shapes for distributions.



For a very difficult exam, most scores tend to be low, with only a few individuals earning high scores. This produces a positively skewed distribution. Similarly, a very easy exam tends to produce a negatively skewed distribution, with most of the students earning high scores and only a few with low values.

LEARNING CHECK

1. Sketch a frequency distribution histogram and a frequency distribution polygon for the data in the following table:

X	*f*
5	4
4	6
3	3
2	1
1	1

2. Describe the shape of the distribution in Exercise 1.

3. A researcher records the gender and academic major for each student at a college basketball game. If the distribution of majors is shown in a frequency distribution graph, what type of graph should be used?

4. If the results from a research study are presented in a frequency distribution histogram, would it also be appropriate to show the same results in a polygon? Explain your answer.

5. A college reports that the youngest registered student is 17 years old, and 20% of the registered students are older than 25. What is the shape of the distribution of ages for registered students?

ANSWERS

1. The graphs are shown in Figure 2.12.
2. The distribution is negatively skewed.
3. A bar graph is used for nominal data.
4. Yes. Histograms and polygons are both used for data from interval or ratio scales.
5. It is positively skewed with most of the distribution around 17–21 and a few scores scattered at 25 and higher.

2.5 PERCENTILES, PERCENTILE RANKS, AND INTERPOLATION

Although the primary purpose of a frequency distribution is to provide a description of an entire set of scores, it also can be used to describe the position of an individual within the set. Individual scores, or X values, are called *raw scores*. By themselves, raw scores do not provide much information. For example, if you are told that your score on an exam is $X = 43$, you cannot tell how well you did relative to other students in the class. To evaluate your score, you need more information, such as the average score or the number of people who had scores above and below you. With this additional information, you would be able to determine your relative position in the class. Because raw scores do not provide much information, it is desirable to transform them into a more meaningful form. One transformation that we consider changes raw scores into *percentiles*.

DEFINITIONS

The **rank** or **percentile rank** of a particular score is defined as the percentage of individuals in the distribution with scores equal to or less than the particular value.

When a score is identified by its percentile rank, the score is called a **percentile**.

Suppose, for example, that you have a score of $X = 43$ on an exam and that you know that exactly 60% of the class had scores of 43 or lower. Then your score $X = 43$ has a percentile rank of 60%, and your score would be called the 60th percentile. Notice that *percentile rank* refers to a percentage and that *percentile* refers to a score. Also notice that your rank or percentile describes your exact position within the distribution.

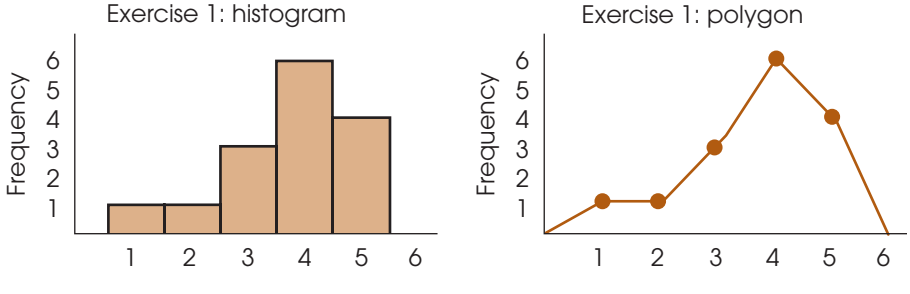


FIGURE 2.12
Answer to the Learning Check Exercise 1.

CUMULATIVE FREQUENCY AND CUMULATIVE PERCENTAGE

To determine percentiles or percentile ranks, the first step is to find the number of individuals who are located at or below each point in the distribution. This can be done most easily with a frequency distribution table by simply counting the number who are in or below each category on the scale. The resulting values are called *cumulative frequencies* because they represent the accumulation of individuals as you move up the scale.

EXAMPLE 2.5

In the following frequency distribution table, we have included a cumulative frequency column headed by *cf*. For each row, the cumulative frequency value is obtained by adding up the frequencies in and below that category. For example, the score $X = 3$ has a cumulative frequency of 14 because exactly 14 individuals had scores of $X = 3$ or less.

X	*f*	*cf*
5	1	20
4	5	19
3	8	14
2	4	6
1	2	2

The cumulative frequencies show the number of individuals located at or below each score. To find percentiles, we must convert these frequencies into percentages. The resulting values are called *cumulative percentages* because they show the percentage of individuals who are accumulated as you move up the scale.

EXAMPLE 2.6

This time we have added a cumulative percentage column (c%) to the frequency distribution table from Example 2.5. The values in this column represent the percentage of individuals who are located in and below each category. For example, 70% of the individuals (14 out of 20) had scores of $X = 3$ or lower. Cumulative percentages can be computed by

$$c\% = \frac{cf}{N}(100\%)$$

X	*f*	*cf*	*c%*
5	1	20	100%
4	5	19	95%
3	8	14	70%
2	4	6	30%
1	2	2	10%

The cumulative percentages in a frequency distribution table give the percentage of individuals with scores at or below each *X* value. However, you must remember that the *X* values in the table are usually measurements of a continuous variable and, therefore, represent intervals on the scale of measurement (see page 22). A score of $X = 2$,

for example, means that the measurement was somewhere between the real limits of 1.5 and 2.5. Thus, when a table shows that a score of $X = 2$ has a cumulative percentage of 30%, you should interpret this as meaning that 30% of the individuals have been accumulated by the time you reach the top of the interval for $X = 2$. Notice that each cumulative percentage value is associated with the upper real limit of its interval. This point is demonstrated in Figure 2.13, which shows the same data that were used in Example 2.6. Figure 2.13 shows that two people, or 10%, had scores of $X = 1$; that is, two people had scores between 0.5 and 1.5. You cannot be sure that both individuals have been accumulated until you reach 1.5, the upper real limit of the interval. Similarly, a cumulative percentage of 30% is reached at 2.5 on the scale, a percentage of 70% is reached at 3.5, and so on.

INTERPOLATION

It is possible to determine some percentiles and percentile ranks directly from a frequency distribution table, provided that the percentiles are upper real limits and the ranks are percentages that appear in the table. Using the table in Example 2.6, for example, you should be able to answer the following questions:

1. What is the 95th percentile? (Answer: $X = 4.5$.)
2. What is the percentile rank for $X = 3.5$? (Answer: 70%.)

However, there are many values that do not appear directly in the table, and it is impossible to determine these values precisely. Referring to the table in Example 2.6 again,

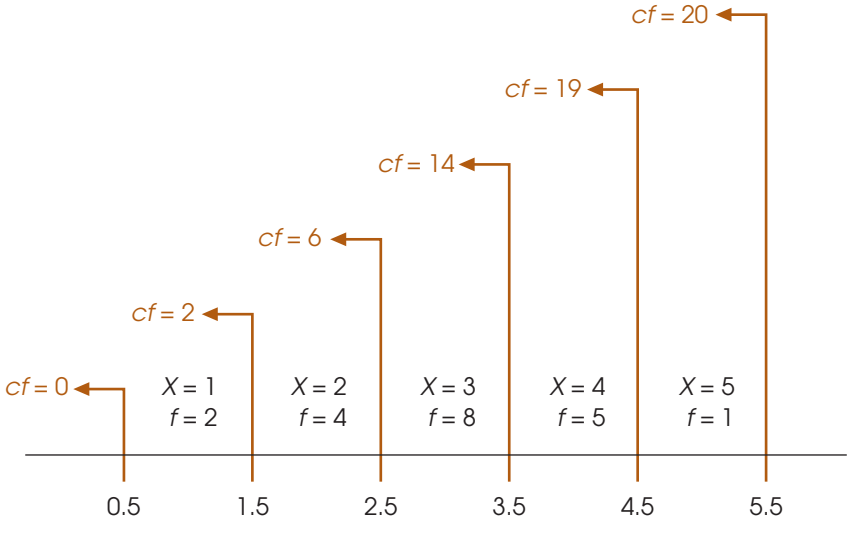
1. What is the 50th percentile?
2. What is the percentile rank for $X = 4$?

Because these values are not specifically reported in the table, you cannot answer the questions. However, it is possible to estimate these intermediate values by using a standard procedure known as *interpolation*.

Before we apply the process of interpolation to percentiles and percentile ranks, we use a simple, commonsense example to introduce this method. Suppose that Bob walks

FIGURE 2.13

The relationship between cumulative frequencies (*cf* values) and upper real limits. Notice that two people have scores of $X = 1$. These two individuals are located between the real limits of 0.5 and 1.5. Although their exact locations are not known, you can be certain that both had scores below the upper limit of 1.



to work each day. The total distance is 2 miles and the trip takes Bob 40 minutes. What is your estimate of how far Bob has walked after 20 minutes? To help, we have created a table showing the time and distance for the start and finish of Bob's trip.

	Time	Distance
Start	0	0
Finish	40	2

If you estimated that Bob walked 1 mile in 20 minutes, you have done interpolation. You probably went through the following logical steps:

1. The total time is 40 minutes.
2. 20 minutes represents half of the total time.
3. Assuming that Bob walks at a steady pace, he should walk half of the total distance in half of the total time.
4. The total distance is 2 miles and half of the total distance is 1.

The process of interpolation is pictured in Figure 2.14. In the figure, the top line shows the time for Bob's walk, from 0 to 40 minutes, and the bottom line shows the time, from 0 to 2 miles. The middle line shows different fractions along the way. Using the figure, try answering the following questions about time and distance.

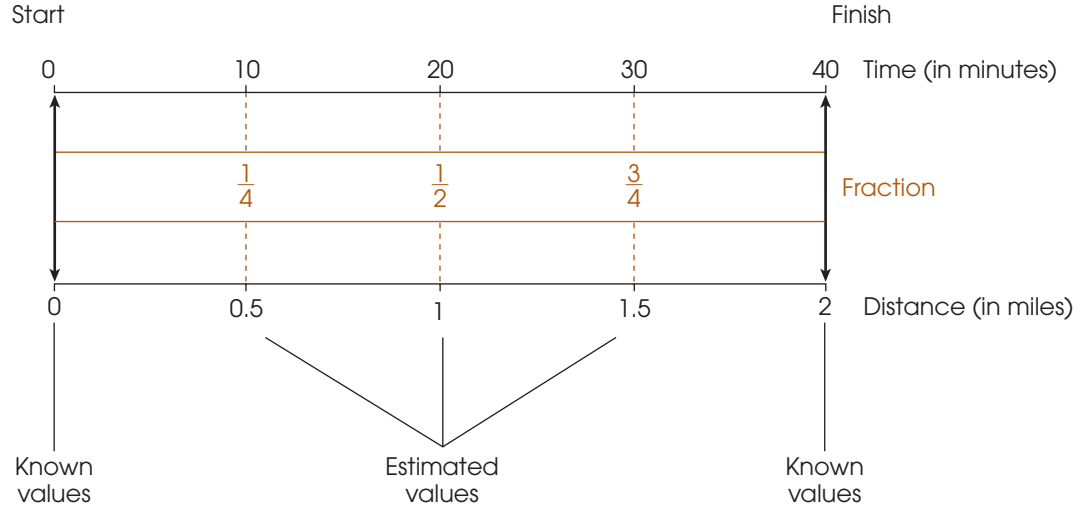


FIGURE 2.14

A graphic representation of the process of interpolation. The same interval is shown on two separate scales, time and distance. Only the endpoints of the scales are known—Bob starts at 0 for both time and distance, and he ends at 40 minutes and 2 miles. Interpolation is used to estimate values within the interval by assuming that fractional portions of one scale correspond to the same fractional portions of the other. For example, it is assumed that halfway through the time scale corresponds to halfway through the distance scale.

1. How much time does it take for Bob to walk 1.5 miles?
2. How far has Bob walked after 10 minutes?

If you got answers of 30 minutes and $\frac{1}{2}$ mile, you have mastered the process of interpolation.

Notice that interpolation provides a method for finding intermediate values—that is, values that are located between two specified numbers. This is exactly the problem we faced with percentiles and percentile ranks. Some values are given in the table, but others are not. Also notice that interpolation only *estimates* the intermediate values. The basic assumption underlying interpolation is that there is a constant rate of change from one end of the interval to the other. In Bob's walking example, we assume that he is walking at a constant rate for the entire trip. Because interpolation is based on this assumption, the values that we calculate are only estimates. The general process of interpolation can be summarized as follows:

1. A single interval is measured on two separate scales (for example, time and distance). The endpoints of the interval are known for each scale.
2. You are given an intermediate value on one of the scales. The problem is to find the corresponding intermediate value on the other scale.
3. The interpolation process requires four steps:
   a. Find the width of the interval on both scales.
   b. Locate the position of the intermediate value in the interval. This position corresponds to a fraction of the whole interval:

$$\text{fraction} = \frac{\text{distance from the top of the interval}}{\text{interval width}}$$

   c. Use the same fraction to determine the corresponding position on the other scale. First, determine the distance from the top of the interval:

$$\text{distance} = (\text{fraction}) \times (\text{width})$$

   d. Use the distance from the top to determine the position on the other scale.

You may notice that in each of these problems we use interpolation working from the *top* of the interval. However, this choice is arbitrary, and you should realize that interpolation can be done just as easily working from the bottom of the interval.

The following examples demonstrate the process of interpolation as it is applied to percentiles and percentile ranks. The key to success in solving these problems is that each cumulative percentage in the table is associated with the upper real limit of its score interval.

EXAMPLE 2.7

Using the following distribution of scores, we will find the percentile rank corresponding to $X = 7.0$:

X	f	cf	$c\%$
10	2	25	100%
9	8	23	92%
8	4	15	60%
7	6	11	44%
6	4	5	20%
5	1	1	4%

Notice that $X = 7.0$ is located in the interval bounded by the real limits of 6.5 and 7.5. The cumulative percentages corresponding to these real limits are 20% and 44%, respectively. These values are shown in the following table:

	Scores (X)	Percentages
Top	7.5	44%
Intermediate value	7.0	?
Bottom	6.5	20%

For interpolation problems, it is always helpful to create a table showing the range on both scales.

STEP 1 For the scores, the width of the interval is 1 point (from 6.5 to 7.5). For the percentages, the width is 24 points (from 20% to 44%).

STEP 2 Our particular score is located 0.5 point from the top of the interval. This is exactly halfway down in the interval.

STEP 3 On the percentage scale, halfway down is

$$\frac{1}{2}\ (24 \text{ points}) = 12 \text{ points}$$

STEP 4 For the percentages, the top of the interval is 44%, so 12 points down would be

$$44\% - 12\% = 32\%$$

This is the answer. A score of $X = 7.0$ corresponds to a percentile rank of 32%

This same interpolation procedure can be used with data that have been grouped into class intervals. Once again, you must remember that the cumulative percentage values are associated with the upper real limits of each interval. The following example demonstrates the calculation of percentiles and percentile ranks using data in a grouped frequency distribution.

EXAMPLE 2.8 Using the following distribution of scores, we can use interpolation to find the 50th percentile:

X	f	cf	$c\%$
20–24	2	20	100%
15–19	3	18	90%
10–14	3	15	75%
5–9	10	12	60%
0–4	2	2	10%

A percentage value of 50% is not given in the table; however, it is located between 10% and 60%, which are given. These two percentage values are associated with the

upper real limits of 4.5 and 9.5, respectively. These values are shown in the following table:

	Scores (X)	Percentages	
Top	9.5	60%	
	?	50%	Intermediate value
Bottom	4.5	10%	

STEP 1 For the scores, the width of the interval is 5 points. For the percentages, the width is 50 points.

STEP 2 The value of 50% is located 10 points from the top of the percentage interval. As a fraction of the whole interval, this is 10 out of 50, or $\frac{1}{5}$ of the total interval.

STEP 3 Using this same fraction for the scores, we obtain a distance of

$$\frac{1}{5}(5 \text{ points}) = 1 \text{ point}$$

The location we want is 1 point down from the top of the score interval.

STEP 4 Because the top of the interval is 9.5, the position we want is

$$9.5 - 1 = 8.5$$

This is the answer. The 50th percentile is $X = 8.5$.

LEARNING CHECK

1. On a statistics exam, would you rather score at the 80th percentile or at the 20th percentile?
2. For the distribution of scores presented in the following table,
   a. Find the 70th percentile.
   b. Find the percentile rank for $X = 9.5$.

X	f	cf	$c\%$
20–24	1	20	100%
15–19	5	19	95%
10–14	8	14	70%
5–9	4	6	20%
0–4	2	2	10%

3. Using the distribution of scores from Exercise 2 and interpolation,
   a. Find the 15th percentile.
   b. Find the percentile rank for $X = 13$.

ANSWERS

1. The 80th percentile is the higher score.
2. **a.** $X = 14.5$ is the 70th percentile. **b.** $X = 9.5$ has a rank of 20%.
3. **a.** Because 15% is between the values of 10% and 20% in the table, you must use interpolation. The score corresponding to a rank of 15% is $X = 7$.
   b. Because $X = 13$ is between the real limits of 9.5 and 14.5, you must use interpolation. The percentile rank for $X = 13$ is 55%.

2.6 STEM AND LEAF DISPLAYS

In 1977, J.W. Tukey presented a technique for organizing data that provides a simple alternative to a grouped frequency distribution table or graph (Tukey, 1977). This technique, called a *stem and leaf display*, requires that each score be separated into two parts: The first digit (or digits) is called the *stem*, and the last digit is called the *leaf*. For example, $X = 85$ would be separated into a stem of 8 and a leaf of 5. Similarly, $X = 42$ would have a stem of 4 and a leaf of 2. To construct a stem and leaf display for a set of data, the first step is to list all the stems in a column. For the data in Table 2.3, for example, the lowest scores are in the 30s and the highest scores are in the 90s, so the list of stems would be

Stems
3
4
5
6
7
8
9

The next step is to go through the data, one score at a time, and write the leaf for each score beside its stem. For the data in Table 2.3, the first score is $X = 83$, so you would write 3 in the leaf column beside the 8 in the column of stems. This process is continued for the entire set of scores. The complete stem and leaf display is shown with the original data in Table 2.3.

COMPARING STEM AND LEAF DISPLAYS WITH FREQUENCY DISTRIBUTIONS

Notice that the stem and leaf display is very similar to a grouped frequency distribution. Each of the stem values corresponds to a class interval. For example, the stem 3 represents all scores in the 30s—that is, all scores in the interval 30–39. The number of leaves in the display shows the frequency associated with each stem. It also should be clear that the stem and leaf display has one important advantage over a traditional grouped frequency distribution. Specifically, the stem and leaf display allows you to identify every individual score in the data. In the display shown in Table 2.3, for example, you know that there were three scores in the 60s and that the specific values were 62, 68, and 63. A frequency distribution would tell you only the frequency, not

TABLE 2.3
A set of $N = 24$ scores presented as raw data and organized in a stem and leaf display.

Data			Stem and Leaf Display	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	6279
76	52	97	6	283
85	42	46	7	1643846
32	57	59	8	3521
56	73	74	9	37
74	81	76		

the specific values. This advantage can be very valuable, especially if you need to do any calculations with the original scores. For example, if you need to add all the scores, you can recover the actual values from the stem and leaf display and compute the total. With a grouped frequency distribution, however, the individual scores are not available.

LEARNING CHECK

1. Use a stem and leaf display to organize the following set of scores:

74, 103, 95, 98, 81, 117, 105, 99, 63, 86, 94, 107

96, 100, 98, 118, 107, 82, 84, 71, 91, 107, 84, 77

2. Explain how a stem and leaf display contains more information than a grouped frequency distribution.

ANSWERS

1. The stem and leaf display for these data is as follows:

6	3
7	417
8	16244
9	5894681
10	357077
11	78

2. A grouped frequency distribution table tells only the number of scores in each interval; it does not identify the exact value for each score. The stem and leaf display identifies the individual scores as well as the number of scores in each interval.

SUMMARY

1. The goal of descriptive statistics is to simplify the organization and presentation of data. One descriptive technique is to place the data in a frequency distribution table or graph that shows exactly how many individuals (or scores) are located in each category on the scale of measurement.

2. A frequency distribution table lists the categories that make up the scale of measurement (the X values) in one column. Beside each X value, in a second column, is the frequency or number of individuals in that category. The table may include a proportion column showing the relative frequency for each category:

$$\text{proportion} = p = \frac{f}{n}$$

The table may include a percentage column showing the percentage associated with each X value:

$$\text{percentage} = p(100) = \frac{f}{n}(100)$$

3. It is recommended that a frequency distribution table have a maximum of 10 to 15 rows to keep it simple. If the scores cover a range that is wider than this suggested maximum, it is customary to divide the range into sections called *class intervals*. These intervals are then listed in the frequency distribution table along with the frequency or number of individuals with scores in each interval. The result is called a *grouped frequency distribution*. The guidelines for constructing a grouped frequency distribution table are as follows:

a. There should be about 10 intervals.

b. The width of each interval should be a simple number (e.g., 2, 5, or 10).

c. The bottom score in each interval should be a multiple of the width.

d. All intervals should be the same width, and they should cover the range of scores with no gaps.

4. A frequency distribution graph lists scores on the horizontal axis and frequencies on the vertical axis. The type of graph used to display a distribution depends on the scale of measurement used. For interval or ratio scales, you should use a histogram or a polygon. For a histogram, a bar is drawn above each score so that the height of the bar corresponds to the frequency. Each bar extends to the real limits of the score, so that adjacent bars touch. For a polygon, a dot is placed above the midpoint of each score or class interval so that the height of the dot corresponds to the frequency; then lines are drawn to connect the dots. Bar graphs are used with nominal or ordinal scales. Bar graphs are similar to histograms except that gaps are left between adjacent bars.
5. Shape is one of the basic characteristics used to describe a distribution of scores. Most distributions can be classified as either symmetrical or skewed. A skewed distribution with the tail on the right is said to be positively skewed. If it has the tail on the left, it is negatively skewed.
6. The cumulative percentage is the percentage of individuals with scores at or below a particular point in the distribution. The cumulative percentage values are associated with the upper real limits of the corresponding scores or intervals.
7. Percentiles and percentile ranks are used to describe the position of individual scores within a distribution. Percentile rank gives the cumulative percentage associated with a particular score. A score that is identified by its rank is called a *percentile*.
8. When a desired percentile or percentile rank is located between two known values, it is possible to estimate the desired value using the process of interpolation. Interpolation assumes a regular linear change between the two known values.
9. A stem and leaf display is an alternative procedure for organizing data. Each score is separated into a stem (the first digit or digits) and a leaf (the last digit or digits). The display consists of the stems listed in a column with the leaf for each score written beside its stem. A stem and leaf display combines the characteristics of a table and a graph and produces a concise, well-organized picture of the data.

KEY TERMS

frequency distribution (39)
range (42)
grouped frequency distribution (42)
class interval (42)
apparent limits (44)
histogram (46)
polygon (47)
bar graph (48)
relative frequency (49)
symmetrical distribution (50)
tail(s) of a distribution (51)
positively skewed distribution (51)
negatively skewed distribution (51)
percentile rank (53)
percentile (53)
cumulative frequency (*cf*) (54)
cumulative percentage (*c*%) (54)
interpolation (55)
stem and leaf display (60)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 2 on the book companion website.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to produce **Frequency Distribution Tables or Graphs**.

Frequency Distribution Tables

Data Entry

1. Enter all the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Frequencies.**
2. Highlight the column label for the set of scores (VAR00001) in the left box and click the arrow to move it into the **Variable** box.
3. Be sure that the option to **Display Frequency Table** is selected.
4. Click **OK.**

SPSS Output

The frequency distribution table lists the score values in a column from smallest to largest, with the percentage and cumulative percentage also listed for each score. Score values that do not occur (zero frequencies) are not included in the table, and the program does not group scores into class intervals (all values are listed).

Frequency Distribution Histograms or Bar Graphs

Data Entry

1. Enter all the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Frequencies.**
2. Highlight the column label for the set of scores (VAR00001) in the left box and click the arrow to move it into the **Variable** box.
3. Click **Charts.**
4. Select either **Bar Graphs** or **Histogram.**
5. Click **Continue.**
6. Click **OK**.

SPSS Output

After a brief delay, SPSS displays a frequency distribution table and a graph. Note that SPSS often produces a histogram that groups the scores in unpredictable intervals. A bar graph usually produces a clearer picture of the actual frequency associated with each score.

FOCUS ON PROBLEM SOLVING

1. The reason for constructing frequency distributions is to put a disorganized set of raw data into a comprehensible, organized format. Because several different types of frequency distribution tables and graphs are available, one problem is deciding which type to use. Tables have the advantage of being easier to construct, but graphs generally give a better picture of the data and are easier to understand.

   To help you decide which type of frequency distribution is best, consider the following points:

   a. What is the range of scores? With a wide range, you need to group the scores into class intervals.

   b. What is the scale of measurement? With an interval or a ratio scale, you can use a polygon or a histogram. With a nominal or an ordinal scale, you must use a bar graph.

2. When using a grouped frequency distribution table, a common mistake is to calculate the interval width by using the highest and lowest values that define each interval. For example, some students are tricked into thinking that an interval identified as 20–24 is only 4 points wide. To determine the correct interval width, you can:

   a. Count the individual scores in the interval. For this example, the scores are 20, 21, 22, 23, and 24, for a total of 5 values. Thus, the interval width is 5 points.

   b. Use the real limits to determine the real width of the interval. For example, an interval identified as 20–24 has a lower real limit of 19.5 and an upper real limit of 24.5 (halfway to the next score). Using the real limits, the interval width is

   $$24.5 - 19.5 = 5 \text{ points}$$

3. Percentiles and percentile ranks are intended to identify specific locations within a distribution of scores. When solving percentile problems, especially with interpolation, it is helpful to sketch a frequency distribution graph. Use the graph to make a preliminary estimate of the answer before you begin any calculations. For example, to find the 60th percentile, draw a vertical line through the graph so that slightly more than half (60%) of the distribution is on the left-hand side of the line. Locating this position in your sketch gives you a rough estimate of what the final answer should be. When doing interpolation problems, you should keep several points in mind:

   a. Remember that the cumulative percentage values correspond to the upper real limits of each score or interval.

   b. You should always identify the interval with which you are working. The easiest way to do this is to create a table showing the endpoints on both scales (scores and cumulative percentages). This is illustrated in Example 2.7 on pages 57–58.

   c. The word *interpolation* means *between two poles*. Remember: Your goal is to find an intermediate value between the two ends of the interval. Check your answer to be sure that it is located between the two endpoints. If it is not, then check your calculations.

DEMONSTRATION 2.1

A GROUPED FREQUENCY DISTRIBUTION TABLE

For the following set of $N = 20$ scores, construct a grouped frequency distribution table using an interval width of 5 points. The scores are:

14, 8, 27, 16, 10, 22, 9, 13, 16, 12,
10, 9, 15, 17, 6, 14, 11, 18, 14, 11

STEP 1 Set up the class intervals.

The largest score in this distribution is $X = 27$, and the lowest is $X = 6$. Therefore, a frequency distribution table for these data would have 22 rows and would be too large. A grouped frequency distribution table would be better. We have asked specifically for an interval width of 5 points, and the resulting table has five rows.

X
25–29
20–24
15–19
10–14
5–9

Remember that the interval width is determined by the real limits of the interval. For example, the class interval 25–29 has an upper real limit of 29.5 and a lower real limit of 24.5. The difference between these two values is the width of the interval—namely, 5.

STEP 2 Determine the frequencies for each interval.

Examine the scores, and count how many fall into the class interval of 25–29. Cross out each score that you have already counted. Record the frequency for this class interval. Now repeat this process for the remaining intervals. The result is the following table:

X	*f*	
25–29	1	(the score $X = 27$)
20–24	1	($X = 22$)
15–19	5	(the scores $X = 16$, 16, 15, 17, and 18)
10–14	9	($X = 14$, 10, 13, 12, 10, 14, 11, 14, and 11)
5–9	4	($X = 8$, 9, 9, and 6)

DEMONSTRATION 2.2

USING INTERPOLATION TO FIND PERCENTILES AND PERCENTILE RANKS

Find the 50th percentile for the set of scores in the grouped frequency distribution table that was constructed in Demonstration 2.1.

STEP 1 Find the cumulative frequency (*cf*) and cumulative percentage values, and add these values to the basic frequency distribution table.

Cumulative frequencies indicate the number of individuals located in or below each category (class interval). To find these frequencies, begin with the bottom interval, and then accumulate the frequencies as you move up the scale.

Cumulative percentages are determined from the cumulative frequencies by the relationship

$$c\% = \left(\frac{cf}{N}\right)100\%$$

For example, the *cf* column shows that 4 individuals (out of the total set of $N = 20$) have scores in or below the 5–9 interval. The corresponding cumulative percentage is

$$c\% = \left(\frac{4}{20}\right)100\% = \left(\frac{1}{5}\right)100\% = 20\%$$

The complete set of cumulative frequencies and cumulative percentages is shown in the following table:

X	*f*	*cf*	*c%*
25–29	1	20	100%
20–24	1	19	95%
15–19	5	18	90%
10–14	9	13	65%
5–9	4	4	20%

STEP 2 Locate the interval that contains the value that you want to calculate.

We are looking for the 50th percentile, which is located between the values of 20% and 65% in the table. The scores (upper real limits) corresponding to these two percentages are 9.5 and 14.5, respectively. The interval, measured in terms of scores and percentages, is shown in the following table:

X	*c%*
14.5	65%
??	50%
9.5	20%

STEP 3 Locate the intermediate value as a fraction of the total interval.

Our intermediate value is 50%, which is located in the interval between 65% and 20%. The total width of the interval is 45 points (65 – 20 = 45), and the value of 50% is located 15 points down from the top of the interval. As a fraction, the 50th percentile is located $\frac{15}{45} = \frac{1}{3}$ down from the top of the interval.

STEP 4 Use the fraction to determine the corresponding location on the other scale.

Our intermediate value, 50%, is located $\frac{1}{3}$ of the way down from the top of the interval. Our goal is to find the score, the *X* value, that also is located $\frac{1}{3}$ of the way down from the top of the interval.

On the score (*X*) side of the interval, the top value is 14.5, and the bottom value is 9.5, so the total interval width is 5 points (14.5 – 9.5 = 5). The position we are seeking is $\frac{1}{3}$ of the way from the top of the interval. One-third of the total interval is

$$\left(\frac{1}{3}\right)5 = \frac{5}{3} = 1.67 \text{ points}$$

To find this location, begin at the top of the interval, and come down 1.67 points:

$$14.5 - 1.67 = 12.83$$

This is our answer. The 50th percentile is $X = 12.83$.

PROBLEMS

1. Place the following sample of $n = 20$ scores in a frequency distribution table.

6, 9, 9, 10, 8, 9, 4, 7, 10, 9
5, 8, 10, 6, 9, 6, 8, 8, 7, 9

2. Construct a frequency distribution table for the following set of scores. Include columns for proportion and percentage in your table.

Scores: 5, 7, 8, 4, 7, 9, 6, 6, 5, 3
9, 6, 4, 7, 7, 8, 6, 7, 8, 5

3. Find each value requested for the distribution of scores in the following table.

a. n
b. ΣX
c. ΣX^2

X	f
5	2
4	3
3	5
2	1
1	1

4. Find each value requested for the distribution of scores in the following table.

a. n
b. ΣX
c. ΣX^2

X	f
5	1
4	2
3	3
2	5
1	3

5. For the following scores, the smallest value is $X = 8$ and the largest value is $X = 29$. Place the scores in a grouped frequency distribution table

a. using an interval width of 2 points.
b. using an interval width of 5 points.

24, 19, 23, 10, 25, 27, 22, 26
25, 20, 8, 24, 29, 21, 24, 13
23, 27, 24, 16, 22, 18, 26, 25

6. The following scores are the ages for a random sample of $n = 30$ drivers who were issued speeding tickets in New York during 2008. Determine the best interval width and place the scores in a grouped frequency distribution table. From looking at your table, does it appear that tickets are issued equally across age groups?

17, 30, 45, 20, 39, 53, 28, 19,
24, 21, 34, 38, 22, 29, 64,
22, 44, 36, 16, 56, 20, 23, 58,
32, 25, 28, 22, 51, 26, 43

7. For each of the following samples, determine the interval width that is most appropriate for a grouped frequency distribution and identify the approximate number of intervals needed to cover the range of scores.

a. Sample scores range from $X = 24$ to $X = 41$
b. Sample scores range from $X = 46$ to $X = 103$
c. Sample scores range from $X = 46$ to $X = 133$

8. What information can you obtain about the scores in a regular frequency distribution table that is not available from a grouped table?

9. Describe the difference in appearance between a bar graph and a histogram and describe the circumstances in which each type of graph is used.

10. For the following set of quiz scores:

3, 5, 4, 6, 2, 3, 4, 1, 4, 3
7, 7, 3, 4, 5, 8, 2, 4, 7, 10

a. Construct a frequency distribution table to organize the scores.
b. Draw a frequency distribution histogram for these data.

11. Sketch a histogram and a polygon showing the distribution of scores presented in the following table:

X	*f*
7	1
6	1
5	3
4	6
3	4
2	1

12. A survey given to a sample of 200 college students contained questions about the following variables. For each variable, identify the kind of graph that should be used to display the distribution of scores (histogram, polygon, or bar graph).
a. number of pizzas consumed during the previous week
b. size of T-shirt worn (S, M, L, XL)
c. gender (male/female)
d. grade point average for the previous semester
e. college class (freshman, sophomore, junior, senior)

13. Each year the college gives away T-shirts to new students during freshman orientation. The students are allowed to pick the shirt sizes that they want. To determine how many of each size shirt they should order, college officials look at the distribution from last year. The following table shows the distribution of shirt sizes selected last year.

Size	*f*
S	27
M	48
L	136
XL	120
XXL	39

a. What kind of graph would be appropriate for showing this distribution?
b. Sketch the frequency distribution graph.

14. A report from the college dean indicates that for the previous semester, the grade distribution for the Department of Psychology included 135 As, 158 Bs, 140 Cs, 94 Ds, and 53 Fs. Determine what kind of graph would be appropriate for showing this distribution and sketch the frequency distribution graph.

15. For the following set of scores

Scores: 5, 8, 5, 7, 6, 6, 5, 7, 4, 6
6, 9, 5, 5, 4, 6, 7, 5, 7, 5

a. Place the scores in a frequency distribution table.
b. Identify the shape of the distribution.

16. Place the following scores in a frequency distribution table. Based on the frequencies, what is the shape of the distribution?

5, 6, 4, 7, 7, 6, 8, 2, 5, 6
3, 1, 7, 4, 6, 8, 2, 6, 5, 7

17. For the following set of scores:

3, 7, 6, 5, 5, 9, 6, 4, 6, 8
10, 2, 7, 4, 9, 5, 6, 3, 8

a. Construct a frequency distribution table.
b. Sketch a polygon showing the distribution.
c. Describe the distribution using the following characteristics:
(1) What is the shape of the distribution?
(2) What score best identifies the center (average) for the distribution?
(3) Are the scores clustered together, or are they spread out across the scale?

18. Fowler and Christakis (2008) report that personal happiness tends to be associated with having a social network including many other happy friends. To test this claim, a researcher obtains a sample of $n = 16$ adults who claim to be happy people and a similar sample of $n = 16$ adults who describe themselves as neutral or unhappy. Each individual is then asked to identify the number of their close friends whom they consider to be happy people. The scores are as follows:

Happy:
8, 7, 4, 10, 6, 6, 8, 9, 8, 8,
7, 5, 6, 9, 8, 9

Unhappy:
5, 8, 4, 6, 6, 7, 9, 6, 2, 8,
5, 6, 4, 7, 5, 6

Sketch a polygon showing the frequency distribution for the happy people. In the same graph, sketch a polygon for the unhappy people. (Use two different colors, or use a solid line for one polygon and a dashed line for the other.) Does one group seem to have more happy friends?

19. Complete the final two columns in the following frequency distribution table and then find the percentiles and percentile ranks requested.

X	*f*	*cf*	*c%*
7	2		
6	3		
5	6		
4	9		
3	4		
2	1		

a. What is the percentile rank for $X = 2.5$?
b. What is the percentile rank for $X = 6.5$?
c. What is the 20th percentile?
d. What is the 80th percentile?

20. Complete the final two columns in the following frequency distribution table and then find the percentiles and percentile ranks requested.

X	*f*	*cf*	*c%*
50–59	1		
40–49	3		
30–39	6		
20–29	5		
10–19	3		
0–9	2		

a. What is the percentile rank for $X = 9.5$?
b. What is the percentile rank for $X = 39.5$?
c. What is the 25th percentile?
d. What is the 50th percentile?

21. Complete the final two columns in the following frequency distribution table and then use interpolation to find the percentiles and percentile ranks requested.

X	*f*	*cf*	*c%*
10	2		
9	5		
8	8		
7	15		
6	10		
5	6		
4	4		

a. What is the percentile rank for $X = 6$?
b. What is the percentile rank for $X = 9$?
c. What is the 25th percentile?
d. What is the 90th percentile?

22. Find the requested percentiles and percentile ranks for the following distribution of quiz scores for a class of $N = 40$ students.

X	*f*	*cf*	*c%*
20	2	40	100.0
19	4	38	95.0
18	6	34	85.0
17	13	28	70.0
16	6	15	37.5
15	4	9	22.5
14	3	5	12.5
13	2	2	5.0

a. What is the percentile rank for $X = 15$?
b. What is the percentile rank for $X = 18$?
c. What is the 15th percentile?
d. What is the 90th percentile?

23. Use interpolation to find the requested percentiles and percentile ranks requested for the following distribution of scores.

X	*f*	*cf*	*c%*
14–15	3	50	100
12–13	6	47	94
10–11	8	41	82
8–9	18	33	66
6–7	10	15	30
4–5	4	5	10
2–3	1	1	2

a. What is the percentile rank for $X = 5$?
b. What is the percentile rank for $X = 12$?
c. What is the 25th percentile?
d. What is the 70th percentile?

24. The following frequency distribution presents a set of exam scores for a class of $N = 20$ students.

X	*f*	*cf*	*c%*
90–99	4	20	100
80–89	7	16	80
70–79	4	9	45
60–69	3	5	25
50–59	2	2	10

a. Find the 30th percentile.
b. Find the 88th percentile.
c. What is the percentile rank for $X = 77$?
d. What is the percentile rank for $X = 90$?

25. Construct a stem and leaf display for the data in problem 6 using one stem for the scores in the 60s, one for scores in the 50s, and so on.

26. A set of scores has been organized into the following stem and leaf display. For this set of scores:
a. How many scores are in the 70s?
b. Identify the individual scores in the 70s.
c. How many scores are in the 40s?
d. Identify the individual scores in the 40s.

3	8
4	60
5	734
6	81469
7	2184
8	247

27. Use a stem and leaf display to organize the following distribution of scores. Use seven stems with each stem corresponding to a 10-point interval.

Scores:

28,	54,	65,	53,	81
45,	44,	51,	72,	34
43,	59,	65,	39,	20
53,	74,	24,	30,	49
36,	58,	60,	27,	47
22,	52,	46,	39,	65

CHAPTER

3

Central Tendency

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Summation notation (Chapter 1)
- Frequency distributions (Chapter 2)

Preview

Research has now confirmed what you already suspected to be true—alcohol consumption increases the attractiveness of opposite-sex individuals (Jones, Jones, Thomas, & Piper, 2003). In the study, college-age participants were recruited from bars and restaurants near campus and asked to participate in a "market research" study. During the introductory conversation, they were asked to report their alcohol consumption for the day and were told that moderate consumption would not prevent them from taking part in the study. Participants were then shown a series of photographs of male and female faces and asked to rate the attractiveness of each face on a scale from 1 to 7. Figure 3.1 shows the general pattern of results obtained in the study. The two polygons in the figure show the distributions of attractiveness ratings for one female photograph obtained from two groups of males: those who had no alcohol and those with moderate alcohol consumption. Note that the attractiveness ratings from the alcohol group are noticeably higher than the ratings from the no-alcohol group. Incidentally, the same pattern of results was obtained for the female's ratings of male photographs.

The Problem: Although it seems obvious that the moderate-alcohol ratings are noticeably higher than the no-alcohol ratings, this conclusion is based on a general impression, or a subjective interpretation, of the figure. In fact, this conclusion is not always true. For example, there is overlap between the two groups so that some of the no-alcohol males actually rate the photograph as more attractive than some of the moderate-alcohol males. What we need is a method to summarize each group as a whole so that we can objectively describe how much difference exists between the two groups.

The Solution: A measure of *central tendency* identifies the average, or typical, score to serve as a representative value for each group. Then we can use the two averages to describe the two groups and to measure the difference between them. The results should show that average attractiveness rating from males consuming alcohol really is higher than the average rating from males who have not consumed alcohol.

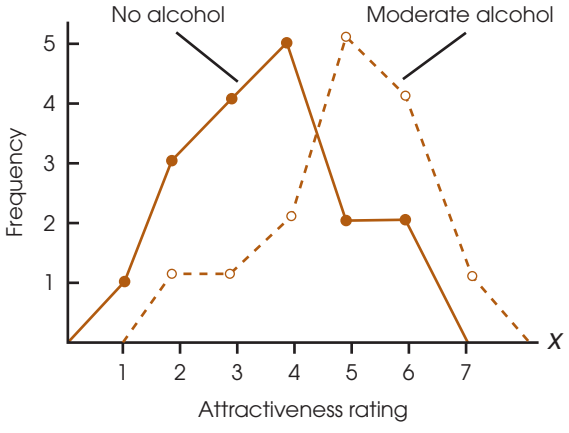


FIGURE 3.1

Frequency distributions for ratings of attractiveness of a female face shown in a photograph for two groups of male participants: those who had consumed no alcohol and those who had consumed moderate amounts of alcohol.

3.1 OVERVIEW

The general purpose of descriptive statistical methods is to organize and summarize a set of scores. Perhaps the most common method for summarizing and describing a distribution is to find a single value that defines the average score and can serve as a representative for the entire distribution. In statistics, the concept of an average, or representative, score is called *central tendency*. The goal in measuring central tendency is to describe a distribution of scores by determining a single value that identifies the center of the distribution. Ideally, this central value is the score that is the best representative value for all of the individuals in the distribution.

DEFINITION

Central tendency is a statistical measure to determine a single score that defines the center of a distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group.

In everyday language, central tendency attempts to identify the "average" or "typical" individual. This average value can then be used to provide a simple description of an entire population or a sample. In addition to describing an entire distribution, measures of central tendency are also useful for making comparisons between groups of individuals or between sets of figures. For example, weather data indicate that for Seattle, Washington, the average yearly temperature is 53° and the average annual precipitation is 34 inches. By comparison, the average temperature in Phoenix, Arizona, is 71° and the average precipitation is 7.4 inches. The point of these examples is to demonstrate the great advantage of being able to describe a large set of data with a single, representative number. Central tendency characterizes what is typical for a large population and, in doing so, makes large amounts of data more digestible. Statisticians sometimes use the expression *number crunching* to illustrate this aspect of data description. That is, we take a distribution consisting of many scores and "crunch" them down to a single value that describes them all.

Unfortunately, there is no single, standard procedure for determining central tendency. The problem is that no single measure produces a central, representative value in every situation. The three distributions shown in Figure 3.2 should help demonstrate this fact. Before we discuss the three distributions, take a moment to look at the figure and try to identify the center or the most representative score for each distribution.

1. The first distribution [Figure 3.2(a)] is symmetrical, with the scores forming a distinct pile centered around $X = 5$. For this type of distribution, it is easy to identify the center, and most people would agree that the value $X = 5$ is an appropriate measure of central tendency.
2. In the second distribution [Figure 3.2(b)], however, problems begin to appear. Now the scores form a negatively skewed distribution, piling up at the high end of the scale around $X = 8$, but tapering off to the left all the way down to $X = 1$. Where is the center in this case? Some people might select $X = 8$ as the center because more individuals had this score than any other single value. However, $X = 8$ is clearly not in the middle of the distribution. In fact, the majority of the scores (10 out of 16) have values less than 8, so it seems reasonable that the center should be defined by a value that is less than 8.
3. Now consider the third distribution [Figure 3.2(c)]. Again, the distribution is symmetrical, but now there are two distinct piles of scores. Because the distribution is symmetrical with $X = 5$ as the midpoint, you may choose $X = 5$ as the center. However, none of the scores is located at $X = 5$ (or even close), so this value is not particularly good as a representative score. On the other hand, because there are two separate piles of scores with one group centered at $X = 2$ and the other centered at $X = 8$, it is tempting to say that this distribution has two centers. But can one distribution have two centers?

Clearly, there can be problems defining the center of a distribution. Occasionally, you will find a nice, neat distribution like the one shown in Figure 3.2(a), for which everyone agrees on the center. But you should realize that other distributions are possible and that there may be different opinions concerning the definition of the center. To deal with these problems, statisticians have developed three different methods for measuring central tendency: the mean, the median, and the mode. They are computed

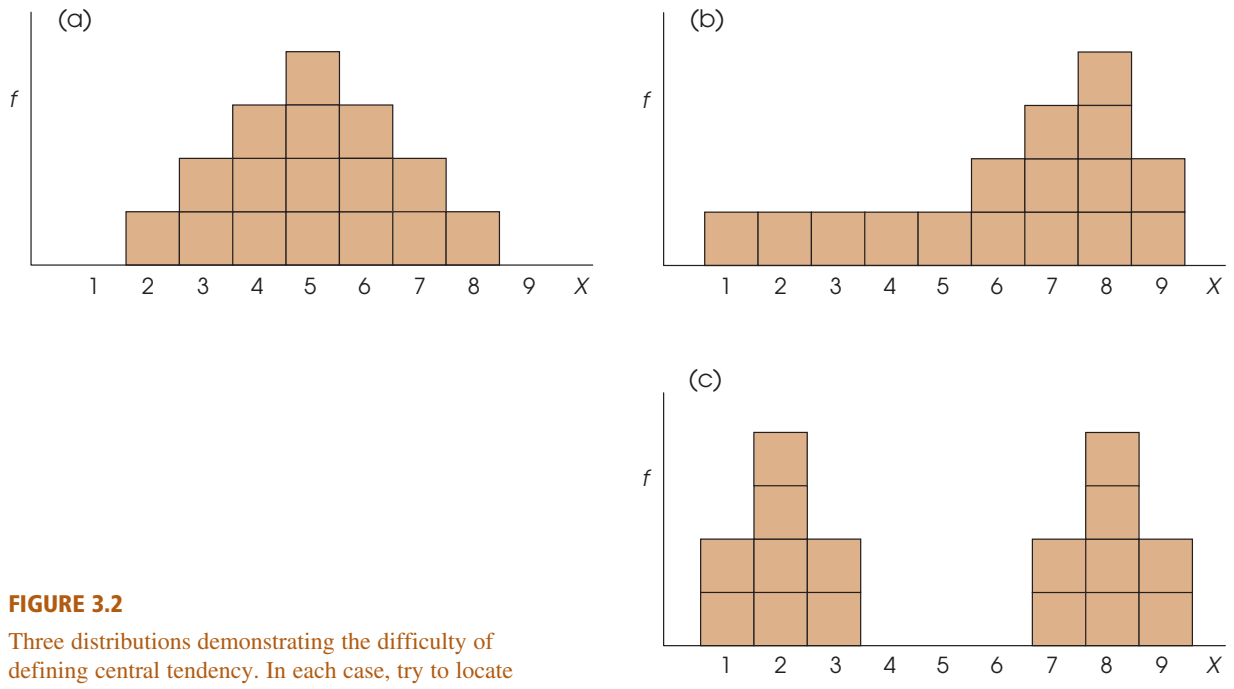


FIGURE 3.2

Three distributions demonstrating the difficulty of defining central tendency. In each case, try to locate the "center" of the distribution.

differently and have different characteristics. To decide which of the three measures is best for any particular distribution, you should keep in mind that the general purpose of central tendency is to find the single most representative score. Each of the three measures we present has been developed to work best in a specific situation. We examine this issue in more detail after we introduce the three measures.

3.2 THE MEAN

The *mean*, also known as the arithmetic average, is computed by adding all the scores in the distribution and dividing by the number of scores. The mean for a population is identified by the Greek letter mu, μ (pronounced "mew"), and the mean for a sample is identified by M or $\overline{X}$ (read "x-bar").

The convention in many statistics textbooks is to use $\overline{X}$ to represent the mean for a sample. However, in manuscripts and in published research reports the letter M is the standard notation for a sample mean. Because you will encounter the letter M when reading research reports and because you should use the letter M when writing research reports, we have decided to use the same notation in this text. Keep in mind that the $\overline{X}$ notation is still appropriate for identifying a sample mean, and you may find it used on occasion, especially in textbooks.

DEFINITION

The **mean** for a distribution is the sum of the scores divided by the number of scores.

The formula for the *population mean* is

$$\mu = \frac{\sum X}{N} \tag{3.1}$$

First, add all the scores in the population, and then divide by N. For a sample, the computation is exactly the same, but the formula for the *sample mean* uses symbols that signify sample values:

$$\text{sample mean} = M = \frac{\sum X}{n} \tag{3.2}$$

In general, we use Greek letters to identify characteristics of a population (parameters) and letters of our own alphabet to stand for sample values (statistics). If a mean is identified with the symbol M, you should realize that we are dealing with a sample. Also note that the equation for the sample mean uses a lowercase n as the symbol for the number of scores in the sample.

EXAMPLE 3.1

For a population of $N = 4$ scores,

3, 7, 4, 6

the mean is

$$\mu = \frac{\Sigma X}{N} = \frac{20}{4} = 5$$

ALTERNATIVE DEFINITIONS FOR THE MEAN

Although the procedure of adding the scores and dividing by the number of scores provides a useful definition of the mean, there are two alternative definitions that may give you a better understanding of this important measure of central tendency.

Dividing the total equally The first alternative is to think of the mean as the amount each individual receives when the total (ΣX) is divided equally among all of the individuals (N) in the distribution. This somewhat socialistic viewpoint is particularly useful in problems for which you know the mean and must find the total. Consider the following example.

EXAMPLE 3.2

A group of $n = 6$ boys buys a box of baseball cards at a garage sale and discovers that the box contains a total of 180 cards. If the boys divide the cards equally among themselves, how many cards will each boy get? You should recognize that this problem represents the standard procedure for computing the mean. Specifically, the total (ΣX) is divided by the number (n) to produce the mean, $\frac{180}{6} = 30$ cards for each boy.

The previous example demonstrates that it is possible to define the mean as the amount that each individual gets when the total is distributed equally. This new definition can be useful for some problems involving the mean. Consider the following example.

EXAMPLE 3.3

Now suppose that the 6 boys from Example 3.2 decide to sell their baseball cards on eBay. If they make an average of $M = \$5$ per boy, what is the total amount of money for the whole group? Although you do not know exactly how much money each boy has, the new definition of the mean tells you that if they pool their money together and then distribute the total equally, each boy will get \$5. For each of $n = 6$ boys to get \$5, the total must be 6(\$5) = \$30. To check this answer, use the formula for the mean:

$$M = \frac{\Sigma X}{n} = \frac{\$30}{6} = \$5$$

The mean as a balance point The second alternative definition of the mean describes the mean as a balance point for the distribution. Consider a population consisting of $N = 5$ scores (1, 2, 6, 6, 10). For this population, $\Sigma X = 25$ and $\mu = \frac{25}{5} = 5$. Figure 3.3 shows this population drawn as a histogram, with each score represented as a box that is sitting on a seesaw. If the seesaw is positioned so that it pivots at a point equal to the mean, then it will be balanced and will rest level.

The reason that the seesaw is balanced over the mean becomes clear when we measures the distance of each box (score) from the mean:

Score	Distance from the Mean
$X = 1$	4 points below the mean
$X = 2$	3 points below the mean
$X = 6$	1 point above the mean
$X = 6$	1 point above the mean
$X = 10$	5 points above the mean

Notice that the mean balances the distances. That is, the total distance below the mean is the same as the total distance above the mean:

below the mean: $4 + 3 = 7$ points

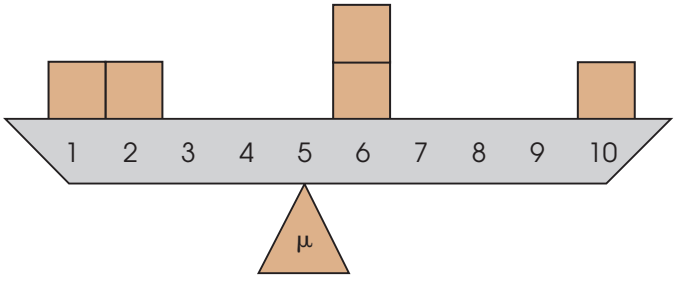
above the mean: $1 + 1 + 5 = 7$ points

Because the mean serves as a balance point, the value of the mean is always located somewhere between the highest score and the lowest score; that is, the mean can never be outside the range of scores. If the lowest score in a distribution is $X = 8$ and the highest is $X = 15$, then the mean *must* be between 8 and 15. If you calculate a value that is outside this range, then you have made an error.

FIGURE 3.3

The frequency distribution shown as a seesaw balanced at the mean.

(Based on G. H. Weinberg, J. A. Schumaker, & D. Oltman (1981). *Statistics: An Intuitive Approach* (p. 14). Belmont, Calif.: Wadsworth.



The image of a seesaw with the mean at the balance point is also useful for determining how a distribution is affected if a new score is added or if an existing score is removed. For the distribution in Figure 3.3, for example, what would happen to the mean (balance point) if a new score were added at $X = 10$?

THE WEIGHTED MEAN

Often it is necessary to combine two sets of scores and then find the overall mean for the combined group. Suppose that we begin with two separate samples. The first sample has $n = 12$ scores and a mean of $M = 6$. The second sample has $n = 8$ and $M = 7$. If the two samples are combined, what is the mean for the total group?

To calculate the overall mean, we need two values:

1. the overall sum of the scores for the combined group (ΣX), and
2. the total number of scores in the combined group (n).

The total number of scores in the combined group can be found easily by adding the number of scores in the first sample (n_1) and the number in the second sample (n_2). In this case, there are $12 + 8 = 20$ scores in the combined group. Similarly, the overall sum for the combined group can be found by adding the sum for the first sample (ΣX_1) and the sum for the second sample (ΣX_2). With these two values, we can compute the mean using the basic equation

$$\text{overall mean} = M = \frac{\Sigma X \text{ (overall sum for the combined group)}}{n \text{ (total number in the combined group)}}$$

$$= \frac{\Sigma X_1 + \Sigma X_2}{n_1 + n_2}$$

To find the sum of the scores for each sample, remember that the mean can be defined as the amount each person receives when the total (ΣX) is distributed equally. The first sample has $n = 12$ and $M = 6$. (Expressed in dollars instead of scores, this sample has $n = 12$ people and each person gets \$6 when the total is divided equally.) For each of 12 people to get $M = 6$, the total must be $\Sigma X = 12 \times 6 = 72$. In the same way, the second sample has $n = 8$ and $M = 7$ so the total must be $\Sigma X = 8 \times 7 = 56$. Using these values, we obtain an overall mean of

$$\text{overall mean} = M = \frac{\Sigma X_1 + \Sigma X_1}{n_1 + n_2} = \frac{72 + 56}{12 + 8} = \frac{128}{20} = 6.4$$

The following table summarizes the calculations.

First Sample	Second Sample	Combined Sample
$n = 12$	$n = 8$	$n = 20$ (12 + 8)
$\Sigma X = 72$	$\Sigma X = 56$	$\Sigma X = 128$ (72 + 56)
$M = 6$	$M = 7$	$M = 6.4$

Note that the overall mean is not halfway between the original two sample means. Because the samples are not the same size, one makes a larger contribution to the total group and, therefore, carries more weight in determining the overall mean. For this reason, the overall mean we have calculated is called the *weighted mean*. In this example, the overall mean of $M = 6.4$ is closer to the value of $M = 6$ (the larger sample) than it is to $M = 7$ (the smaller sample). An alternative method for finding the weighted mean is presented in Box 3.1.

BOX 3.1 AN ALTERNATIVE PROCEDURE FOR FINDING THE WEIGHTED MEAN

In the text, the weighted mean was obtained by first determining the total number of scores (n) for the two combined samples and then determining the overall sum (ΣX) for the two combined samples. The following example demonstrates how the same result can be obtained using a slightly different conceptual approach.

We begin with the same two samples that were used in the text: One sample has $M = 6$ for $n = 12$ students, and the second sample has $M = 7$ for $n = 8$ students. The goal is to determine the mean for the overall group when the two samples are combined.

Logically, when these two samples are combined, the larger sample (with $n = 12$ scores) will make a greater contribution to the combined group than the smaller sample (with $n = 8$ scores). Thus, the larger sample will carry more weight in determining the mean for the combined group. We accommodate this fact by assigning a weight to each sample mean so that the weight is determined by the size of the sample. To determine how much weight should be assigned to each sample mean, you simply consider the sample's contribution to the combined group. When the two samples are combined, the resulting group will have a total of 20 scores ($n = 12$ from the first sample and $n = 8$ from the second). The first sample contributes 12 out of 20 scores and, therefore, is assigned a weight of $\frac{12}{20}$. The second sample contributes 8 out of 20 scores, and its weight is $\frac{8}{20}$. Each sample mean is then multiplied by its weight, and the results are added to find the weighted mean for the combined sample. For this example,

$$\text{weighted mean} = \left(\frac{12}{20}\right)(6) + \left(\frac{8}{20}\right)(7)$$

$$= \frac{72}{20} + \frac{56}{20}$$

$$= 3.6 + 2.8$$

$$= 6.4$$

Note that this is the same result obtained using the method described in the text.

COMPUTING THE MEAN FROM A FREQUENCY DISTRIBUTION TABLE

When a set of scores has been organized in a frequency distribution table, the calculation of the mean is usually easier if you first remove the individual scores from the table. Table 3.1 shows a distribution of scores organized in a frequency distribution table. To compute the mean for this distribution you must be careful to use both the X values in the first column and the frequencies in the second column. The values in the table show that the distribution consists of one 10, two 9s, four 8s, and one 6, for a total of $n = 8$ scores. Remember that you can determine the number of scores by adding the frequencies, $n = \Sigma f$. To find the sum of the scores, you must be careful to add all eight scores:

$$\Sigma X = 10 + 9 + 9 + 8 + 8 + 8 + 8 + 6 = 66$$

Note that you can also find the sum of the scores by computing ΣfX as we demonstrated in Chapter 2 (pp. 40–41). Once you have found ΣX and n, you compute the mean as usual. For these data,

$$M = \frac{\sum X}{n} = \frac{66}{8} = 8.25$$

TABLE 3.1

Statistics quiz scores for a sample of $n = 8$ students.

Quiz Score (X)	f	fX
10	1	10
9	2	18
8	4	32
7	0	0
6	1	6

LEARNING CHECK

1. Find the mean for the following sample of $n = 5$ scores: 1, 8, 7, 5, 9
2. A sample of $n = 6$ scores has a mean of $M = 8$. What is the value of ΣX for this sample?
3. One sample has $n = 5$ scores with a mean of $M = 4$. A second sample has $n = 3$ scores with a mean of $M = 10$. If the two samples are combined, what is the mean for the combined sample?
4. A sample of $n = 6$ scores has a mean of $M = 40$. One new score is added to the sample and the new mean is found to be $M = 35$. What can you conclude about the value of the new score?
   a. It must be greater 40.
   b. It must be less than 40.
5. Find the values for n, ΣX, and M for the sample that is summarized in the following frequency distribution table.

X	f
5	1
4	2
3	3
2	5
1	1

ANSWERS

1. $\Sigma X = 30$ and $M = 6$
2. $\Sigma X = 48$
3. The combined sample has $n = 8$ scores that total $\Sigma X = 50$. The mean is $M = 6.25$.
4. b
5. For this sample, $n = 12$, $\Sigma X = 33$, and $M = \frac{33}{12} = 2.75$.

CHARACTERISTICS OF THE MEAN

The mean has many characteristics that will be important in future discussions. In general, these characteristics result from the fact that every score in the distribution contributes to the value of the mean. Specifically, every score adds to the total (ΣX) and every score contributes one point to the number of scores (n). These two values (ΣX and n) determine the value of the mean. We now discuss four of the more important characteristics of the mean.

Changing a score Changing the value of any score changes the mean. For example, a sample of quiz scores for a psychology lab section consists of 9, 8, 7, 5, and 1. Note that the sample consists of $n = 5$ scores with $\Sigma X = 30$. The mean for this sample is

$$M = \frac{\sum X}{n} = \frac{30}{5} = 6.00$$

Now suppose that the score of $X = 1$ is changed to $X = 8$. Note that we have added 7 points to this individual's score, which also adds 7 points to the total (ΣX). After changing the score, the new distribution consists of

9, 8, 7, 5, 8

There are still $n = 5$ scores, but now the total is $\Sigma X = 37$. Thus, the new mean is

$$M = \frac{\Sigma X}{n} = \frac{37}{5} = 7.40$$

Notice that changing a single score in the sample has produced a new mean. You should recognize that changing any score also changes the value of ΣX (the sum of the scores), and, thus, always changes the value of the mean.

Introducing a new score or removing a score Adding a new score to a distribution, or removing an existing score, usually changes the mean. The exception is when the new score (or the removed score) is exactly equal to the mean. It is easy to visualize the effect of adding or removing a score if you remember that the mean is defined as the balance point for the distribution. Figure 3.4 shows a distribution of scores represented as boxes on a seesaw that is balanced at the mean, $\mu = 7$. Imagine what would happen if we added a new score (a new box) at $X = 10$. Clearly, the seesaw would tip to the right and we would need to move the pivot point (the mean) to the right to restore balance.

Now imagine what would happen if we removed the score (the box) at $X = 9$. This time the seesaw would tip to the left and, once again, we would need to change the mean to restore balance.

Finally, consider what would happen if we added a new score of $X = 7$, exactly equal to the mean. It should be clear that the seesaw would not tilt in either direction, so the mean would stay in exactly the same place. Also note that if we removed the new score at $X = 7$, the seesaw would remain balanced and the mean would not change. In general, adding a new score or removing an existing score causes the mean to change unless the that score is located exactly at the mean.

The following example demonstrates exactly how the new mean is computed when a new score is added to an existing sample.

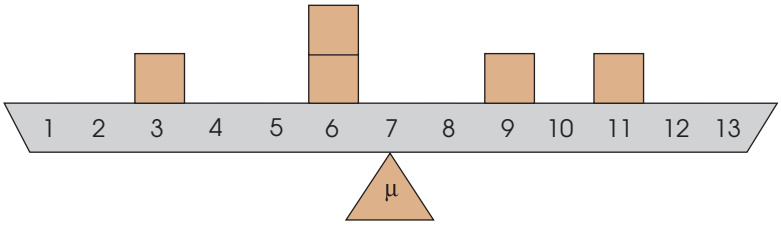
EXAMPLE 3.4

Adding a score (or removing a score) has the same effect on the mean whether the original set of scores is a sample or a population. To demonstrate the calculation of the new mean, we will use the set of scores that is shown in Figure 3.4. This time, however, we will treat the scores as a sample with $n = 5$ and $M = 7$. Note that this sample must have $\Sigma X = 35$. What happens to the mean if a new score of $X = 13$ is added to the sample?

To find the new sample mean, we must determine how the values for n and ΣX are be changed by a new score. We begin with the original sample and then consider the effect of adding the new score. The original sample had $n = 5$ scores, so adding one new score produces $n = 6$. Similarly, the original sample had $\Sigma X = 35$. Adding a score

FIGURE 3.4

A distribution of $N = 5$ scores that is balanced with a mean of $\mu = 7$.



of $X = 13$ increases the sum by 13 points, producing a new sum of $\Sigma X = 35 + 13 = 48$. Finally, the new mean is computed using the new values for n and ΣX.

$$M = \frac{\Sigma X}{n} = \frac{48}{6} = 8$$

The entire process can be summarized as follows:

Original Sample	New Sample, Adding $X = 13$
$n = 5$	$n = 6$
$\Sigma X = 35$	$\Sigma X = 48$
$M = 35/5 = 7$	$M = 48/6 = 8$

Adding or subtracting a constant from each score If a constant value is added to every score in a distribution, the same constant is added to the mean. Similarly, if you subtract a constant from every score, the same constant is subtracted from the mean.

As mentioned in Chapter 2 (p. 38), Schmidt (1994) conducted a set of experiments examining how humor influences memory. In one study, participants were shown lists of sentences, of which half were humorous (I got a bill for my surgery—now I know why those doctors were wearing masks.) and half were nonhumorous (I got a bill for my surgery—those doctors were like robbers with the prices they charged.). The results showed that people consistently recalled more of the humorous sentences.

Table 3.2 shows the results for a sample of $n = 6$ participants. The first column shows their memory scores for nonhumorous sentences. Note that the total number of sentences recalled is $\Sigma X = 17$ for a sample of $n = 6$ participants, so the mean is $M = \frac{17}{6} = 2.83$. Now suppose that the effect of humor is to add a constant amount (2 points) to each individual's memory score. The resulting scores for humorous sentences are shown in the second column of the table. For these scores, the 6 participants recalled a total of $\Sigma X = 29$ sentences, so the mean is $M = \frac{29}{6} = 4.83$. Adding 2 points to each score has also added 2 points to the mean, from $M = 2.83$ to $M = 4.83$. (It is important to note that experimental effects are usually not as simple as adding or subtracting a constant amount. Nonetheless, the concept of adding a constant to every score is important and will be addressed in later chapters when we are using statistics to evaluate the effects of experimental manipulations.)

TABLE 3.2
Number of sentences recalled for humorous and nonhumorous sentences.

Participant	Nonhumorous Sentences	Humorous Sentences
A	4	6
B	2	4
C	3	5
D	3	5
E	2	4
F	3	5
	$\Sigma X = 17$	$\Sigma X = 29$
	$M = 2.83$	$M = 4.83$

Multiplying or dividing each score by a constant If every score in a distribution is multiplied by (or divided by) a constant value, the mean changes in the same way.

Multiplying (or dividing) each score by a constant value is a common method for changing the unit of measurement. To change a set of measurements from minutes to seconds, for example, you multiply by 60; to change from inches to feet, you divide by 12. One common task for researchers is converting measurements into metric units to conform to international standards. For example, publication guidelines of the American Psychological Association call for metric equivalents to be reported in parentheses when most nonmetric units are used. Table 3.3 shows how a sample of $n = 5$ scores measured in inches would be transformed to a set of scores measured in centimeters. (Note that 1 inch equals 2.54 centimeters.) The first column shows the original scores that total $\Sigma X = 50$ with $M = 10$ inches. In the second column, each of the original scores has been multiplied by 2.54 (to convert from inches to centimeters) and the resulting values total $\Sigma X = 127$, with $M = 25.4$. Multiplying each score by 2.54 has also caused the mean to be multiplied by 2.54. You should realize, however, that although the numerical values for the individual scores and the sample mean have changed, the actual measurements have not changed.

LEARNING CHECK

1. Adding a new score to a distribution always changes the mean. (True or false?)
2. Changing the value of a score in a distribution always changes the mean. (True or false?)
3. A population has a mean of $\mu = 40$.
   a. If 5 points were added to every score, what would be the value for the new mean?
   b. If every score were multiplied by 3, what would be the value for the new mean?
4. A sample of $n = 4$ scores has a mean of 9. If one person with a score of $X = 3$ is removed from the sample, what is the value for the new sample mean?

ANSWERS

1. False. If the score is equal to the mean, it does not change the mean.
2. True.
3. **a.** The new mean would be 45. **b.** The new mean would be 120.
4. The original sample has $n = 4$ and $\Sigma X = 36$. The new sample has $n = 3$ scores that total $\Sigma X = 33$. The new mean is $M = 11$.

TABLE 3.3
Measurements converted from inches to centimeters.

Original Measurement in Inches	Conversion to Centimeters (Multiply by 2.54)
10	25.40
9	22.86
12	30.48
8	20.32
11	27.94
$\Sigma X = 50$	$\Sigma X = 127.00$
$M = 10$	$M = 25.40$

3.3 THE MEDIAN

The second measure of central tendency we consider is called the *median*. The goal of the median is to locate the midpoint of the distribution. Unlike the mean, there are no specific symbols or notation to identify the median. Instead, the median is simply identified by the word *median*. In addition, the definition and the computations for the median are identical for a sample and for a population.

DEFINITION

If the scores in a distribution are listed in order from smallest to largest, the **median** is the midpoint of the list. More specifically, the median is the point on the measurement scale below which 50% of the scores in the distribution are located.

FINDING THE MEDIAN FOR MOST DISTRIBUTIONS

Defining the median as the *midpoint* of a distribution means that the scores are divided into two equal-sized groups. We are not locating the midpoint between the highest and lowest X values. To find the median, list the scores in order from smallest to largest. Begin with the smallest score and count the scores as you move up the list. The median is the first point you reach that is greater than 50% of the scores in the distribution. The median can be equal to a score in the list or it can be a point between two scores. Notice that the median is not algebraically defined (there is no equation for computing the median), which means that there is a degree of subjectivity in determining the exact value. However, the following two examples demonstrate the process of finding the median for most distributions.

EXAMPLE 3.5

This example demonstrates the calculation of the median when n is an odd number. With an odd number of scores, you list the scores in order (lowest to highest), and the median is the middle score in the list. Consider the following set of $N = 5$ scores, which have been listed in order:

3, 5, 8, 10, 11

The middle score is $X = 8$, so the median is equal to 8. Using the counting method, with $N = 5$ scores, the 50% point would be $2\frac{1}{2}$ scores. Starting with the smallest scores, we must count the 3, the 5, and the 8 before we reach the target of at least 50%. Again, for this distribution, the median is the middle score, $X = 8$.

EXAMPLE 3.6

This example demonstrates the calculation of the median when n is an even number. With an even number of scores in the distribution, you list the scores in order (lowest to highest) and then locate the median by finding the average of the middle two scores. Consider the following population:

1, 1, 4, 5, 7, 8

Now we select the middle pair of scores (4 and 5), add them together, and divide by 2:

$$\text{median} = \frac{4 + 5}{2} = \frac{9}{2} = 4.5$$

Using the counting procedure, with $N = 6$ scores, the 50% point is 3 scores. Starting with the smallest scores, we must count the first 1, the second 1, and the 4 before we reach the target of at least 50%. Again, the median for this distribution is 4.5, which is the first point on the scale beyond $X = 4$. For this distribution, exactly 3 scores (50%) are located below 4.5. Note: If there is a gap between the middle two scores, the convention is to define the median as the midpoint between the two scores. For example, if the middle two scores are $X = 4$ and $X = 6$, the median would be defined as 5.

The simple technique of listing and counting scores is sufficient to determine the median for most distributions and is always appropriate for discrete variables. Notice that this technique always produces a median that is either a whole number or is halfway between two whole numbers. With a continuous variable, however, it is possible to divide a distribution precisely in half so that *exactly* 50% of the distribution is located below (and above) a specific point. The procedure for locating the precise median is discussed in the following section.

FINDING THE PRECISE MEDIAN FOR A CONTINUOUS VARIABLE

Recall from Chapter 1 that a continuous variable consists of categories that can be split into an infinite number of fractional parts. For example, time can be measured in seconds, tenths of a second, hundredths of a second, and so on. When the scores in a distribution are measurements of a continuous variable, it is possible to split one of the categories into fractional parts and find the median by locating the precise point that separates the bottom 50% of the distribution from the top 50%. The following example demonstrates this process.

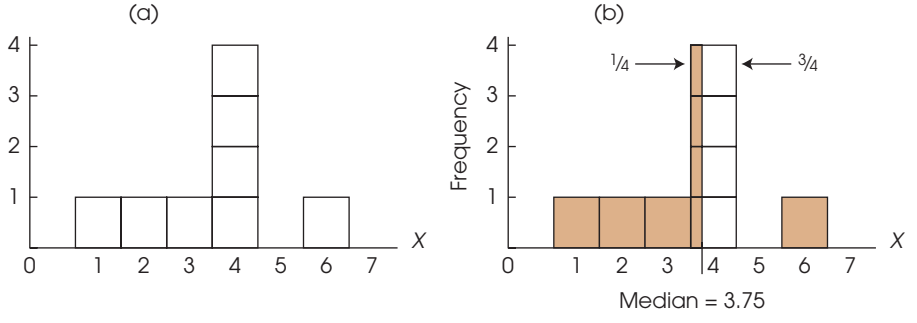
EXAMPLE 3.7

For this example, we will find the precise median for the following sample of $n = 8$ scores: 1, 2, 3, 4, 4, 4, 4, 6

The frequency distribution for this sample is shown in Figure 3.5(a). With an even number of scores, you normally would compute the average of the middle two scores to find the median. This process produces a median of $X = 4$. For a discrete variable,

FIGURE 3.5

A distribution with several scores clustered at the median. The median for this distribution is positioned so that each of the four boxes above $X = 4$ is divided into two sections, with $\frac{1}{4}$ of each box below the median (to the left) and $\frac{3}{4}$ of each box above the median (to the right). As a result, there are exactly four boxes, 50% of the distribution, on each side of the median.



$X = 4$ is the correct value for the median. Recall from Chapter 1 that a discrete variable consists of indivisible categories, such as the number of children in a family. Some families have 4 children and some have 5, but none have 4.31 children. For a discrete variable, the category $X = 4$ cannot be divided and the whole number 4 is the median.

However, if you look at the distribution histogram, the value $X = 4$ does not appear to be the exact midpoint. The problem comes from the tendency to interpret a score of $X = 4$ as meaning exactly 4.00. However, if the scores are measurements of a continuous variable, then the score $X = 4$ actually corresponds to an interval from 3.5 to 4.5, and the median corresponds to a point within this interval.

To find the precise median, we first observe that the distribution contains $n = 8$ scores represented by 8 boxes in the graph. The median is the point that has exactly 4 boxes (50%) on each side. Starting at the left-hand side and moving up the scale of measurement, we accumulate a total of 3 boxes when we reach a value of 3.5 on the X-axis [see Figure 3.5(a)]. What is needed is 1 more box to reach the goal of 4 boxes (50%). The problem is that the next interval contains four boxes. The solution is to take a fraction of each box so that the fractions combine to give you one box. For this example, if we take $\frac{1}{4}$ of each box, the four quarters will combine to make one whole box. This solution is shown in Figure 3.5(b). The fraction is determined by the number of boxes needed to reach 50% and the number that exists in the interval.

$$\text{fraction} = \frac{\text{number needed to reach 50\%}}{\text{number in the interval}}$$

For this example, we needed 1 out of the 4 boxes in the interval, so the fraction is $\frac{1}{4}$. To obtain one-fourth of each box, the median is the point that is located exactly one-fourth of the way into the interval. The interval for $X = 4$ extends from 3.5 to 4.5. The interval width is 1 point, so one-fourth of the interval corresponds to 0.25 points. Starting at the bottom of the interval and moving up 0.25 points produces a value of $3.50 + 0.25 = 3.75$. This is the median, with exactly 50% of the distribution (4 boxes) on each side.

You may recognize that the process used to find the precise median in Example 3.7 is equivalent to the process of interpolation that was introduced in Chapter 2 (pp. 55–59). Specifically, the precise median is identical to the 50th percentile for a distribution, and interpolation can be used to locate the 50th percentile. The process of using interpolation is demonstrated in Box 3.2 using the same scores that were used in Example 3.7.

Remember, finding the precise midpoint by dividing scores into fractional parts is sensible for a continuous variable, however, it is not appropriate for a discrete variable. For example, a median time of 3.75 seconds is reasonable, but a median family size of 3.75 children is not.

THE MEDIAN, THE MEAN, AND THE MIDDLE

Earlier, we defined the mean as the "balance point" for a distribution because the distances above the mean must have the same total as the distances below the mean. One consequence of this definition is that the mean is always located inside the group of scores, somewhere between the smallest score and the largest score. You should notice, however, that the concept of a balance point focuses on distances rather than scores. In particular, it is possible to have a distribution in which the vast majority of the scores

BOX 3.2 USING INTERPOLATION TO LOCATE THE 50TH PERCENTILE (THE MEDIAN)

The precise median and the 50th percentile are both defined as the point that separates the top 50% of a distribution from the bottom 50%. In Chapter 2, we introduced interpolation as a technique for finding specific percentiles. We now use that same process to find the 50th percentile for the scores in Example 3.7.

Looking at the distribution of scores shown in Figure 3.5, exactly 3 of the $n = 8$ scores, or 37.5%, are located below the real limit of 3.5. Also, 7 of the $n = 8$ scores (87.5%) are located below the real limit of 4.5. This interval of scores and percentages is shown in the following table. Note that the median, the 50th percentile, is located within this interval.

	Scores (X)	Percentages	
Top	4.5	87.5%	
	?	50%	← Intermediate value
Bottom	3.5	37.5%	

We will find the 50th percentile (the median) using the 4-step interpolation process that was introduced in Chapter 2.

1. For the scores, the width of the interval is 1 point. For the percentages, the width is 50 points.
2. The value of 50% is located 37.5 points down from the top of the percentage interval. As a fraction of the whole interval, this is 37.5 out of 50, or 0.75 of the total interval.
3. For the scores, the interval width is 1 point and 0.75 of the interval corresponds to a distance of $0.75(1) = 0.75$ points.
4. Because the top of the interval is 4.5, the position we want is $4.5 - 0.75 = 3.75$

For this distribution, the 50% point (the 50th percentile) corresponds to a score of $X = 3.75$. Note that this is exactly the same value that we obtained for the median in Example 3.7.

are located on one side of the mean. Figure 3.6 shows a distribution of $N = 6$ scores in which 5 out of 6 scores have values less than the mean. In this figure, the total of the distances above the mean is 8 points and the total of the distances below the mean is 8 points. Thus, the mean is located in the middle of the distribution if you use the concept of distance to define the middle. However, you should realize that the mean is not necessarily located at the exact center of the group of scores.

The median, on the other hand, defines the middle of the distribution in terms of scores. In particular, the median is located so that half of the scores are on one side and half are on the other side. For the distribution in Figure 3.6, for example, the median is located at $X = 2.5$, with exactly 3 scores above this value and exactly 3 scores below. Thus, it is possible to claim that the median is located in the middle of the distribution, provided that the term *middle* is defined by the number of scores.

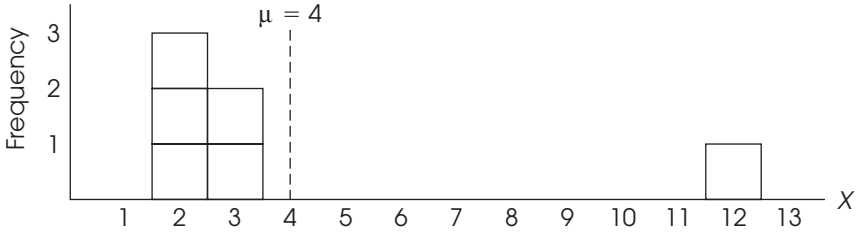


FIGURE 3.6
A population of $N = 6$ scores with a mean of $\mu = 4$. Notice that the mean does not necessarily divide the scores into two equal groups. In this example, 5 out of the 6 scores have values less than the mean.

In summary, the mean and the median are both methods for defining and measuring central tendency. Although they both define the middle of the distribution, they use different definitions of the term *middle*.

LEARNING CHECK

1. Find the median for each distribution of scores:
   a. 3, 4, 6, 7, 9, 10, 11
   b. 8, 10, 11, 12, 14, 15
2. If you have a score of 52 on an 80-point exam, then you definitely scored above the median. (True or false?)
3. The following is a distribution of measurements for a continuous variable. Find the precise median that divides the distribution exactly in half.

   Scores: 1, 2, 2, 3, 4, 4, 4, 4, 4, 5

ANSWERS

1. **a.** The median is $X = 7$. **b.** The median is $X = 11.5$.
2. False. The value of the median would depend on where all of the scores are located.
3. The median is 3.70 (one-fifth of the way into the interval from 3.5 to 4.5).

3.4 THE MODE

The final measure of central tendency that we consider is called the *mode*. In its common usage, the word *mode* means "the customary fashion" or "a popular style." The statistical definition is similar in that the mode is the most common observation among a group of scores.

DEFINITION

In a frequency distribution, the **mode** is the score or category that has the greatest frequency.

As with the median, there are no symbols or special notation used to identify the mode or to differentiate between a sample mode and a population mode. In addition, the definition of the mode is the same for a population and for a sample distribution.

The mode is a useful measure of central tendency because it can be used to determine the typical or average value for any scale of measurement, including a nominal scale (see Chapter 1). Consider, for example, the data shown in Table 3.4. These data were obtained by asking a sample of 100 students to name their favorite restaurants in

TABLE 3.4

Favorite restaurants named by a sample of $n = 100$ students. *Caution:* The mode is a score or category, not a frequency. For this example, the mode is Luigi's, not $f = 42$.

Restaurant	*f*
College Grill	5
George & Harry's	16
Luigi's	42
Oasis Diner	18
Roxbury Inn	7
Sutter's Mill	12

town. The result is a sample of $n = 100$ scores with each score corresponding to the restaurant that the student named.

For these data, the mode is Luigi's, the restaurant (score) that was named most frequently as a favorite place. Although we can identify a modal response for these data, you should notice that it would be impossible to compute a mean or a median. For example, you cannot add the scores to determine a mean (How much is 5 College Grills plus 42 Luigi's?). Also, it is impossible to list the scores in order because the restaurants do not form any natural order. For example, the College Grill is not "more than" or "less than" the Oasis Diner, they are simply two different restaurants. Thus, it is impossible to obtain the median by finding the midpoint of the list. In general, the mode is the only measure of central tendency that can be used with data from a nominal scale of measurement.

The mode also can be useful because it is the only measure of central tendency that corresponds to an actual score in the data; by definition, the mode is the most frequently occurring score. The mean and the median, on the other hand, are both calculated values and often produce an answer that does not equal any score in the distribution. For example, in Figure 3.6 (p. 86) we presented a distribution with a mean of 4 and a median of 2.5. Note that none of the scores is equal to 4 and none of the scores is equal to 2.5. However, the mode for this distribution is $X = 2$; there are three individuals who actually have scores of $X = 2$.

In a frequency distribution graph, the greatest frequency appears as the tallest part of the figure. To find the mode, you simply identify the score located directly beneath the highest point in the distribution.

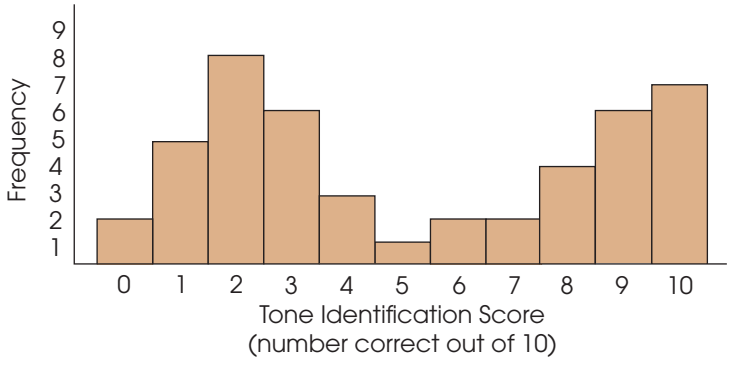
Although a distribution has only one mean and only one median, it is possible to have more than one mode. Specifically, it is possible to have two or more scores that have the same highest frequency. In a frequency distribution graph, the different modes correspond to distinct, equally high peaks. A distribution with two modes is said to be *bimodal,* and a distribution with more than two modes is called *multimodal*. Occasionally, a distribution with several equally high points is said to have no mode.

Incidentally, a bimodal distribution is often an indication that two separate and distinct groups of individuals exist within the same population (or sample). For example, if you measured height for each person in a set of 100 college students, the resulting distribution would probably have two modes, one corresponding primarily to the males in the group and one corresponding primarily to the females.

Technically, the mode is the score with the absolute highest frequency. However, the term *mode* is often used more casually to refer to scores with relatively high frequencies—that is, scores that correspond to peaks in a distribution even though the peaks are not the absolute highest points. For example, Athos, et al. (2007) asked people to identify the pitch for both pure tones and piano tones. Participants were presented with a series of tones and had to name the note corresponding to each tone. Nearly half the participants (44%) had extraordinary pitch-naming ability (absolute pitch), and were able to identify most of the tones correctly. Most of the other participants performed around chance level, apparently guessing the pitch names randomly. Figure 3.7 shows a distribution of scores that is consistent with the results of the study. There are two distinct peaks in the distribution, one located at $X = 2$ (chance performance) and the other located at $X = 10$ (perfect performance). Each of these values is a mode in the distribution. Note, however, that the two modes do not have identical frequencies. Eight people scored at $X = 2$ and only seven had scores of $X = 10$. Nonetheless, both of these points are called modes. When two modes have unequal frequencies, researchers occasionally differentiate the two values by calling the taller peak the *major mode,* and the shorter one the *minor mode*.

FIGURE 3.7

A frequency distribution for tone identification scores. An example of bimodal distributions.



LEARNING CHECK

1. During the month of October, an instructor recorded the number of absences for each student in a class of $n = 20$ and obtained the following distribution.

Number of Absences	*f*
5	1
4	2
3	7
2	5
1	3
0	2

a. Using the mean, what is the average number of absences for the class?

b. Using the median, what is the average number of absences for the class?

c. Using the mode, what is the average number of absences for the class?

ANSWERS

1.

a. The mean is 47/20 = 2.35.

b. The median is 2.5.

c. The mode is 3.

3.5 SELECTING A MEASURE OF CENTRAL TENDENCY

How do you decide which measure of central tendency to use? The answer to this question depends on several factors. Before we discuss these factors, however, note that you usually can compute two or even three measures of central tendency for the same set of data. Although the three measures often produce similar results, there are situations in which they are very different (see Section 3.6). Also note that the mean is usually the preferred measure of central tendency. Because the mean uses every score in the distribution, it typically produces a good representative value. Remember that the goal of central tendency is to find the single value that best represents the entire distribution. Besides being a good representative, the mean has the added advantage of being closely

related to variance and standard deviation, the most common measures of variability (Chapter 4). This relationship makes the mean a valuable measure for purposes of inferential statistics. For these reasons, and others, the mean generally is considered to be the best of the three measures of central tendency. But there are specific situations in which it is impossible to compute a mean or in which the mean is not particularly representative. It is in these situations that the mode and the median are used.

WHEN TO USE THE MEDIAN

We consider four situations in which the median serves as a valuable alternative to the mean. In the first three cases, the data consist of numerical values (interval or ratio scales) for which you would normally compute the mean. However, each case also involves a special problem so that it is either impossible to compute the mean, or the calculation of the mean produces a value that is not central or not representative of the distribution. The fourth situation involves measuring central tendency for ordinal data.

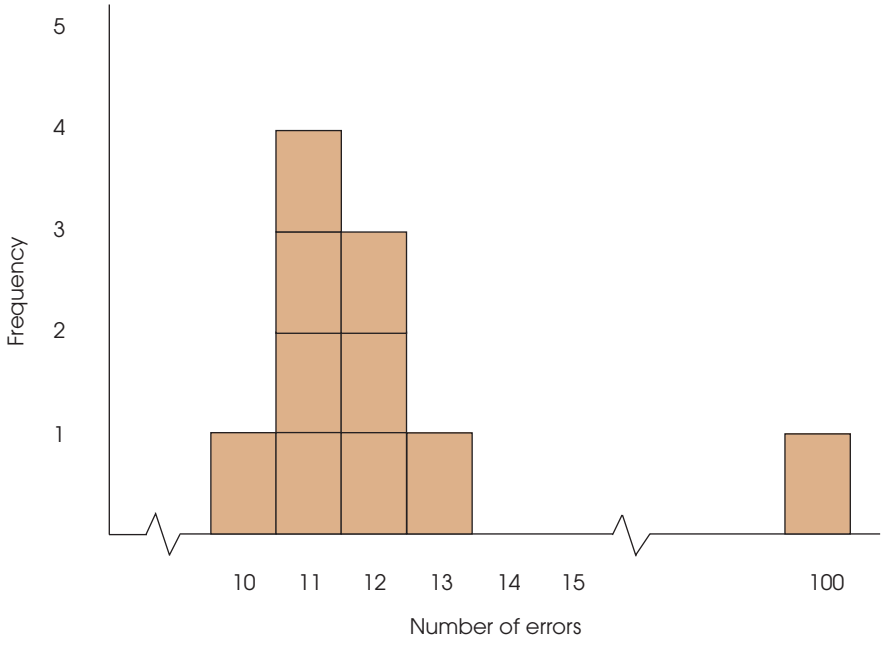
Extreme scores or skewed distributions When a distribution has a few extreme scores, scores that are very different in value from most of the others, then the mean may not be a good representative of the majority of the distribution. The problem comes from the fact that one or two extreme values can have a large influence and cause the mean to be displaced. In this situation, the fact that the mean uses all of the scores equally can be a disadvantage. Consider, for example, the distribution of $n = 10$ scores in Figure 3.8. For this sample, the mean is

$$M = \frac{\Sigma X}{n} = \frac{203}{10} = 20.3$$

Notice that the mean is not very representative of any score in this distribution. Although most of the scores are clustered between 10 and 13, the extreme score of $X = 100$ inflates the value of ΣX and distorts the mean.

FIGURE 3.8

Frequency distribution of errors committed before reaching learning criterion. Notice that the graph shows two *breaks* in the *X*-axis. Rather than listing all the scores from 0 to 100, the graph jumps directly to the first score, which is $X = 10$, and then jumps directly from $X = 15$ to $X = 100$. The breaks shown in the *X*-axis are the conventional way of notifying the reader that some values have been omitted.



The median, on the other hand, is not easily affected by extreme scores. For this sample, $n = 10$, so there should be five scores on either side of the median. The median is 11.50. Notice that this is a very representative value. Also note that the median would be unchanged even if the extreme score were 1000 instead of only 100. Because it is relatively unaffected by extreme scores, the median commonly is used when reporting the average value for a skewed distribution. For example, the distribution of personal incomes is very skewed, with a small segment of the population earning incomes that are astronomical. These extreme values distort the mean, so that it is not very representative of the salaries that most of us earn. The median is the preferred measure of central tendency when extreme scores exist.

Undetermined values Occasionally, you encounter a situation in which an individual has an unknown or undetermined score. In psychology, this often occurs in learning experiments in which you are measuring the number of errors (or amount of time) required for an individual to solve a particular problem. For example, suppose that participants are asked to assemble a wooden puzzle as quickly as possible. The experimenter records how long (in minutes) it takes each individual to arrange all of the pieces to complete the puzzle. Table 3.5 presents results for a sample of $n = 6$ people.

TABLE 3.5
Number of minutes needed to assemble a wooden puzzle.

Person	Time (Min.)
1	8
2	11
3	12
4	13
5	17
6	Never finished

Notice that person 6 never completed the puzzle. After an hour, this person still showed no sign of solving the puzzle, so the experimenter stopped him or her. This person has an undetermined score. (There are two important points to be noted. First, the experimenter should not throw out this individual's score. The whole purpose for using a sample is to gain a picture of the population, and this individual tells us that part of the population cannot solve the puzzle. Second, this person should not be given a score of $X = 60$ minutes. Even though the experimenter stopped the individual after 1 hour, the person did not finish the puzzle. The score that is recorded is the amount of time needed to finish. For this individual, we do not know how long this is.)

It is impossible to compute the mean for these data because of the undetermined value. We cannot calculate the ΣX part of the formula for the mean. However, it is possible to determine the median. For these data, the median is 12.5. Three scores are below the median, and three scores (including the undetermined value) are above the median.

Number of Pizzas (X)	f
5 or more	3
4	2
3	2
2	3
1	6
0	4

Open-ended distributions A distribution is said to be *open-ended* when there is no upper limit (or lower limit) for one of the categories. The table in the margin provides an example of an open-ended distribution, showing the number of pizzas eaten during a 1 month period for a sample of $n = 20$ high school students. The top category in this distribution shows that three of the students consumed "5 or more" pizzas. This is an open-ended category. Notice that it is impossible to compute a mean for these data because you cannot find ΣX (the total number of pizzas for all 20 students). However, you can find the median. Listing the 20 scores in order produces $X = 1$ and $X = 2$ as the middle two scores. For these data, the median is 1.5.

Ordinal scale Many researchers believe that it is not appropriate to use the mean to describe central tendency for ordinal data. When scores are measured on an ordinal scale, the median is always appropriate and is usually the preferred measure of central tendency.

You should recall that ordinal measurements allow you to determine direction (greater than or less than) but do not allow you to determine distance. The median is compatible with this type of measurement because it is defined by direction: half of the scores are above the median and half are below the median. The mean, on the other hand, defines central tendency in terms of distance. Remember that the mean is the balance point for the distribution, so that the distances above the mean are exactly balanced by the distances below the mean. Because the mean is defined in terms of distances, and because ordinal scales do not measure distance, it is not appropriate to compute a mean for scores from an ordinal scale.

WHEN TO USE THE MODE

We consider three situations in which the mode is commonly used as an alternative to the mean, or is used in conjunction with the mean to describe central tendency.

Nominal scales The primary advantage of the mode is that it can be used to measure and describe central tendency for data that are measured on a nominal scale. Recall that the categories that make up a nominal scale are differentiated only by name. Because nominal scales do not measure quantity (distance or direction), it is impossible to compute a mean or a median for data from a nominal scale. Therefore, the mode is the only option for describing central tendency for nominal data.

Discrete variables Recall that discrete variables are those that exist only in whole, indivisible categories. Often, discrete variables are numerical values, such as the number of children in a family or the number of rooms in a house. When these variables produce numerical scores, it is possible to calculate means. In this situation, the calculated means are usually fractional values that cannot actually exist. For example, computing means generates results such as "the average family has 2.4 children and a house with 5.33 rooms." On the other hand, the mode always identifies the most typical case and, therefore, it produces more sensible measures of central tendency. Using the mode, our conclusion would be "the typical, or modal, family has 2 children and a house with 5 rooms." In many situations, especially with discrete variables, people are more comfortable using the realistic, whole-number values produced by the mode.

Describing shape Because the mode requires little or no calculation, it is often included as a supplementary measure along with the mean or median as a no-cost extra. The value of the mode (or modes) in this situation is that it gives an indication of the shape of the distribution as well as a measure of central tendency. Remember that the mode identifies the location of the peak (or peaks) in the frequency distribution graph. For example, if you are told that a set of exam scores has a mean of 72 and a mode of 80, you should have a better picture of the distribution than would be available from the mean alone (see Section 3.6).

IN THE LITERATURE

REPORTING MEASURES OF CENTRAL TENDENCY

Measures of central tendency are commonly used in the behavioral sciences to summarize and describe the results of a research study. For example, a researcher may report the sample means from two different treatments or the median score for a

large sample. These values may be reported in verbal descriptions of the results, in tables, or in graphs.

In reporting results, many behavioral science journals use guidelines adopted by the American Psychological Association (APA), as outlined in the *Publication Manual of the American Psychological Association* (2010). We refer to the APA manual from time to time in describing how data and research results are reported in the scientific literature. The APA style uses the letter M as the symbol for the sample mean. Thus, a study might state:

> The treatment group showed fewer errors ($M = 2.56$) on the task than the control group ($M = 11.76$).

When there are many means to report, tables with headings provide an organized and more easily understood presentation. Table 3.6 illustrates this point.

The median can be reported using the abbreviation *Mdn,* as in "Mdn = 8.5 errors," or it can simply be reported in narrative text, as follows:

> The median number of errors for the treatment group was 8.5, compared to a median of 13 for the control group.

There is no special symbol or convention for reporting the mode. If mentioned at all, the mode is usually just reported in narrative text.

PRESENTING MEANS AND MEDIANS IN GRAPHS

Graphs also can be used to report and compare measures of central tendency. Usually, graphs are used to display values obtained for sample means, but occasionally sample medians are reported in graphs (modes are rarely, if ever, shown in a graph). The value of a graph is that it allows several means (or medians) to be shown simultaneously, so it is possible to make quick comparisons between groups or treatment conditions. When preparing a graph, it is customary to list the different groups or treatment conditions on the horizontal axis. Typically, these are the different values that make up the independent variable or the quasi-independent variable. Values for the dependent variable (the scores) are listed on the vertical axis. The means (or medians) are then displayed using a *line graph,* a *histogram,* or a *bar graph,* depending on the scale of measurement used for the independent variable.

Figure 3.9 shows an example of a *line graph* displaying the relationship between drug dose (the independent variable) and food consumption (the dependent variable). In this study, there were five different drug doses (treatment conditions) and they are listed along the horizontal axis. The five means appear as points in the graph. To construct this graph, a point was placed above each treatment condition so that the vertical position of the point corresponds to the mean score for the treatment condition. The points are then connected with straight lines. A line graph is used when the values on the horizontal axis are measured on an interval or a ratio scale. An alternative to the line graph is a *histogram*. For this example, the histogram would show a bar above each drug dose so that the height of each bar corresponds to the mean food consumption for that group, with no space between adjacent bars.

TABLE 3.6

The mean number of errors made on the task for treatment and control groups, divided by gender.

	Treatment	Control
Females	1.45	8.36
Males	3.83	14.77

FIGURE 3.9

The relationship between an independent variable (drug dose) and a dependent variable (food consumption). Because drug dose is a continuous variable, a continuous line is used to connect the different dose levels.

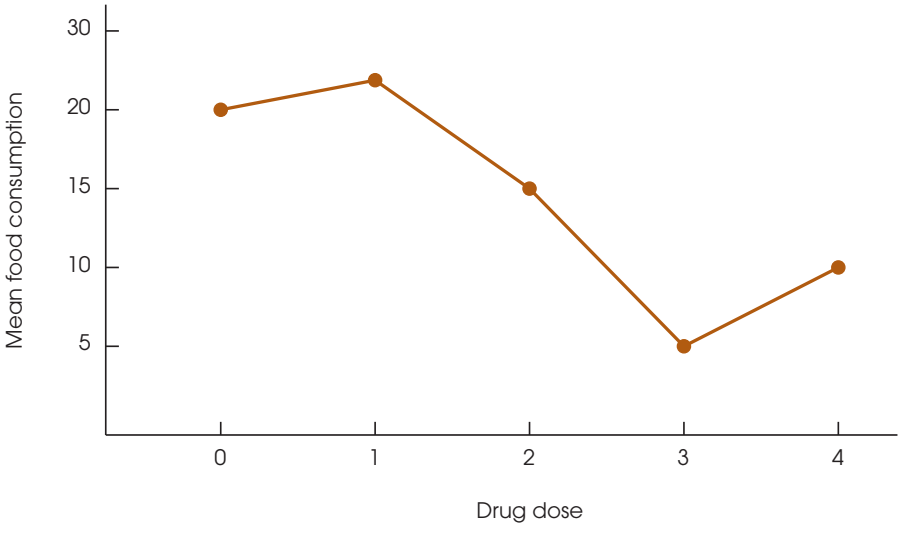


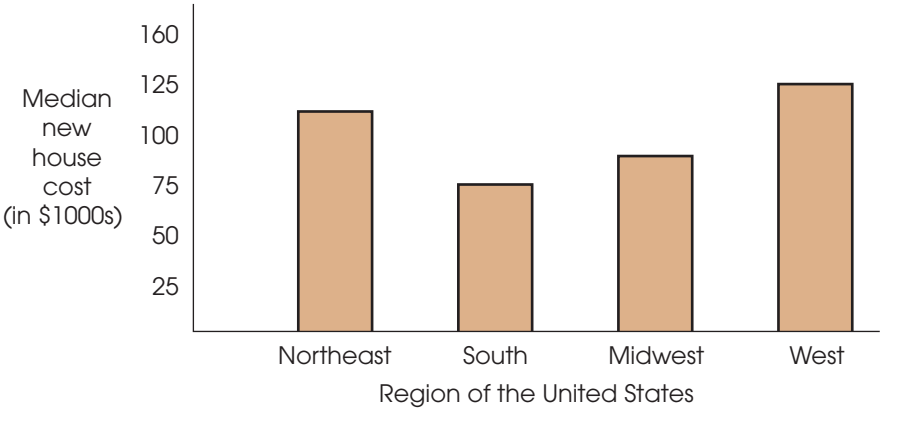
Figure 3.10 shows a bar graph displaying the median selling price for single-family homes in different regions of the United States. Bar graphs are used to present means (or medians) when the groups or treatments shown on the horizontal axis are measured on a nominal or an ordinal scale. To construct a bar graph, you simply draw a bar directly above each group or treatment so that the height of the bar corresponds to the mean (or median) for that group or treatment. For a bar graph, a space is left between adjacent bars to indicate that the scale of measurement is nominal or ordinal.

When constructing graphs of any type, you should recall the basic rules that we introduced in Chapter 2:

1. The height of a graph should be approximately two-thirds to three-quarters of its length.
2. Normally, you start numbering both the *X*-axis and the *Y*-axis with zero at the point where the two axes intersect. However, when a value of zero is part of the

FIGURE 3.10

Median cost of a new, single-family home by region.



data, it is common to move the zero point away from the intersection so that the graph does not overlap the axes (see Figure 3.9).

Following these rules helps to produce a graph that provides an accurate presentation of the information in a set of data. Although it is possible to construct graphs that distort the results of a study (see Box 2.1), researchers have an ethical responsibility to present an honest and accurate report of their research results. ❑

3.6 CENTRAL TENDENCY AND THE SHAPE OF THE DISTRIBUTION

We have identified three different measures of central tendency, and often a researcher calculates all three for a single set of data. Because the mean, the median, and the mode are all trying to measure the same thing, it is reasonable to expect that these three values should be related. In fact, there are some consistent and predictable relationships among the three measures of central tendency. Specifically, there are situations in which all three measures have exactly the same value. On the other hand, there are situations in which the three measures are guaranteed to be different. In part, the relationships among the mean, median, and mode are determined by the shape of the distribution. We consider two general types of distributions.

SYMMETRICAL DISTRIBUTIONS

For a *symmetrical distribution,* the right-hand side of the graph is a mirror image of the left-hand side. If a distribution is perfectly symmetrical, the median is exactly at the center because exactly half of the area in the graph is on either side of the center. The mean also is exactly at the center of a perfectly symmetrical distribution because each score on the left side of the distribution is balanced by a corresponding score (the mirror image) on the right side. As a result, the mean (the balance point) is located at the center of the distribution. Thus, for a perfectly symmetrical distribution, the mean and the median are the same (Figure 3.11). If a distribution is roughly symmetrical, but not perfect, the mean and median are close together in the center of the distribution.

If a symmetrical distribution has only one mode, it is also in the center of the distribution. Thus, for a perfectly symmetrical distribution with one mode, all three measures of central tendency, the mean, the median, and the mode, have the same value. For a roughly symmetrical distribution, the three measures are clustered together in the center of the distribution. On the other hand, a bimodal distribution that is symmetrical [see Figure 3.11(b)] has the mean and median together in the center with the modes on each side. A rectangular distribution [see Figure 3.11(c)] has no mode because all X values occur with the same frequency. Still, the mean and the median are in the center of the distribution.

SKEWED DISTRIBUTIONS

The positions of the mean, median, and mode are not as consistently predictable in distributions of discrete variables (see Von Hippel, 2005).

In *skewed distributions,* especially distributions for continuous variables, there is a strong tendency for the mean, median, and mode to be located in predictably different positions. Figure 3.12(a), for example, shows a positively skewed distribution with the peak (highest frequency) on the left-hand side. This is the position of the mode. However, it should be clear that the vertical line drawn at the mode does not divide the distribution into two equal parts. To have exactly 50% of the distribution

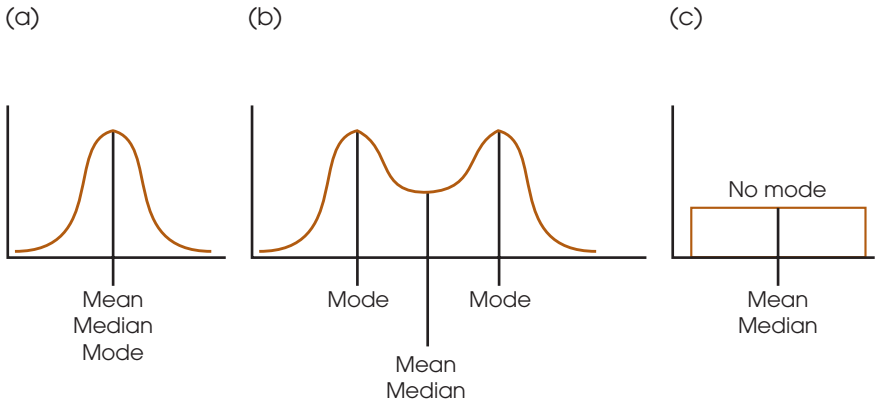


FIGURE 3.11
Measures of central tendency for three symmetrical distributions: normal, bimodal, and rectangular.

on each side, the median must be located to the right of the mode. Finally, the mean is located to the right of the median because it is the measure influenced most by the extreme scores in the tail and is displaced farthest to the right toward the tail of the distribution. Therefore, in a positively skewed distribution, the order of the three measures of central tendency from smallest to largest (left to right) is the mode, the median, and the mean.

Negatively skewed distributions are lopsided in the opposite direction, with the scores piling up on the right-hand side and the tail tapering off to the left. The grades on an easy exam, for example, tend to form a negatively skewed distribution [see Figure 3.12(b)]. For a distribution with negative skew, the mode is on the right-hand side (with the peak), whereas the mean is displaced toward the left by the extreme scores in the tail. As before, the median is located between the mean and the mode. In order from smallest value to largest value (left to right), the three measures of central tendency for a negatively skewed distribution are the mean, the median, and the mode.

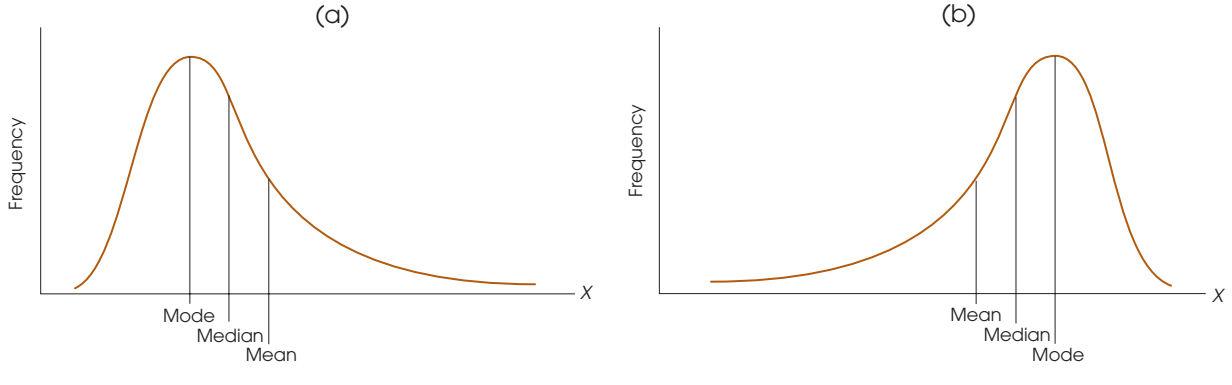


FIGURE 3.12
Measures of central tendency for skewed distributions.

LEARNING CHECK

1. Which measure of central tendency is most affected if one extremely large score is added to a distribution? (mean, median, mode)
2. Why is it usually considered inappropriate to compute a mean for scores measured on an ordinal scale?
3. In a perfectly symmetrical distribution, the mean, the median, and the mode will all have the same value. (True or false?)
4. A distribution with a mean of 70 and a median of 75 is probably positively skewed. (True or false?)

ANSWERS

1. mean
2. The definition of the mean is based on distances (the mean balances the distances) and ordinal scales do not measure distance.
3. False, if the distribution is bimodal.
4. False. The mean is displaced toward the tail on the left-hand side.

SUMMARY

1. The purpose of central tendency is to determine the single value that identifies the center of the distribution and best represents the entire set of scores. The three standard measures of central tendency are the mode, the median, and the mean.
2. The mean is the arithmetic average. It is computed by adding all of the scores and then dividing by the number of scores. Conceptually, the mean is obtained by dividing the total (ΣX) equally among the number of individuals (N or n). The mean can also be defined as the balance point for the distribution. The distances above the mean are exactly balanced by the distances below the mean. Although the calculation is the same for a population or a sample mean, a population mean is identified by the symbol μ, and a sample mean is identified by M. In most situations with numerical scores from an interval or a ratio scale, the mean is the preferred measure of central tendency.
3. Changing any score in the distribution causes the mean to be changed. When a constant value is added to (or subtracted from) every score in a distribution, the same constant value is added to (or subtracted from) the mean. If every score is multiplied by a constant, the mean is multiplied by the same constant.
4. The median is the midpoint of a distribution of scores. The median is the preferred measure of central tendency when a distribution has a few extreme scores that displace the value of the mean. The median also is used for open-ended distributions and when there are undetermined (infinite) scores that make it impossible to compute a mean. Finally, the median is the preferred measure of central tendency for data from an ordinal scale.
5. The mode is the most frequently occurring score in a distribution. It is easily located by finding the peak in a frequency distribution graph. For data measured on a nominal scale, the mode is the appropriate measure of central tendency. It is possible for a distribution to have more than one mode.
6. For symmetrical distributions, the mean is equal to the median. If there is only one mode, then it has the same value, too.
7. For skewed distributions, the mode is located toward the side where the scores pile up, and the mean is pulled toward the extreme scores in the tail. The median is usually located between these two values.

KEY TERMS

central tendency (73)
population mean (μ) (75)
sample mean (M) (75)
weighted mean (77)
median (83)
mode (87)
bimodal (88)
multimodal (88)
major mode (88)
minor mode (88)
line graph (93)
symmetrical distribution (95)
skewed distribution (95)
positive skew (95)
negative skew (96)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 3 on the book companion website. The website also includes a workshop entitled *Central Tendency and Variability* that reviews the basic concept of the mean and introduces the concept of variability that is presented in Chapter 4.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to compute the **Mean** and ΣX for a set of scores.

Data Entry

1. Enter all of the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Descriptives.**
2. Highlight the column label for the set of scores (VAR00001) in the left box and click the arrow to move it into the **Variable** box.
3. If you want ΣX as well as the mean, click on the **Options** box, select **Sum,** then click **Continue.**
4. Click **OK.**

SPSS Output

SPSS produces a summary table listing the number of scores (N), the maximum and minimum scores, the sum of the scores (if you selected this option), the mean, and the standard deviation. Note: The standard deviation is a measure of variability that is presented in Chapter 4.

FOCUS ON PROBLEM SOLVING

X	f
4	1
3	4
2	3
1	2

1. Although the three measures of central tendency appear to be very simple to calculate, there is always a chance for errors. The most common sources of error are listed next.
   - **a.** Many students find it very difficult to compute the mean for data presented in a frequency distribution table. They tend to ignore the frequencies in the table and simply average the score values listed in the X column. You must use the frequencies *and* the scores! Remember that the number of scores is found by $N = \Sigma f$, and the sum of all N scores is found by ΣfX. For the distribution shown in the margin, the mean is $\frac{24}{10} = 2.40$.
   - **b.** The median is the midpoint of the distribution of scores, not the midpoint of the scale of measurement. For a 100-point test, for example, many students incorrectly assume that the median must be $X = 50$. To find the median, you must have the *complete set* of individual scores. The median separates the individuals into two equal-sized groups.
   - **c.** The most common error with the mode is for students to report the highest frequency in a distribution rather than the *score* with the highest frequency. Remember that the purpose of central tendency is to find the most representative score. For the distribution in the margin, the mode is $X = 3$, not $f = 4$.

DEMONSTRATION 3.1

COMPUTING MEASURES OF CENTRAL TENDENCY

For the following sample, find the mean, the median, and the mode. The scores are:

5, 6, 9, 11, 5, 11, 8, 14, 2, 11

Compute the mean The calculation of the mean requires two pieces of information: the sum of the scores, ΣX; and the number of scores, n. For this sample, $n = 10$ and

$$\Sigma X = 5 + 6 + 9 + 11 + 5 + 11 + 8 + 14 + 2 + 11 = 82$$

Therefore, the sample mean is

$$M = \frac{\Sigma X}{n} = \frac{82}{10} = 8.2$$

Find the median To find the median, first list the scores in order from smallest to largest. With an even number of scores, the median is the average of the middle two scores in the list. Listed in order, the scores are:

2, 5, 5, 6, 8, 9, 11, 11, 11, 14

The middle two scores are 8 and 9, and the median is 8.5.

Find the mode For this sample, $X = 11$ is the score that occurs most frequently. The mode is $X = 11$.

PROBLEMS

1. What general purpose is served by a good measure of central tendency?

2. Why is it necessary to have more than one method for measuring central tendency?

3. Find the mean, median, and mode for the following sample of scores:

6, 2, 4, 1, 2, 2, 3, 4, 3, 2

4. Find the mean, median, and mode for the following sample of scores:

8, 7, 8, 8, 4, 9, 10, 7, 8, 8, 9, 8

5. Find the mean, median, and mode for the scores in the following frequency distribution table:

X	f
8	1
7	4
6	2
5	2
4	2
3	1

6. Find the mean, median, and mode for the scores in the following frequency distribution table:

X	f
10	1
9	2
8	3
7	3
6	4
5	2

7. For the following sample

a. Assume that the scores are measurements of a continuous variable and find the median by locating the precise midpoint of the distribution.

b. Assume that the scores are measurements of a discrete variable and find the median.

Scores: 1, 2, 3, 3, 3, 4

8. A sample of $n = 7$ scores has a mean of $M = 9$. What is the value of ΣX for this sample?

9. A population with a mean of $\mu = 10$ has $\Sigma X = 250$. How many scores are in the population?

10. A sample of $n = 8$ scores has a mean of $M = 10$. If one new person with a score of $X = 1$ is added to the sample, what is the value for the new mean?

11. A sample of $n = 5$ scores has a mean of $M = 12$. If one person with a score of $X = 8$ is removed from the sample, what is the value for the new mean?

12. A sample of $n = 11$ scores has a mean of $M = 4$. One person with a score of $X = 16$ is added to the sample. What is the value for the new sample mean?

13. A sample of $n = 9$ scores has a mean of $M = 10$. One person with a score of $X = 2$ is removed from the sample. What is the value for the new sample mean?

14. A population of $N = 20$ scores has a mean of $\mu = 15$. One score in the population is changed from $X = 8$ to $X = 28$. What is the value for the new population mean?

15. A sample of $n = 7$ scores has a mean of $M = 9$. One score in the sample is changed from $X = 19$ to $X = 5$. What is the value for the new sample mean?

16. A sample of $n = 7$ scores has a mean of $M = 5$. After one new score is added to the sample, the new mean is found to be $M = 6$. What is the value of the new score? (Hint: Compare the values for ΣX before and after the score was added.)

17. A population of $N = 16$ scores has a mean of $\mu = 20$. After one score is removed from the population, the new mean is found to be $\mu = 19$. What is the value of the score that was removed? (Hint: Compare the values for ΣX before and after the score was removed.)

18. One sample has a mean of $M = 4$ and a second sample has a mean of $M = 8$. The two samples are combined into a single set of scores.

a. What is the mean for the combined set if both of the original samples have $n = 7$ scores?

b. What is the mean for the combined set if the first sample has $n = 3$ and the second sample has $n = 7$?

c. What is the mean for the combined set if the first sample has $n = 7$ and the second sample has $n = 3$?

19. One sample has a mean of $M = 5$ and a second sample has a mean of $M = 10$. The two samples are combined into a single set of scores.

a. What is the mean for the combined set if both of the original samples have $n = 5$ scores?

b. What is the mean for the combined set if the first sample has $n = 4$ scores and the second sample has $n = 6$?

c. What is the mean for the combined set if the first sample has $n = 6$ scores and the second sample has $n = 4$?

20. Explain why the mean is often not a good measure of central tendency for a skewed distribution.

21. Identify the circumstances in which the median rather than the mean is the preferred measure of central tendency.

22. For each of the following situations, identify the measure of central tendency (mean, median, or mode) that would provide the best description of the average score:

a. A news reporter interviewed people shopping in a local mall and asked how much they spent on summer vacations. Most people traveled locally and reported modest amounts but one couple had flown to Paris for a month and paid a small fortune.

b. A marketing researcher asked consumers to select their favorite from a set of four designs for a new product logo.

c. A driving instructor recorded the number of orange cones that each student ran over during the first attempt at parallel parking.

23. One question on a student survey asks: In a typical week, how many times do you eat at a fast-food restaurant? The following frequency distribution table summarizes the results for a sample of $n = 20$ students.

Number of times per week	f
5 or more	2
4	2
3	3
2	6
1	4
0	3

a. Find the mode for this distribution.

b. Find the median for the distribution.

c. Explain why you cannot compute the mean using the data in the table.

24. A nutritionist studying weight gain for college freshmen obtains a sample of $n = 20$ first-year students at the state college. Each student is weighed on the first day of school and again on the last day of the semester. The following scores measure the change in weight, in pounds, for each student. A positive score indicates a weight gain during the semester.

+5, +6, +3, +1, +8, +5, +4, +4, +3, –1
+2, +7, +1, +5, +8, 0, +4, +6, +5, +3

a. Sketch a histogram showing the distribution of weight-change scores.
b. Calculate the mean weight-change score for this sample.
c. Does there appear to be a consistent trend in weight change during the semester?

25. Does it ever seem to you that the weather is nice during the work week, but lousy on the weekend? Cerveny and Balling (1998) have confirmed that this is not your imagination—pollution accumulating during the work week most likely spoils the weekend weather for people on the Atlantic coast. Consider the following hypothetical data showing the daily amount of rainfall for 10 weeks during the summer.

Week	Average Daily Rainfall on Weekdays (Mon.–Fri.)	Average Daily Rainfall on Weekends (Sat.–Sun.)
1	1.2	1.5
2	0.6	2.0
3	0.0	1.8
4	1.6	1.5
5	0.8	2.2
6	2.1	2.4
7	0.2	0.8
8	0.9	1.6
9	1.1	1.2
10	1.4	1.7

a. Calculate the average daily rainfall (the mean) during the week, and the average daily rainfall for weekends.
b. Based on the two means, does there appear to be a pattern in the data?

CHAPTER

4

Variability

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Summation notation (Chapter 1)
- Central tendency (Chapter 3)
  - Mean
  - Median

Preview

Although measures of central tendency, such as the mean and median, are handy ways to summarize large sets of data, these measures do not tell the whole story. Specifically, not everyone is average. Many people may perform near average, but others demonstrate performance that is far above (or below) average. In simple terms, people are different.

The differences that exist from one person to another are often called *diversity*. Researchers comparing cognitive skills for younger adults and older adults typically find that differences between people tend to increase as people age. For example, Morse (1993) reviewed hundreds of research studies published in *Psychology and Aging* and in the *Journal of Gerontology* from 1986 to 1990, and found increased diversity in older adults on measures of reaction time, memory, and some measures of intelligence. One possible explanation for the increased diversity is that different people respond differently to the aging process; some are essentially unchanged and others show a rapid decline. As a result, the differences from one person to another are larger for older people than for those who are younger.

It also is possible to measure differences in performance for the same person. These differences provide a measure of *consistency*. Often, large differences from trial to trial for the same person are viewed as evidence of poor performance. For example, the ability to consistently hit a target is an indication of skilled performance in many sports, whereas inconsistent performance indicates a lack of skill. Researchers in the field of aging have also found that older participants tend to have larger differences from trial to trial than younger participants. That is, older people seem to lose the ability to perform consistently on many tasks. For example, in a study comparing older and younger women, Wegesin and Stern (2004) found lower consistency for older women on a recognition memory task.

The Problem: To study phenomena such as diversity and consistency, it is necessary to devise a method to measure and objectively describe the differences that exist from one score to another within a distribution.

The Solution: A measure of *variability* provides an objective description of the differences between the scores in a distribution by measuring the degree to which the scores are spread out or are clustered together.

4.1 OVERVIEW

The term *variability* has much the same meaning in statistics as it has in everyday language; to say that things are variable means that they are not all the same. In statistics, our goal is to measure the amount of variability for a particular set of scores, a distribution. In simple terms, if the scores in a distribution are all the same, then there is no variability. If there are small differences between scores, then the variability is small, and if there are large differences between scores, then the variability is large.

DEFINITION

Variability provides a quantitative measure of the differences between scores in a distribution and describes the degree to which the scores are spread out or clustered together.

Figure 4.1 shows two distributions of familiar values for the population of adult males: Part (a) shows the distribution of men's heights (in inches), and part (b) shows the distribution of men's weights (in pounds). Notice that the two distributions differ in terms of central tendency. The mean height is 70 inches (5 feet, 10 inches) and the mean weight is 170 pounds. In addition, notice that the distributions differ in terms of variability. For example, most heights are clustered close together, within 5 or 6 inches of

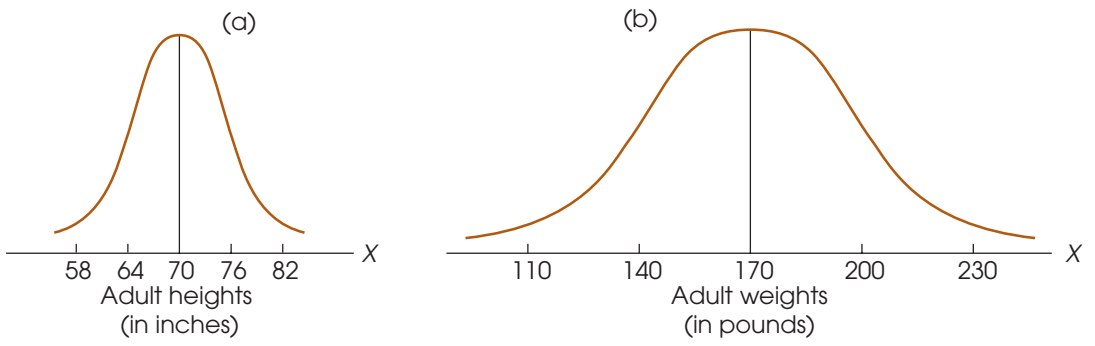


FIGURE 4.1

Population distributions of adult heights and adult weights.

the mean. On the other hand, weights are spread over a much wider range. In the weight distribution it is not unusual to find individuals who are located more than 30 pounds away from the mean, and it would not be surprising to find two individuals whose weights differ by more than 30 or 40 pounds. The purpose for measuring variability is to obtain an objective measure of how the scores are spread out in a distribution. In general, a good measure of variability serves two purposes:

1. Variability describes the distribution. Specifically, it tells whether the scores are clustered close together or are spread out over a large distance. Usually, variability is defined in terms of *distance*. It tells how much distance to expect between one score and another, or how much distance to expect between an individual score and the mean. For example, we know that the heights for most adult males are clustered close together, within 5 or 6 inches of the average. Although more extreme heights exist, they are relatively rare.
2. Variability measures how well an individual score (or group of scores) represents the entire distribution. This aspect of variability is very important for inferential statistics, in which relatively small samples are used to answer questions about populations. For example, suppose that you selected a sample of one person to represent the entire population. Because most adult males have heights that are within a few inches of the population average (the distances are small), there is a very good chance that you would select someone whose height is within 6 inches of the population mean. On the other hand, the scores are much more spread out (greater distances) in the distribution of weights. In this case, you probably would *not* obtain someone whose weight was within 6 pounds of the population mean. Thus, variability provides information about how much error to expect if you are using a sample to represent a population.

In this chapter, we consider three different measures of variability: the range, standard deviation, and the variance. Of these three, the standard deviation and the related measure of variance are by far the most important.

4.2 THE RANGE

The *range* is the distance covered by the scores in a distribution, from the smallest score to the largest score. When the scores are measurements of a continuous variable, the range can be defined as the difference between the upper real limit (URL) for the largest score (X_{max}) and the lower real limit (LRL) for the smallest score (X_{min}).

$$\text{range} = \text{URL for } X_{max} - \text{LRL for } X_{min}$$

If the scores have values from 1 to 5, for example, the range is $5.5 - 0.5 = 5$ points. When the scores are whole numbers, the range is also a measure of the number of measurement categories. If every individual is classified as either 1, 2, 3, 4, or 5, then there are five measurement categories and the range is 5 points.

Continuous and discrete variables were discussed in Chapter 1 on pages 21–22.

Defining the range as the number of measurement categories also works for discrete variables that are measured with numerical scores. For example, if you are measuring the number of children in a family and the data produce values from 0 to 4, then there are five measurement categories (0, 1, 2, 3, and 4) and the range is 5 points. By this definition, when the scores are all whole numbers, the range can be obtained by

$$X_{max} - X_{min} + 1.$$

A commonly used alternative definition of the range simply measures the difference between the largest score (X_{max}) and the smallest score (X_{min}), without any reference to real limits.

$$\text{range} = X_{max} - X_{max}$$

By this definition, scores having values from 1 to 5 cover a range of only 4 points. Many computer programs, such as SPSS, use this definition. For discrete variables, which do not have real limits, this definition is often considered more appropriate. Also, this definition works well for variables with precisely defined upper and lower boundaries. For example, if you are measuring proportions of an object, like pieces of a pizza, you can obtain values such as $\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and so on. Expressed as decimal values, the proportions range from 0 to 1. You can never have a value less than 0 (none of the pizza) and you can never have a value greater than 1 (all of the pizza). Thus, the complete set of proportions is bounded by 0 at one end and by 1 at the other. As a result, the proportions cover a range of 1 point.

Using either definition, the range is probably the most obvious way to describe how spread out the scores are—simply find the distance between the maximum and the minimum scores. The problem with using the range as a measure of variability is that it is completely determined by the two extreme values and ignores the other scores in the distribution. Thus, a distribution with one unusually large (or small) score has a large range even if the other scores are all clustered close together.

Because the range does not consider all of the scores in the distribution, it often does not give an accurate description of the variability for the entire distribution. For this reason, the range is considered to be a crude and unreliable measure of variability. Therefore, in most situations, it does not matter which definition you use to determine the range.

4.3 STANDARD DEVIATION AND VARIANCE FOR A POPULATION

The standard deviation is the most commonly used and the most important measure of variability. Standard deviation uses the mean of the distribution as a reference point and measures variability by considering the distance between each score and the mean.

In simple terms, the standard deviation provides a measure of the standard, or average, distance from the mean, and describes whether the scores are clustered closely around the mean or are widely scattered. The fundamental definition of the standard deviation is the same for both samples and populations, but the calculations differ slightly. We look first at the standard deviation as it is computed for a population, and then turn our attention to samples in Section 4.4.

Although the concept of standard deviation is straightforward, the actual equations appear complex. Therefore, we begin by looking at the logic that leads to these equations. If you remember that our goal is to measure the standard, or typical, distance from the mean, then this logic and the equations that follow should be easier to remember.

STEP 1 The first step in finding the standard distance from the mean is to determine the *deviation,* or distance from the mean, for each individual score. By definition, the deviation for each score is the difference between the score and the mean.

DEFINITION

Deviation is distance from the mean:

$$\text{deviation score} = X - \mu$$

A deviation score is often represented by a lowercase letter x.

For a distribution of scores with $\mu = 50$, if your score is $X = 53$, then your *deviation score* is

$$X - \mu = 53 - 50 = 3$$

If your score is $X = 45$, then your deviation score is

$$X - \mu = 45 - 50 = -5$$

Notice that there are two parts to a deviation score: the sign (+ or –) and the number. The sign tells the direction from the mean—that is, whether the score is located above (+) or below (–) the mean. The number gives the actual distance from the mean. For example, a deviation score of –6 corresponds to a score that is below the mean by a distance of 6 points.

STEP 2 Because our goal is to compute a measure of the standard distance from the mean, the obvious next step is to calculate the mean of the deviation scores. To compute this mean, you first add up the deviation scores and then divide by N. This process is demonstrated in the following example.

EXAMPLE 4.1 We start with the following set of $N = 4$ scores. These scores add up to $\Sigma X = 12$, so the mean is $\mu = \frac{12}{4} = 3$. For each score, we have computed the deviation.

X	$X - \mu$
8	+5
1	–2
3	0
0	–3
	$0 = \Sigma(X - \mu)$

Note that the deviation scores add up to zero. This should not be surprising if you remember that the mean serves as a balance point for the distribution. The total of the

distances above the mean is exactly equal to the total of the distances below the mean (see page 76). Thus, the total for the positive deviations is exactly equal to the total for the negative deviations, and the complete set of deviations always adds up to zero.

Because the sum of the deviations is always zero, the mean of the deviations is also zero and is of no value as a measure of variability. The mean of the deviations is zero if the scores are closely clustered and it is zero if the scores are widely scattered. (You should note, however, that the constant value of zero can be useful in other ways. Whenever you are working with deviation scores, you can check your calculations by making sure that the deviation scores add up to zero.)

STEP 3 The average of the deviation scores does not work as a measure of variability because it is always zero. Clearly, this problem results from the positive and negative values canceling each other out. The solution is to get rid of the signs (+ and –). The standard procedure for accomplishing this is to square each deviation score. Using the squared values, you then compute the *mean squared deviation,* which is called *variance.*

DEFINITION

Population variance equals the mean squared deviation. Variance is the average squared distance from the mean.

Note that the process of squaring deviation scores does more than simply get rid of plus and minus signs. It results in a measure of variability based on *squared* distances. Although variance is valuable for some of the *inferential* statistical methods covered later, the concept of squared distance is not an intuitive or easy to understand *descriptive* measure. For example, it is not particularly useful to know that the squared distance from New York City to Boston is 26,244 miles squared. The squared value becomes meaningful, however, if you take the square root. Therefore, we continue the process with one more step.

STEP 4 Remember that our goal is to compute a measure of the standard distance from the mean. Variance, which measures the average squared distance from the mean, is not exactly what we want. The final step simply takes the square root of the variance to obtain the *standard deviation,* which measures the standard distance from the mean.

DEFINITION

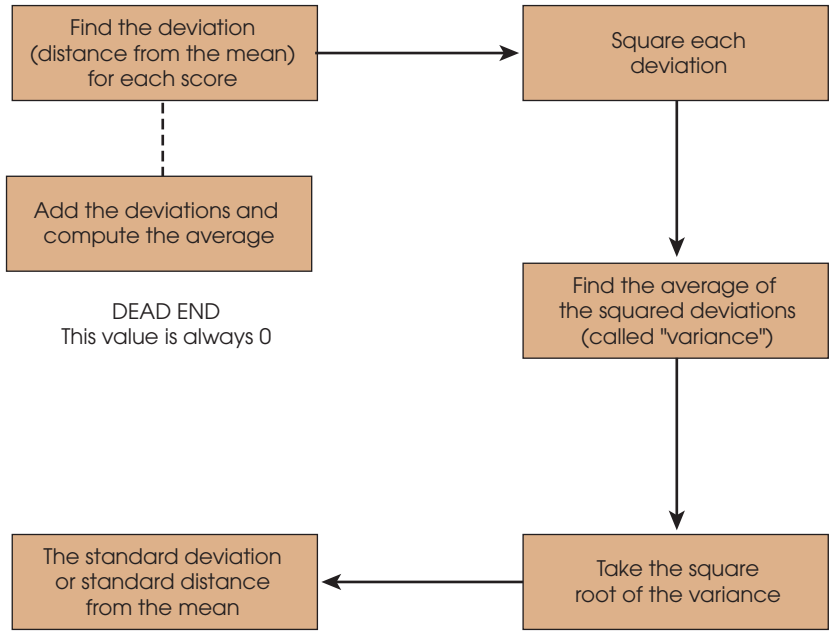
Standard deviation is the square root of the variance and provides a measure of the standard, or average, distance from the mean.

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

Figure 4.2 shows the overall process of computing variance and standard deviation. Remember that our goal is to measure variability by finding the standard distance from the mean. However, we cannot simply calculate the average of the distances because this value will always be zero. Therefore, we begin by squaring each distance, then we find the average of the squared distances, and finally we take the square root to obtain a measure of the standard distance. Technically, the standard deviation is the square root of the average squared deviation. Conceptually, however, the standard deviation provides a measure of the average distance from the mean.

Because the standard deviation and variance are defined in terms of distance from the mean, these measures of variability are used only with numerical scores that are obtained from measurements on an interval or a ratio scale. Recall from Chapter 1 (p. 24) that these two scales are the only ones that provide information about distance; nominal and ordinal scales do not. Also, recall from Chapter 3 (p. 92) that it is inappropriate to compute a mean for ordinal data and impossible to compute a mean for nominal data. Because the mean is a critical component in the calculation of standard deviation

FIGURE 4.2
The calculation of variance and standard deviation.



and variance, the same restrictions that apply to the mean also apply to these two measures of variability. Specifically, the mean, the standard deviation, and the variance should be used only with numerical scores from interval or ordinal scales of measurement.

Although we still have not presented any formulas for variance or standard deviation, you should be able to compute these two statistical values from their definitions. The following example demonstrates this process.

EXAMPLE 4.2

We will calculate the variance and standard deviation for the following population of $N = 5$ scores:

1, 9, 5, 8, 7

Remember that the purpose of standard deviation is to measure the standard distance from the mean, so we begin by computing the population mean. These five scores add up to $\Sigma X = 30$ so the mean is $\mu = \frac{30}{5} = 6$. Next, we find the deviation, (distance from the mean) for each score and then square the deviations. Using the population mean $\mu = 6$, these calculations are shown in the following table.

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$
1	–5	25
9	3	9
5	–1	1
8	2	4
7	1	1
		40 = the sum of the squared deviations

For this set of $N = 5$ scores, the squared deviations add up to 40. The mean of the squared deviations, the variance, is $\frac{40}{5} = 8$, and the standard deviation is $\sqrt{8} = 2.83$.

You should note that a standard deviation of 2.83 is a sensible answer for this distribution. The five scores in the population are shown in a histogram in Figure 4.3 so that you can see the distances more clearly. Note that the scores closest to the mean are only 1 point away. Also, the score farthest from the mean is 5 points away. For this distribution, the largest distance from the mean is 5 points and the smallest distance is 1 point. Thus, the standard distance should be somewhere between 1 and 5. By looking at a distribution in this way, you should be able to make a rough estimate of the standard deviation. In this case, the standard deviation should be between 1 and 5, probably around 3 points. The value we calculated for the standard deviation is in excellent agreement with this estimate.

Making a quick estimate of the standard deviation can help you avoid errors in calculation. For example, if you calculated the standard deviation for the scores in Figure 4.3 and obtained a value of 12, you should realize immediately that you have made an error. (If the biggest deviation is only 5 points, then it is impossible for the standard deviation to be 12.)

LEARNING CHECK

1. Briefly explain what is measured by the standard deviation and what is measured by the variance.
2. The deviation scores are calculated for each individual in a population of $N = 4$. The first three individuals have deviations of +2, +4, and –1. What is the deviation for the fourth individual?
3. What is the standard deviation for the following set of $N = 5$ scores: 10, 10, 10, 10, and 10? (*Note:* You should be able to answer this question directly from the definition of standard deviation, without doing any calculations.)
4. Calculate the variance for the following population of $N = 5$ scores: 4, 0, 7, 1, 3.

ANSWERS

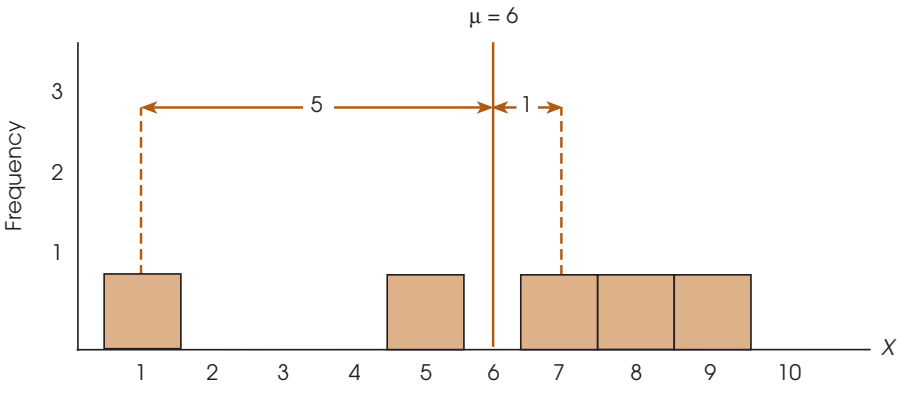
1. Standard deviation measures the standard distance from the mean and variance measures the average squared distance from the mean.
2. The deviation scores for the entire set must add up to zero. The first four deviations add to +5 so the fifth deviation must be –5.
3. Because there is no variability (the scores are all the same), the standard deviation is zero.
4. For these scores, the sum of the squared deviations is 30 and the variance is 30/5 = 6.

FORMULAS FOR POPULATION VARIANCE AND STANDARD DEVIATION

The concepts of standard deviation and variance are the same for both samples and populations. However, the details of the calculations differ slightly, depending on whether you have data from a sample or from a complete population. We first consider the formulas for populations and then look at samples in Section 4.4.

FIGURE 4.3

A frequency distribution histogram for a population of $N = 5$ scores. The mean for this population is $\mu = 6$. The smallest distance from the mean is 1 point, and the largest distance is 5 points. The standard distance (or standard deviation) should be between 1 and 5 points.



The sum of squared deviations (*SS*) Recall that variance is defined as the mean of the squared deviations. This mean is computed in exactly the same way you compute any mean: First find the sum, and then divide by the number of scores.

$$\text{variance} = \text{mean squared deviation} = \frac{\text{sum of squared deviations}}{\text{number of scores}}$$

The value in the numerator of this equation, the sum of the squared deviations, is a basic component of variability, and we focus on it. To simplify things, it is identified by the notation *SS* (for sum of squared deviations), and it generally is referred to as the *sum of squares.*

DEFINITION

SS, or **sum of squares,** is the sum of the squared deviation scores.

You need to know two formulas to compute *SS*. These formulas are algebraically equivalent (they always produce the same answer), but they look different and are used in different situations.

The first of these formulas is called the definitional formula because the symbols in the formula literally define the process of adding up the squared deviations:

$$\text{Definitional formula: } SS = \Sigma(X - \mu)^2 \tag{4.1}$$

To find the sum of the squared deviations, the formula instructs you to perform the following sequence of calculations:

1. Find each deviation score $(X - \mu)$.
2. Square each deviation score $(X - \mu)^2$.
3. Add the squared deviations.

The result is *SS*, the sum of the squared deviations. The following example demonstrates using this formula.

EXAMPLE 4.3

We compute *SS* for the following set of $N = 4$ scores. These scores have a sum of $\Sigma X = 8$, so the mean is $\mu = \frac{8}{4} = 2$. The following table shows the deviation and the squared deviation for each score. The sum of the squared deviation is $SS = 22$.

Score X	Deviation $X - \mu$	Squared Deviation $(X - \mu)^2$	
1	–1	1	$\Sigma X = 8$
0	–2	4	$\mu = 2$
6	+4	16	
1	–1	1	
		$22 = \Sigma(X - \mu)^2$	

Although the definitional formula is the most direct method for computing *SS*, it can be awkward to use. In particular, when the mean is not a whole number, the deviations all contain decimals or fractions, and the calculations become difficult. In addition, calculations with decimal values introduce the opportunity for rounding error, which can make the result less accurate. For these reasons, an alternative formula has been developed for computing *SS*. The alternative, known as the computational formula, performs calculations with the scores (not the deviations) and therefore minimizes the complications of decimals and fractions.

$$\text{computational formula: } SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \tag{4.2}$$

The first part of this formula directs you to square each score and then add the squared values, ΣX^2. In the second part of the formula, you find the sum of the scores, ΣX, then square this total and divide the result by N. Finally, subtract the second part from the first. The use of this formula is shown in Example 4.4 with the same scores that we used to demonstrate the definitional formula.

EXAMPLE 4.4

The computational formula can be used to calculate *SS* for the same set of $N = 4$ scores we used in Example 4.3. Note that the formula requires the calculation of two sums: first, compute ΣX, and then square each score and compute ΣX^2. These calculations are shown in the following table. The two sums are used in the formula to compute *SS*.

X	X^2
1	1
0	0
6	36
1	1
$\Sigma X = 8$	$\Sigma X^2 = 38$

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$
$$= 38 - \frac{(8)^2}{4}$$
$$= 38 - \frac{64}{4}$$
$$= 38 - 16$$
$$= 22$$

Note that the two formulas produce exactly the same value for *SS*. Although the formulas look different, they are in fact equivalent. The definitional formula provides the most direct representation of the concept of *SS*; however, this formula can be awkward to use, especially if the mean includes a fraction or decimal value. If you have a small group of scores and the mean is a whole number, then the definitional formula is fine; otherwise the computational formula is usually easier to use.

FINAL FORMULAS AND NOTATION

In the same way that sum of squares, or *SS*, is used to refer to the sum of squared deviations, the term *mean square,* or *MS*, is often used to refer to variance, which is the mean squared deviation.

With the definition and calculation of *SS* behind you, the equations for variance and standard deviation become relatively simple. Remember that variance is defined as the mean squared deviation. The mean is the sum of the squared deviations divided by *N*, so the equation for the *population variance* is

$$\text{variance} = \frac{SS}{N}$$

Standard deviation is the square root of variance, so the equation for the *population standard deviation* is

$$\text{standard deviation} = \sqrt{\frac{SS}{N}}$$

There is one final bit of notation before we work completely through an example computing *SS*, variance, and standard deviation. Like the mean (μ), variance and standard deviation are parameters of a population and are identified by Greek letters. To identify the standard deviation, we use the Greek letter sigma (the Greek letter *s,* standing for standard deviation). The capital letter sigma (Σ) has been used already, so we now use the lowercase sigma, σ, as the symbol for the population standard deviation. To emphasize the relationship between standard deviation and variance, we use σ^2 as the symbol for population variance (standard deviation is the square root of the variance). Thus,

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}} \tag{4.3}$$

$$\text{population variance} = \sigma^2 = \frac{SS}{N} \tag{4.4}$$

Earlier, in Examples 4.3 and 4.4, we computed the sum of squared deviations for a simple population of $N = 4$ scores (1, 0, 6, 1) and obtained $SS = 22$. For this population, the variance is

$$\sigma^2 = \frac{SS}{N} = \frac{22}{4} = 5.50$$

and the standard deviation is $\sigma = \sqrt{5.50} = 2.345$

LEARNING CHECK

1. Find the sum of the squared deviations, *SS*, for each of the following populations. Note that the definitional formula works well for one population but the computational formula is better for the other.

Population 1: 3, 1, 5, 1
Population 2: 6, 4, 2, 0, 9, 3

2. a. Sketch a histogram showing the frequency distribution for the following population of $N = 6$ scores: 12, 0, 1, 7, 4, 6. Locate the mean in your sketch, and estimate the value of the standard deviation.

b. Calculate *SS*, variance, and the standard deviation for these scores. How well does your estimate compare with the actual standard deviation?

ANSWERS

1. For population 1, the mean is not a whole number ($M = 2.5$) and the computational formula is better and produces $SS = 11$. The mean is a whole number ($M = 4$) and definitional formula works well for population 2, which has $SS = 50$.

2. **a.** Your sketch should show a mean of $\mu = 5$. The scores closest to the mean are $X = 4$ and $X = 6$, both of which are only 1 point away. The score farthest from the mean is $X = 12$, which is 7 points away. The standard deviation should have a value between 1 and 7, probably around 4 points.

   b. For these scores, $SS = 96$, the variance is $96/6 = 16$, and the standard deviation is $\sigma = 4$.

4.4 STANDARD DEVIATION AND VARIANCE FOR SAMPLES

A sample statistic is said to be *biased* if, on average, it consistently overestimates or underestimates the corresponding population parameter.

The goal of inferential statistics is to use the limited information from samples to draw general conclusions about populations. The basic assumption of this process is that samples should be representative of the populations from which they come. This assumption poses a special problem for variability because samples consistently tend to be less variable than their populations. An example of this general tendency is shown in Figure 4.4. Notice that a few extreme scores in the population tend to make the population variability relatively large. However, these extreme values are unlikely to be obtained when you are selecting a sample, which means that the sample variability is relatively small. The fact that a sample tends to be less variable than its population means that sample variability gives a *biased* estimate of population variability. This bias is in the direction of underestimating the population value rather than being right on the mark. (The concept of a biased statistic is discussed in more detail in Section 4.5.)

Fortunately, the bias in sample variability is consistent and predictable, which means it can be corrected. For example, if the speedometer in your car consistently shows speeds that are 5 mph slower than you are actually going, it does not mean that

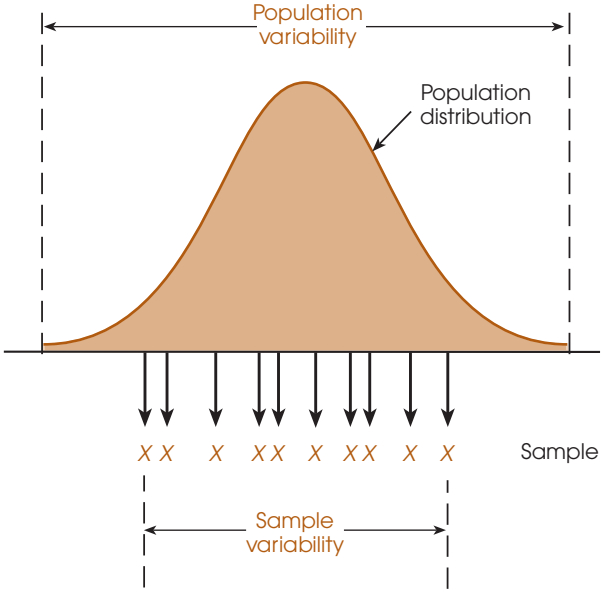


FIGURE 4.4

The population of adult heights forms a normal distribution. If you select a sample from this population, you are most likely to obtain individuals who are near average in height. As a result, the scores in the sample are less variable (spread out) than the scores in the population.

the speedometer is useless. It simply means that you must make an adjustment to the speedometer reading to get an accurate speed. In the same way, we make an adjustment in the calculation of sample variance. The purpose of the adjustment is to make the resulting value for sample variance an accurate and unbiased representative of the population variance.

The calculations of variance and standard deviation for a sample follow the same steps that were used to find population variance and standard deviation. Except for minor changes in notation, the first three steps in this process are exactly the same for a sample as they were for a population. That is, calculating the sum of the squared deviations, *SS*, is the same for a sample as it is for a population. The changes in notation involve using M for the sample mean instead of μ, and using n (instead of N) for the number of scores. Thus, to find the *SS* for a sample

1. Find the deviation from the mean for each score: deviation $= X - M$
2. Square each deviation: squared deviation $= (X - M)^2$
3. Add the squared deviations: $SS = \Sigma(X - M)^2$

These three steps can be summarized in a definitional formula for *SS*:

$$\text{Definitional formula: } SS = \Sigma(X - M)^2 \tag{4.5}$$

The value of *SS* also can be obtained using a computational formula. Except for one minor difference in notation (using n in place of N), the computational formula for *SS* is the same for a sample as it was for a population (see Equation 4.2). Using sample notation, this formula is:

$$\text{Computational formula: } SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} \tag{4.6}$$

Again, calculating *SS* for a sample is exactly the same as for a population, except for minor changes in notation. After you compute *SS*, however, it becomes critical to differentiate between samples and populations. To correct for the bias in sample variability, it is necessary to make an adjustment in the formulas for sample variance and standard deviation. With this in mind, *sample variance* (identified by the symbol s^2) is defined as

$$\text{sample variance} = s^2 = \frac{SS}{n - 1} \tag{4.7}$$

Sample standard deviation (identified by the symbol s) is simply the square root of the variance.

$$\text{sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{SS}{n - 1}} \tag{4.8}$$

Notice that the sample formulas divide by $n - 1$, unlike the population formulas, which divide by N (see Equations 4.3 and 4.4). This is the adjustment that is necessary to correct for the bias in sample variability. The effect of the adjustment is to increase the value that you obtain. Dividing by a smaller number ($n - 1$ instead of n) produces a larger result and makes sample variance an accurate and unbiased estimator of population variance. The following example demonstrates the calculation of variance and standard deviation for a sample.

Remember, sample variability tends to underestimate population variability unless some correction is made.

EXAMPLE 4.5

We have selected a sample of $n = 7$ scores from a population. The scores are 1, 6, 4, 3, 8, 7, 6. The frequency distribution histogram for this sample is shown in Figure 4.5. Before we begin any calculations, you should be able to look at the sample distribution and make a preliminary estimate of the outcome. Remember that standard deviation measures the standard distance from the mean. For this sample, the mean is $M = \frac{35}{7} = 5$. The scores closest to the mean are $X = 4$ and $X = 6$, both of which are exactly 1 point away. The score farthest from the mean is $X = 1$, which is 4 points away. With the smallest distance from the mean equal to 1 and the largest distance equal to 4, we should obtain a standard distance somewhere between 1 and 4, probably around 2.5.

We begin the calculations by finding the value of SS for this sample. Because there are only a few scores and the mean is a whole number ($M = 5$), the definitional formula is easy to use. The scores, the deviations, and the squared deviations are shown in the following table.

Squared Score X	Deviation $X - M$	Deviation $(X - M)^2$
1	–4	16
6	1	1
4	–1	1
3	–2	4
8	3	9
7	2	4
6	1	1
		$36 = SS = \Sigma(X - M)^2$

The sum of squared deviations for this sample is $SS = 36$. Continuing the calculations,

$$\text{sample variance} = s^2 = \frac{SS}{n-1} = \frac{36}{7-1} = 6$$

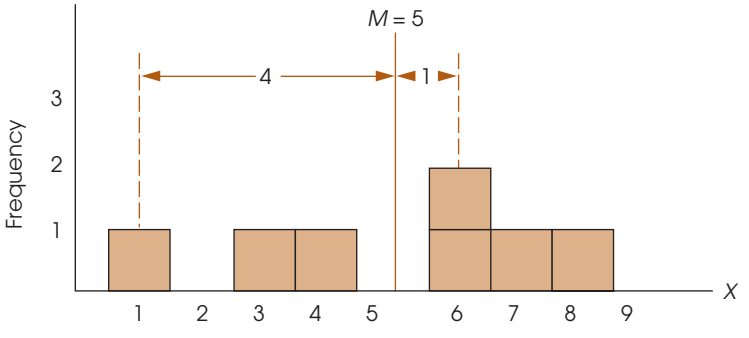
Finally, the standard deviation is

$$s = \sqrt{s^2} = \sqrt{6} = 2.45$$

Note that the value we obtained is in excellent agreement with our preliminary prediction (see Figure 4.5).

FIGURE 4.5

The frequency distribution histogram for a sample of $n = 7$ scores. The sample mean is $M = 5$. The smallest distance from the mean is 1 point, and the largest distance from the mean is 4 points. The standard distance (standard deviation) should be between 1 and 4 points, or about 2.5.



Remember that the formulas for sample variance and standard deviation were constructed so that the sample variability would provide a good estimate of population variability. For this reason, the sample variance is often called *estimated population variance,* and the sample standard deviation is called *estimated population standard deviation.* When you have only a sample to work with, the variance and standard deviation for the sample provide the best possible estimates of the population variability.

SAMPLE VARIABILITY AND DEGREES OF FREEDOM

Although the concept of a deviation score and the calculation of SS are almost exactly the same for samples and populations, the minor differences in notation are really very important. Specifically, with a population, you find the deviation for each score by measuring its distance from the population mean, μ. With a sample, on the other hand, the value of μ is unknown and you must measure distances from the sample mean. Because the value of the sample mean varies from one sample to another, you must first compute the sample mean before you can begin to compute deviations. However, calculating the value of M places a restriction on the variability of the scores in the sample. This restriction is demonstrated in the following example.

EXAMPLE 4.6

Suppose we select a sample of $n = 3$ scores and compute a mean of $M = 5$. The first two scores in the sample have no restrictions; they are independent of each other and they can have any values. For this demonstration, we assume that we obtained $X = 2$ for the first score and $X = 9$ for the second. At this point, however, the third score in the sample is restricted.

X	A sample of $n = 3$ scores with a mean of $M = 5$.
2	
9	
—	← What is the third score?

For this example, the third score must be $X = 4$. The reason that the third score is restricted to $X = 4$ is that the entire sample of $n = 3$ scores has a mean of $M = 5$. For 3 scores to have a mean of 5, the scores must have a total of $\Sigma X = 15$. Because the first two scores add up to 11 (9 + 2), the third score must be $X = 4$.

In Example 4.6, the first two out of three scores were free to have any values, but the final score was dependent on the values chosen for the first two. In general, with a sample of n scores, the first $n - 1$ scores are free to vary, but the final score is restricted. As a result, the sample is said to have $n - 1$ *degrees of freedom*.

DEFINITION

For a sample of n scores, the **degrees of freedom,** or ***df,*** for the sample variance are defined as $df = n - 1$. The degrees of freedom determine the number of scores in the sample that are independent and free to vary.

The $n - 1$ degrees of freedom for a sample is the same $n - 1$ that is used in the formulas for sample variance and standard deviation. Remember that variance is defined as the mean squared deviation. As always, this mean is computed by finding the sum and dividing by the number of scores:

$$\text{mean} = \frac{\text{sum}}{\text{number}}$$

To calculate sample variance (mean squared deviation), we find the sum of the squared deviations (*SS*) and divide by the number of scores that are free to vary. This number is $n - 1 = df$. Thus, the formula for sample variance is

$$s^2 = \frac{\text{sum of squared deviations}}{\text{number of scores free to vary}} = \frac{SS}{df} = \frac{SS}{n - 1}$$

Later in this book, we use the concept of degrees of freedom in other situations. For now, remember that knowing the sample mean places a restriction on sample variability. Only $n - 1$ of the scores are free to vary; $df = n - 1$.

LEARNING CHECK

1. **a.** Sketch a histogram showing the frequency distribution for the following sample of $n = 5$ scores: 3, 1, 9, 4, 3. Locate the mean in your sketch, and estimate the value of the sample standard deviation.
   b. Calculate *SS*, variance, and standard deviation for this sample. How well does your estimate from part a compare with the real standard deviation?
2. For the following set of scores: 1, 5, 7, 3, 4
   a. Assume that this is a population of $N = 5$ scores and compute *SS* and variance for the population.
   b. Assume that this is a sample of $n = 5$ scores and compute *SS* and variance for the sample.
3. Explain why the formula for sample variance divides *SS* by $n - 1$ instead of dividing by n.

ANSWERS

1. **a.** Your graph should show a sample mean of $M = 4$. The score farthest from the mean is $X = 9$ (which is 5 points away), and the closest score is $X = 4$ (which is 0 points away). You should estimate the standard deviation to be between 1 and 5 points, probably around 3 points.
   b. For this sample, $SS = 36$; the sample variance is $36/4 = 9$; the sample standard deviation is $\sqrt{9} = 3$.
2. **a.** $SS = 20$ and the population variance is $20/5 = 4$.
   b. $SS = 20$ and the sample variance is $20/4 = 5$.
3. Without some correction, sample variability consistently underestimates the population variability. Dividing by a smaller number ($n - 1$ instead of n) increases the value of the sample variance and makes it an unbiased estimate of the population variance.

4.5 MORE ABOUT VARIANCE AND STANDARD DEVIATION

PRESENTING THE MEAN AND STANDARD DEVIATION IN A FREQUENCY DISTRIBUTION GRAPH

In frequency distribution graphs, we identify the position of the mean by drawing a vertical line and labeling it with μ or M. Because the standard deviation measures distance from the mean, it is represented by a line or an arrow drawn from the mean outward for a distance equal to the standard deviation and labeled with a σ or an s. Figure 4.6(a) shows an example of a population distribution with a mean of $\mu = 80$ and a standard

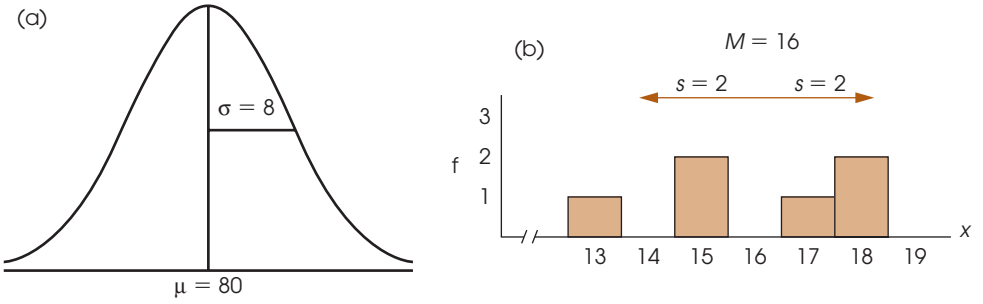


FIGURE 4.6

Showing means and standard deviations in frequency distribution graphs. (a) A population distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 8$. (b) A sample with a mean of $M = 16$ and a standard deviation of $s = 2$.

deviation of $\sigma = 8$, and Figure 4.6(b) shows the frequency distribution for a sample with a mean of $M = 16$ and a standard deviation of $s = 2$. For rough sketches, you can identify the mean with a vertical line in the middle of the distribution. The standard deviation line should extend approximately halfway from the mean to the most extreme score. [*Note:* In Figure 4.6(a) we show the standard deviation as a line to the right of the mean. You should realize that we could have drawn the line pointing to the left, or we could have drawn two lines (or arrows), with one pointing to the right and one pointing to the left, as in Figure 4.6(b). In each case, the goal is to show the standard distance from the mean.]

SAMPLE VARIANCE AS AN UNBIASED STATISTIC

Earlier we noted that sample variability tends to underestimate the variability in the corresponding population. To correct for this problem we adjusted the formula for sample variance by dividing by $n - 1$ instead of dividing by n. The result of the adjustment is that sample variance provides a much more accurate representation of the population variance. Specifically, dividing by $n - 1$ produces a sample variance that provides an *unbiased* estimate of the corresponding population variance. This does not mean that each individual sample variance is exactly equal to its population variance. In fact, some sample variances overestimate the population value and some underestimate it. However, the average of all the sample variances produces an accurate estimate of the population variance. This is the idea behind the concept of an unbiased statistic.

DEFINITIONS

A sample statistic is **unbiased** if the average value of the statistic is equal to the population parameter. (The average value of the statistic is obtained from all the possible samples for a specific sample size, n.)

A sample statistic is **biased** if the average value of the statistic either underestimates or overestimates the corresponding population parameter.

The following example demonstrates the concept of biased and unbiased statistics.

EXAMPLE 4.7

We have structured this example to mimic "sampling with replacement," which is covered in Chapter 6.

We begin with a population that consists of exactly $N = 6$ scores: 0, 0, 3, 3, 9, 9. With a few calculations you should be able to verify that this population has a mean of $\mu = 4$ and a variance of $\sigma^2 = 14$.

Next, we select samples of $n = 2$ scores from this population. In fact, we obtain every single possible sample with $n = 2$. The complete set of samples is listed in Table 4.1. Notice that the samples are listed systematically to ensure that every possible sample is included. We begin by listing all the samples that have $X = 0$ as the first score, then all the samples with $X = 3$ as the first score, and so on. Notice that the table shows a total of 9 samples.

Finally, we have computed the mean and the variance for each sample. Note that the sample variance has been computed two different ways. First, we examine what happens if there is no correction for bias and the sample variance is computed by simply dividing *SS* by n. Second, we examine the correct sample variances for which *SS* is divided by $n - 1$ to produce an unbiased measure of variance. You should verify our calculations by computing one or two of the values for yourself. The complete set of sample means and sample variances is presented in Table 4.1.

First, consider the column of biased sample variances, which were calculated by dividing by n. These 9 sample variances add up to a total of 63, which produces an average value of $\frac{63}{9} = 7$. The original population variance, however, is $\sigma^2 = 14$. Note that the average of the sample variances is *not* equal to the population variance. If the sample variance is computed by dividing by n, the resulting values do not produce an accurate estimate of the population variance. On average, these sample variances underestimate the population variance and, therefore, are biased statistics.

Next, consider the column of sample variances that are computed using $n - 1$. Although the population has a variance of $\sigma^2 = 14$, you should notice that none of the samples has a variance exactly equal to 14. However, if you consider the complete set of sample variances, you will find that the 9 values add up to a total of 126, which produces an average value of $\frac{126}{9} = 14$. Thus, the average of the sample variances is exactly equal to the original population variance. On average, the sample variance (computed using $n - 1$) produces an accurate, unbiased estimate of the population variance.

Finally, direct your attention to the column of sample means. For this example, the original population has a mean of $\mu = 4$. Although none of the samples has a

TABLE 4.1

The set of all the possible samples for $n = 2$ selected from the population described in Example 4.7. The mean is computed for each sample, and the variance is computed two different ways: (1) dividing by n, which is incorrect and produces a biased statistic; and (2) dividing by $n - 1$, which is correct and produces an unbiased statistic.

Sample	First Score	Second Score	Sample Statistics: Mean *M*	Biased Variance (Using *n*)	Unbiased Variance (Using *n* – 1)
1	0	0	0.00	0.00	0.00
2	0	3	1.50	2.25	4.50
3	0	9	4.50	20.25	40.50
4	3	0	1.50	2.25	4.50
5	3	3	3.00	0.00	0.00
6	3	9	6.00	9.00	18.00
7	9	0	4.50	20.25	40.50
8	9	3	6.00	9.00	18.00
9	9	9	9.00	0.00	0.00
		Totals	36.00	63.00	126.00

mean exactly equal to 4, if you consider the complete set of sample means, you will find that the 9 sample means add up to a total of 36, so the average of the sample means is $\frac{36}{9} = 4$. Note that the average of the sample means is exactly equal to the population mean. Again, this is what is meant by the concept of an unbiased statistic. On average, the sample values provide an accurate representation of the population. In this example, the average of the 9 sample means is exactly equal to the population mean.

In summary, both the sample mean and the sample variance (using $n - 1$) are examples of unbiased statistics. This fact makes the sample mean and sample variance extremely valuable for use as inferential statistics. Although no individual sample is likely to have a mean and variance exactly equal to the population values, both the sample mean and the sample variance, on average, do provide accurate estimates of the corresponding population values.

STANDARD DEVIATION AND DESCRIPTIVE STATISTICS

Because standard deviation requires extensive calculations, there is a tendency to get lost in the arithmetic and forget what standard deviation is and why it is important. Standard deviation is primarily a descriptive measure; it describes how variable, or how spread out, the scores are in a distribution. Behavioral scientists must deal with the variability that comes from studying people and animals. People are not all the same; they have different attitudes, opinions, talents, IQs, and personalities. Although we can calculate the average value for any of these variables, it is equally important to describe the variability. Standard deviation describes variability by measuring *distance from the mean*. In any distribution, some individuals are close to the mean, and others are relatively far from the mean. Standard deviation provides a measure of the typical, or standard, distance from the mean.

Describing an entire distribution Rather than listing all of the individual scores in a distribution, research reports typically summarize the data by reporting only the mean and the standard deviation. When you are given these two descriptive statistics, however, you should be able to visualize the entire set of data. For example, consider a sample with a mean of $M = 36$ and a standard deviation of $s = 4$. Although there are several different ways to picture the data, one simple technique is to imagine (or sketch) a histogram in which each score is represented by a box in the graph. For this sample, the data can be pictured as a pile of boxes (scores) with the center of the pile located at a value of $M = 36$. The individual scores, or boxes, are scattered on both sides of the mean with some of the boxes relatively close to the mean and some farther away. As a rule of thumb, roughly 70% of the scores in a distribution are located within a distance of one standard deviation from the mean, and almost all of the scores (roughly 95%) are within two standard deviations of the mean. In this example, the standard distance from the mean is $s = 4$ points, so your image should have most of the boxes within 4 points of the mean, and nearly all of the boxes within 8 points. One possibility for the resulting image is shown in Figure 4.7.

Describing the location of individual scores Notice that Figure 4.7 not only shows the mean and the standard deviation, but also uses these two values to reconstruct the underlying scale of measurement (the X values along the horizontal line). The scale of measurement helps to complete the picture of the entire distribution and relate each individual score to the rest of the group. In this example, you should realize that a score of $X = 34$ is located near the center of the distribution, only slightly below the mean.

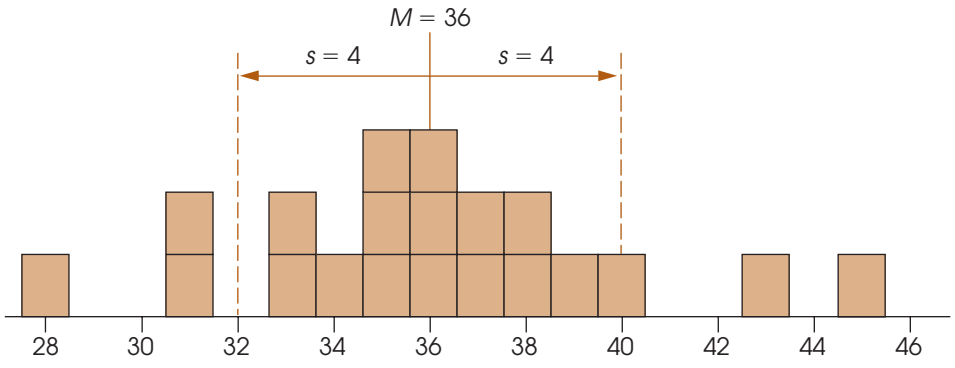


FIGURE 4.7

A sample of $n = 20$ scores with a mean of $M = 36$ and a standard deviation of $s = 4$.

On the other hand, a score of $X = 45$ is an extremely high score, located far out in the right-hand tail of the distribution.

Notice that the relative position of a score depends in part on the size of the standard deviation. In Figure 4.6 (p. 119), for example, we show a population distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 8$, and a sample distribution with a mean of $M = 16$ and a standard deviation of $s = 2$. In the population distribution, a score that is 4 points above the mean is slightly above average but is certainly not an extreme value. In the sample distribution, however, a score that is 4 points above the mean is an extremely high score. In each case, the relative position of the score depends on the size of the standard deviation. For the population, a deviation of 4 points from the mean is relatively small, corresponding to only half of the standard deviation. For the sample, on the other hand, a 4-point deviation is very large, twice the size of the standard deviation.

The general point of this discussion is that the mean and standard deviation are not simply abstract concepts or mathematical equations. Instead, these two values should be concrete and meaningful, especially in the context of a set of scores. The mean and standard deviation are central concepts for most of the statistics that are presented in the following chapters. A good understanding of these two statistics will help you with the more complex procedures that follow. (Box 4.1.)

TRANSFORMATIONS OF SCALE

Occasionally a set of scores is transformed by adding a constant to each score or by multiplying each score by a constant value. This happens, for example, when exposure to a treatment adds a fixed amount to each participant's score or when you want to change the unit of measurement (to convert from minutes to seconds, multiply each score by 60). What happens to the standard deviation when the scores are transformed in this manner?

The easiest way to determine the effect of a transformation is to remember that the standard deviation is a measure of distance. If you select any two scores and see what happens to the distance between them, you also find out what happens to the standard deviation.

1. Adding a constant to each score does not change the standard deviation If you begin with a distribution that has $\mu = 40$ and $\sigma = 10$, what happens to σ if you add 5 points to every score? Consider any two scores in this distribution: Suppose, for

BOX 4.1 AN ANALOGY FOR THE MEAN AND THE STANDARD DEVIATION

Although the basic concepts of the mean and the standard deviation are not overly complex, the following analogy often helps students gain a more complete understanding of these two statistical measures.

In our local community, the site for a new high school was selected because it provides a central location. An alternative site on the western edge of the community was considered, but this site was rejected because it would require extensive busing for students living on the east side. In this example, the location of the high school is analogous to the concept of the mean; just as the high school is located in the center of the community, the mean is located in the center of the distribution of scores.

For each student in the community, it is possible to measure the distance between home and the new high school. Some students live only a few blocks from the new school and others live as much as 3 miles away. The average distance that a student must travel to school was calculated to be 0.80 miles. The average distance from the school is analogous to the concept of the standard deviation; that is, the standard deviation measures the standard distance from an individual score to the mean.

example, that these are exam scores and that you had a score of $X = 41$ and your friend had $X = 43$. The distance between these two scores is $43 - 41 = 2$ points. After adding the constant, 5 points, to each score, your score would be $X = 46$, and your friend would have $X = 48$. The distance between scores is still 2 points. Adding a constant to every score does not affect any of the distances and, therefore, does not change the standard deviation. This fact can be seen clearly if you imagine a frequency distribution graph. If, for example, you add 10 points to each score, then every score in the graph is moved 10 points to the right. The result is that the entire distribution is shifted to a new position 10 points up the scale. Note that the mean moves along with the scores and is increased by 10 points. However, the variability does not change because each of the deviation scores $(X - \mu)$ does not change.

2. Multiplying each score by a constant causes the standard deviation to be multiplied by the same constant Consider the same distribution of exam scores we looked at earlier. If $\mu = 40$ and $\sigma = 10$, what would happen to σ if each score were multiplied by 2? Again, we look at two scores, $X = 41$ and $X = 43$, with a distance between them equal to 2 points. After all of the scores have been multiplied by 2, these scores become $X = 82$ and $X = 86$. Now the distance between scores is 4 points, twice the original distance. Multiplying each score causes each distance to be multiplied, so the standard deviation also is multiplied by the same amount.

IN THE LITERATURE

REPORTING THE STANDARD DEVIATION

In reporting the results of a study, the researcher often provides descriptive information for both central tendency and variability. The dependent variables in psychology research are often numerical values obtained from measurements on interval or ratio scales. With numerical scores, the most common descriptive statistics are the mean (central tendency) and the standard deviation (variability), which are usually reported together. In many journals, especially those following APA style,

the symbol *SD* is used for the sample standard deviation. For example, the results might state:

> Children who viewed the violent cartoon displayed more aggressive responses ($M = 12.45$, $SD = 3.7$) than those who viewed the control cartoon ($M = 4.22$, $SD = 1.04$).

When reporting the descriptive measures for several groups, the findings may be summarized in a table. Table 4.2 illustrates the results of hypothetical data.

TABLE 4.2

The number of aggressive responses in male and female children after viewing cartoons.

	Type of Cartoon	
	Violent	Control
Males	$M = 15.72$	$M = 6.94$
	$SD = 4.43$	$SD = 2.26$
Females	$M = 3.47$	$M = 2.61$
	$SD = 1.12$	$SD = 0.98$

Sometimes the table also indicates the sample size, *n*, for each group. You should remember that the purpose of the table is to present the data in an organized, concise, and accurate manner. ❑

VARIANCE AND INFERENTIAL STATISTICS

In very general terms, the goal of inferential statistics is to detect meaningful and significant patterns in research results. The basic question is whether the patterns observed in the sample data reflect corresponding patterns that exist in the population, or are simply random fluctuations that occur by chance. Variability plays an important role in the inferential process because the variability in the data influences how easy it is to see patterns. In general, low variability means that existing patterns can be seen clearly, whereas high variability tends to obscure any patterns that might exist. The following example provides a simple demonstration of how variance can influence the perception of patterns.

EXAMPLE 4.8

In most research studies the goal is to compare means for two (or more) sets of data. For example:

> Is the mean level of depression lower after therapy than it was before therapy?
>
> Is the mean attitude score for men different from the mean score for women?
>
> Is the mean reading achievement score higher for students in a special program than for students in regular classrooms?

In each of these situations, the goal is to find a clear difference between two means that would demonstrate a significant, meaningful pattern in the results. Variability plays an important role in determining whether a clear pattern exists. Consider the following data representing hypothetical results from two experiments, each comparing two treatment conditions. For both experiments, your task is to determine whether there appears to be any consistent difference between the scores in treatment 1 and the scores in treatment 2.

Experiment A		Experiment B	
Treatment 1	Treatment 2	Treatment 1	Treatment 2
35	39	31	46
34	40	15	21
36	41	57	61
35	40	37	32

For each experiment, the data have been constructed so that there is a 5-point mean difference between the two treatments: On average, the scores in treatment 2 are 5 points higher than the scores in treatment 1. The 5-point difference is relatively easy to see in experiment A, where the variability is low, but the same 5-point difference is difficult to see in experiment B, where the variability is large. Again, high variability tends to obscure any patterns in the data. This general fact is perhaps even more convincing when the data are presented in a graph. Figure 4.8 shows the two sets of data from experiments A and B. Notice that the results from experiment A clearly show the 5-point difference between treatments. One group of scores piles up around 35 and the second group piles up around 40. On the other hand, the scores from experiment B [Figure 4.8(b)] seem to be mixed together randomly with no clear difference between the two treatments.

In the context of inferential statistics, the variance that exists in a set of sample data is often classified as *error variance.* This term is used to indicate that the sample variance represents unexplained and uncontrolled differences between scores. As the error variance increases, it becomes more difficult to see any systematic differences or patterns that might exist in the data. An analogy is to think of variance as the static that appears on a radio station or a cell phone when you enter an area of poor reception. In general, variance makes it difficult to get a clear signal from the data. High variance can make it difficult or impossible to see a mean difference between two sets of scores, or to see any other meaningful patterns in the results from a research study.

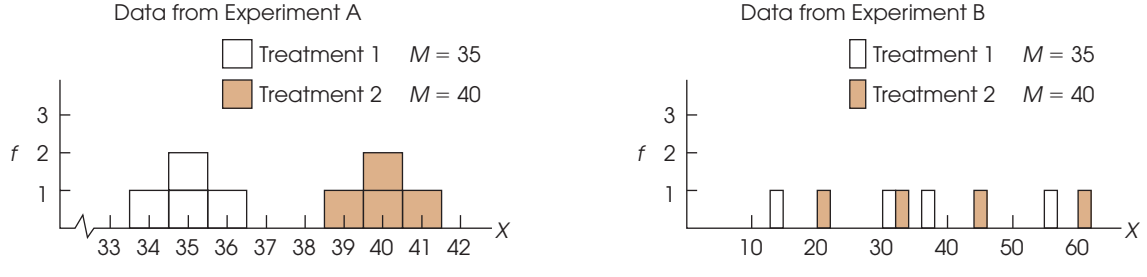


FIGURE 4.8

Graphs showing the results from two experiments. In experiment A, the variability is small and it is easy to see the 5-point mean difference between the two treatments. In experiment B, however, the 5-point mean difference between treatments is obscured by the large variability.

LEARNING CHECK

1. Explain the difference between a biased and an unbiased statistic.
2. In a population with a mean of $\mu = 50$ and a standard deviation of $\sigma = 10$, would a score of $X = 58$ be considered an extreme value (far out in the tail of the distribution)? What if the standard deviation were $\sigma = 3$?
3. A population has a mean of $\mu = 70$ and a standard deviation of $\sigma = 5$.
   a. If 10 points were added to every score in the population, what would be the new values for the population mean and standard deviation?
   b. If every score in the population were multiplied by 2, what would be the new values for the population mean and standard deviation?

ANSWERS

1. If a statistic is biased, it means that the average value of the statistic does not accurately represent the corresponding population parameter. Instead, the average value of the statistic either overestimates or underestimates the parameter. If a statistic is unbiased, it means that the average value of the statistic is an accurate representation of the corresponding population parameter.
2. With $\sigma = 10$, a score of $X = 58$ would be located in the central section of the distribution (within one standard deviation). With $\sigma = 3$, a score of $X = 58$ would be an extreme value, located more than two standard deviations above the mean.
3. a. The new mean would be $\mu = 80$ but the standard deviation would still be $\sigma = 5$.
   b. The new mean would be $\mu = 140$ and the new standard deviation would be $\sigma = 10$.

SUMMARY

1. The purpose of variability is to measure and describe the degree to which the scores in a distribution are spread out or clustered together. There are three basic measures of variability: the range, the variance, and the standard deviation.

   The range is the distance covered by the set of scores, from the smallest score to the largest score. The range is completely determined by the two extreme scores and is considered to be a relatively crude measure of variability.

   Standard deviation and variance are the most commonly used measures of variability. Both of these measures are based on the idea that each score can be described in terms of its deviation, or distance, from the mean. The variance is the mean of the squared deviations. The standard deviation is the square root of the variance and provides a measure of the standard distance from the mean.

2. To calculate variance or standard deviation, you first need to find the sum of the squared deviations, *SS*. Except for minor changes in notation, the calculation of *SS* is identical for samples and populations. There are two methods for calculating *SS*:

   I. By definition, you can find *SS* using the following steps:
      a. Find the deviation $(X - \mu)$ for each score.
      b. Square each deviation.
      c. Add the squared deviations.

      This process can be summarized in a formula as follows:

      $$\text{Definitional formula:} \quad SS = \Sigma(X - \mu)^2$$

   II. The sum of the squared deviations can also be found using a computational formula, which is especially useful when the mean is not a whole number:

      $$\text{Computational formula: } SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

3. Variance is the mean squared deviation and is obtained by finding the sum of the squared deviations and then dividing by the number of scores. For a population, variance is

$$\sigma^2 = \frac{SS}{N}$$

For a sample, only $n - 1$ of the scores are free to vary (degrees of freedom or $df = n - 1$), so sample variance is

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df}$$

Using $n - 1$ in the sample formula makes the sample variance an accurate and unbiased estimate of the population variance.

4. Standard deviation is the square root of the variance. For a population, this is

$$\sigma = \sqrt{\frac{SS}{N}}$$

Sample standard deviation is

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

5. Adding a constant value to every score in a distribution does not change the standard deviation. Multiplying every score by a constant, however, causes the standard deviation to be multiplied by the same constant.

KEY TERMS

variability (104)
range (106)
deviation score (107)
sum of squares (*SS*) (111)
population variance (σ^2) (113)
population standard deviation (σ) (113)
sample variance (s^2) (115)
sample standard deviation (s) (115)
degrees of freedom (df) (117)
unbiased statistic (119)
biased statistic (119)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 4 on the book companion website. The website also includes a workshop entitled *Central Tendency and Variability,* which examines the basic concepts of variability and the standard deviation, and reviews the concept of central tendency, which was covered in Chapter 3.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to compute the **Range, Standard Deviation,** and **Variance** for a sample of scores.

Data Entry

1. Enter all of the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Descriptives.**
2. Highlight the column label for the set of scores (VAR00001) in the left box and click the arrow to move it into the **Variable** box.
3. If you want the variance and/or the range reported along with the standard deviation, click on the **Options** box, select **Variance** and/or **Range,** then click **Continue.**
4. Click **OK.**

SPSS Output

The SPSS output is shown in Figure 4.9. The summary table lists the number of scores, the maximum and minimum scores, the mean, the range, the standard deviation, and the variance. Note that the range and variance are included because these values were selected using the Options box during data analysis. Caution: SPSS computes the *sample* standard deviation and *sample* variance using $n - 1$. If your scores are intended to be a population, you can multiply the sample standard deviation by the square root of $(n - 1)/n$ to obtain the population standard deviation.

Note: You can also obtain the mean and standard deviation for a sample if you use SPSS to display the scores in a frequency distribution histogram (see the SPSS section at the end of Chapter 2). The mean and standard deviation are displayed beside the graph.

FOCUS ON PROBLEM SOLVING

1. The purpose of variability is to provide a measure of how spread out the scores in a distribution are. Usually this is described by the standard deviation. Because the calculations are relatively complicated, it is wise to make a preliminary estimate of

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
VAR00001	7	7.00	1.00	8.00	5.0000	2.44949	6.000
Valid N (listwise)	7						

FIGURE 4.9

The SPSS summary table showing descriptive statistics for a sample of $n = 7$ scores.

the standard deviation before you begin. Remember that standard deviation provides a measure of the typical, or standard, distance from the mean. Therefore, the standard deviation must have a value somewhere between the largest and the smallest deviation scores. As a rule of thumb, the standard deviation should be about one-fourth of the range.

2. Rather than trying to memorize all of the formulas for *SS*, variance, and standard deviation, you should focus on the definitions of these values and the logic that relates them to each other:

   SS is the sum of squared deviations.

   Variance is the mean squared deviation.

   Standard deviation is the square root of variance.

   The only formula you should need to memorize is the computational formula for *SS*.

3. A common error is to use $n - 1$ in the computational formula for *SS* when you have scores from a sample. Remember that the *SS* formula always uses n (or N). After you compute *SS* for a sample, you must correct for the sample bias by using $n - 1$ in the formulas for variance and standard deviation.

DEMONSTRATION 4.1

COMPUTING MEASURES OF VARIABILITY

For the following sample data, compute the variance and standard deviation. The scores are:

10, 7, 6, 10, 6, 15

STEP 1 **Compute *SS*, the sum of squared deviations** We use the computational formula. For this sample, $n = 6$ and

$$\Sigma X = 10 + 7 + 6 + 10 + 6 + 15 = 54$$

$$\Sigma X^2 = 10^2 + 7^2 + 6^2 + 10^2 + 6^2 + 15^2 = 546$$

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 546 - \frac{(54)^2}{6}$$

$$= 546 - 486$$

$$= 60$$

STEP 2 **Compute the sample variance** For sample variance, *SS* is divided by the degrees of freedom, $df = n - 1$

$$s^2 = \frac{SS}{n-1} = \frac{60}{5} = 12$$

STEP 3 **Compute the sample standard deviation** Standard deviation is simply the square root of the variance.

PROBLEMS

1. In words, explain what is measured by each of the following:
- **a.** *SS*
- **b.** Variance
- **c.** Standard deviation

2. Can *SS* ever have a value less than zero? Explain your answer.

3. Is it possible to obtain a negative value for the variance or the standard deviation?

4. What does it mean for a sample to have a standard deviation of zero? Describe the scores in such a sample.

5. Explain why the formulas for sample variance and population variance are different.

6. A population has a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$.
- **a.** Would a score of $X = 70$ be considered an extreme value (out in the tail) in this sample?
- **b.** If the standard deviation were $\sigma = 5$, would a score of $X = 70$ be considered an extreme value?

7. On an exam with a mean of $M = 78$, you obtain a score of $X = 84$.
- **a.** Would you prefer a standard deviation of $s = 2$ or $s = 10$? (*Hint:* Sketch each distribution and find the location of your score.)
- **b.** If your score were $X = 72$, would you prefer $s = 2$ or $s = 10$? Explain your answer.

8. A population has a mean of $\mu = 30$ and a standard deviation of $\sigma = 5$.
- **a.** If 5 points were added to every score in the population, what would be the new values for the mean and standard deviation?
- **b.** If every score in the population were multiplied by 3 what would be the new values for the mean and standard deviation?

9.
- **a.** After 3 points have been added to every score in a sample, the mean is found to be $M = 83$ and the standard deviation is $s = 8$. What were the values for the mean and standard deviation for the original sample?
- **b.** After every score in a sample has been multiplied by 4, the mean is found to be $M = 48$ and the standard deviation is $s = 12$. What were the values for the mean and standard deviation for the original sample?

10. A student was asked to compute the mean and standard deviation for the following sample of $n = 5$ scores: 81, 87, 89, 86, and 87. To simplify the arithmetic, the student first subtracted 80 points from each score to obtain a new sample consisting of 1, 7, 9, 6, and 7. The mean and standard deviation for the new sample were then calculated to be $M = 6$ and $s = 3$. What are the values of the mean and standard deviation for the original sample?

11. For the following population of $N = 6$ scores:

11, 0, 2, 9, 9, 5

- **a.** Calculate the range and the standard deviation. (Use either definition for the range.)
- **b.** Add 2 points to each score and compute the range and standard deviation again. Describe how adding a constant to each score influences measures of variability.

12. There are two different formulas or methods that can be used to calculate *SS*.
- **a.** Under what circumstances is the definitional formula easy to use?
- **b.** Under what circumstances is the computational formula preferred?

13. Calculate the mean and *SS* (sum of squared deviations) for each of the following samples. Based on the value for the mean, you should be able to decide which *SS* formula is better to use.

Sample A: 1, 4, 8, 5
Sample B: 3, 0, 9, 4

14. The range is completely determined by the two extreme scores in a distribution. The standard deviation, on the other hand, uses every score.
- **a.** Compute the range (choose either definition) and the standard deviation for the following sample of $n = 5$ scores. Note that there are three scores clustered around the mean in the center of the distribution, and two extreme values.

Scores: 0, 6, 7, 8, 14.

- **b.** Now we break up the cluster in the center of the distribution by moving two of the central scores out to the extremes. Once again compute the range and the standard deviation.

New scores: 0, 0, 7, 14, 14.

- **c.** According to the range, how do the two distributions compare in variability? How do they compare according to the standard deviation?

15. For the data in the following sample:

8, 1, 5, 1, 5

a. Find the mean and the standard deviation.
b. Now change the score of $X = 8$ to $X = 18$, and find the new mean and standard deviation.
c. Describe how one extreme score influences the mean and standard deviation.

16. Calculate *SS*, variance, and standard deviation for the following sample of $n = 4$ scores: 7, 4, 2, 1. (*Note:* The computational formula works well with these scores.)

17. Calculate *SS*, variance, and standard deviation for the following population of $N = 8$ scores: 0, 0, 5, 0, 3, 0, 0, 4. (*Note:* The computational formula works well with these scores.)

18. Calculate *SS*, variance, and standard deviation for the following population of $N = 7$ scores: 8, 1, 4, 3, 5, 3, 4. (*Note:* The definitional formula works well with these scores.)

19. Calculate *SS*, variance, and standard deviation for the following sample of $n = 5$ scores: 9, 6, 2, 2, 6. (Note: The definitional formula works well with these scores.)

20. For the following population of $N = 6$ scores:

3, 1, 4, 3, 3, 4

a. Sketch a histogram showing the population distribution.
b. Locate the value of the population mean in your sketch, and make an estimate of the standard deviation (as done in Example 4.2).
c. Compute *SS*, variance, and standard deviation for the population. (How well does your estimate compare with the actual value of σ?)

21. For the following sample of $n = 7$ scores:

8, 6, 5, 2, 6, 3, 5

a. Sketch a histogram showing the sample distribution.
b. Locate the value of the sample mean in your sketch, and make an estimate of the standard deviation (as done in Example 4.5).
c. Compute *SS*, variance, and standard deviation for the sample. (How well does your estimate compare with the actual value of *s*?)

22. In an extensive study involving thousands of British children, Arden and Plomin (2006) found significantly higher variance in the intelligence scores for males than for females. Following are hypothetical data, similar to the results obtained in the study. Note that the scores are not regular IQ scores but have been standardized so that the entire sample has a mean of $M = 10$ and a standard deviation of $s = 2$.

a. Calculate the mean and the standard deviation for the sample of $n = 8$ females and for the sample of $n = 8$ males.
b. Based on the means and the standard deviations, describe the differences in intelligence scores for males and females.

Female	Male
9	8
11	10
10	11
13	12
8	6
9	10
11	14
9	9

23. In the Preview section at the beginning of this chapter we reported a study by Wegesin and Stern (2004) that found greater consistency (less variability) in the memory performance scores for younger women than for older women. The following data represent memory scores obtained for two women, one older and one younger, over a series of memory trials.

a. Calculate the variance of the scores for each woman.
b. Are the scores for the younger woman more consistent (less variable)?

Younger	Older
8	7
6	5
6	8
7	5
8	7
7	6
8	8
8	5

REVIEW

By completing this part, you should understand and be able to perform basic descriptive statistical procedures. These include:

1. Familiarity with statistical terminology and notation (Chapter 1).
2. The ability to organize a set of scores in a frequency distribution table or a frequency distribution graph (Chapter 2).
3. The ability to summarize and describe a distribution of scores by computing a measure of central tendency (Chapter 3).
4. The ability to summarize and describe a distribution of scores by computing a measure of variability (Chapter 4).

The general goal of descriptive statistics is to simplify a set of data by organizing or summarizing a large set of scores. A frequency distribution table or graph organizes the entire set of scores so that it is possible to see the complete distribution all at once. Measures of central tendency describe the distribution by identifying its center. They also summarize the distribution by condensing all of the individual scores into one value that represents the entire group. Measures of variability describe whether the scores in a distribution are widely scattered or closely clustered. Variability also provides an indication of how accurately a measure of central tendency represents the entire group.

Of the basic skills presented in this part, the most commonly used are calculating the mean and standard deviation for a sample of numerical scores. The following exercises should provide an opportunity to use and reinforce these statistical skills.

REVIEW EXERCISES

1. **a.** What is the general goal for descriptive statistics?

b. How is the goal served by putting scores in a frequency distribution?

c. How is the goal served by computing a measure of central tendency?

d. How is the goal served by computing a measure of variability?

2. In a classic study examining the relationship between heredity and intelligence, Robert Tryon (1940) used a selective breeding program to develop separate strains of "smart rats" and "dumb rats." Tryon started with a large sample of laboratory rats and tested each animal on a maze-learning problem. Based on their error scores for the maze, Tryon selected the brightest rats and the dullest rats from the sample. The brightest males were mated with the brightest females. Similarly, the dullest rats were interbred. This process of testing and selective breeding was continued for several generations until Tryon had established a line of maze-bright rats and a separate line of maze-dull rats. The following data represent results similar to those obtained by Tryon. The data consist of maze-learning error scores for the original sample of laboratory rats and the seventh generation of the maze-bright rats.

Errors Before Solving Maze					
Original Rats			Seventh Generation Maze-Bright Rats		
10	14	7	5	8	7
17	13	12	8	8	6
11	9	20	6	10	4
13	6	15	6	9	8
4	18	10	5	7	9
13	21	6	10	8	6
17	11	14	9	7	8

a. Sketch a polygon showing the distribution of error scores for the sample of original rats. On the same graph, sketch a polygon for the sample of maze-bright rats. (Use two different colors or use a dashed line for one group and a solid line for the other.) Based on the appearance of your graph, describe the differences between the two samples.

b. Calculate the mean error score for each sample. Does the mean difference support your description from part a?

c. Calculate the variance and standard deviation for each sample. Based on the measures of variability, is one group more diverse than the other? Is one group more homogeneous than the other?

This page intentionally left blank

PART

II

Foundations of Inferential Statistics

You should recall from Chapter 1 that statistical methods are classified into two general categories: descriptive statistics, which attempt to organize and summarize data, and inferential statistics, which use the limited information from samples to answer general questions about populations. In most research situations, both kinds of statistics are used to gain a complete understanding of the research results. In Part I of this book we introduced the techniques of descriptive statistics. We now are ready to turn our attention to inferential statistics.

Before we proceed with inferential statistics, however, it is necessary to present some additional information about samples. We know that it is possible to obtain hundreds or even thousands of different samples from the same population. We need to determine how all the different samples are related to each other and how individual samples are related to the population from which they were obtained. Finally, we need a system for designating which samples are representative of their populations and which are not.

In the next four chapters we develop the concepts and skills that form the foundation for inferential statistics. In general, these chapters establish formal, quantitative relationships between samples and populations and introduce a standardized procedure for determining whether the data from a sample justify a conclusion about the population. After we have developed this foundation, we will be prepared to begin inferential statistics. That is, we can begin to look at statistical techniques that use the sample data obtained in research studies as the basis for answering questions about populations.

This page intentionally left blank

CHAPTER

5

z-Scores: Location of Scores and Standardized Distributions

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter and section before proceeding.

- The mean (Chapter 3)
- The standard deviation (Chapter 4)
- Basic algebra (math review, Appendix A)

Preview

A common test of cognitive ability requires participants to search through a visual display and respond to specific targets as quickly as possible. This kind of test is called a perceptual-speed test. Measures of perceptual speed are commonly used for predicting performance on jobs that demand a high level of speed and accuracy. Although many different tests are used, a typical example is shown in Figure 5.1. This task requires the participant to search through the display of digit pairs as quickly as possible and circle each pair that adds up 10. Your score is determined by the amount of time required to complete the task with a correction for the number of errors that you make. One complaint about this kind of paper-and-pencil test is that it is tedious and time consuming to score because a researcher must also search through the entire display to identify errors to determine the participant's level of accuracy. An alternative, proposed by Ackerman and Beier (2007), is a computerized version of the task. The computer version presents a series of digit pairs and participants respond on a touch-sensitive monitor. The computerized test is very reliable and the scores are equivalent to the paper-and-pencil tests in terms of assessing cognitive skill. The advantage of the computerized test is that the computer produces a test score immediately when a participant finishes the test.

Suppose that you took Ackerman and Beier's test and your combined time and errors produced a score of 92. How did you do? Are you faster than average, fairly normal in perceptual speed, or does your score indicate a serious deficit in cognitive skill?

FIGURE 5.1

An example of a perceptual speed task. The participant is asked to search through the display as quickly as possible and circle each pair of digits that add up to 10.

Circle every pair of adjacent numbers that add up to 10.

64	23	19	31	19	46	31	91	83	82	82	46	19	87
11	42	94	87	64	44	19	55	82	46	57	98	39	46
78	73	72	66	63	71	67	42	62	73	45	22	62	99
73	91	52	37	55	97	91	51	44	23	46	64	97	62
97	31	21	49	93	91	89	46	73	82	55	98	12	56
73	82	37	55	89	83	73	27	83	82	73	46	97	62
57	96	46	55	46	19	13	67	73	26	58	64	32	73
23	94	66	55	91	73	67	73	82	55	64	62	46	39
87	11	99	73	56	73	63	73	91	82	63	33	16	88
19	42	62	91	12	82	32	92	73	46	68	19	11	64
93	91	32	82	63	91	46	46	36	55	19	92	62	71

The Problem: Without any frame of reference, a simple raw score provides relatively little information. Specifically, you have no idea how your test score of 92 compares with others who took the same test.

The Solution: Transforming your test score into a z-*score* will identify exactly where you are located in the distribution of scores. In this case, the distribution of scores has a mean of 86.75 and a standard deviation of 10.50. With this additional information, you should realize that your score ($X = 92$) is somewhat higher than average but not extreme. The z-score combines all of this information (your score, the mean, and the standard deviation) into a single number that precisely describes your location relative to the other scores in the distribution.

5.1 INTRODUCTION TO *z*-SCORES

In the previous two chapters, we introduced the concepts of the mean and the standard deviation as methods for describing an entire distribution of scores. Now we shift attention to the individual scores within a distribution. In this chapter, we introduce a statistical technique that uses the mean and the standard deviation to transform each score (X value) into a z-*score,* or a *standard score*. The purpose of z-scores, or standard scores, is to identify and describe the exact location of each score in a distribution.

The following example demonstrates why z-scores are useful and introduces the general concept of transforming X values into z-scores.

EXAMPLE 5.1

Suppose you received a score of $X = 76$ on a statistics exam. How did you do? It should be clear that you need more information to predict your grade. Your score of $X = 76$ could be one of the best scores in the class, or it might be the lowest score in the distribution. To find the location of your score, you must have information about the other scores in the distribution. It would be useful, for example, to know the mean for the class. If the mean were $\mu = 70$, you would be in a much better position than if the mean were $\mu = 85$. Obviously, your position relative to the rest of the class depends on the mean. However, the mean by itself is not sufficient to tell you the exact location of your score. Suppose you know that the mean for the statistics exam is $\mu = 70$ and your score is $X = 76$. At this point, you know that your score is 6 points above the mean, but you still do not know exactly where it is located. Six points may be a relatively big distance and you may have one of the highest scores in the class, or 6 points may be a relatively small distance and you may be only slightly above the average. Figure 5.2 shows two possible distributions of exam scores. Both distributions have a mean of $\mu = 70$, but for one distribution, the standard deviation is $\sigma = 3$, and for the other, $\sigma = 12$. The location of $X = 76$ is highlighted in each of the two distributions. When the standard deviation is $\sigma = 3$, your score of $X = 76$ is in the extreme right-hand tail, the highest score in the distribution. However, in the other distribution, where $\sigma = 12$, your score is only slightly above average. Thus, the relative location of your score within the distribution depends on the standard deviation as well as the mean.

The purpose of the preceding example is to demonstrate that a score *by itself* does not necessarily provide much information about its position within a distribution. These original, unchanged scores that are the direct result of measurement are called *raw scores*. To make raw scores more meaningful, they are often transformed into new values that contain more information. This transformation is one purpose for z-scores. In particular, we transform X values into z-scores so that the resulting z-scores tell exactly where the original scores are located.

A second purpose for z-scores is to *standardize* an entire distribution. A common example of a standardized distribution is the distribution of IQ scores. Although there are several different tests for measuring IQ, the tests usually are standardized so that they have a mean of 100 and a standard deviation of 15. Because all the different tests are standardized, it is possible to understand and compare IQ scores even though they come from different tests. For example, we all understand that an IQ score of 95 is a little below average, *no matter which IQ test was used*. Similarly, an IQ of 145 is extremely high, *no matter which IQ test was used*. In general terms, the process of

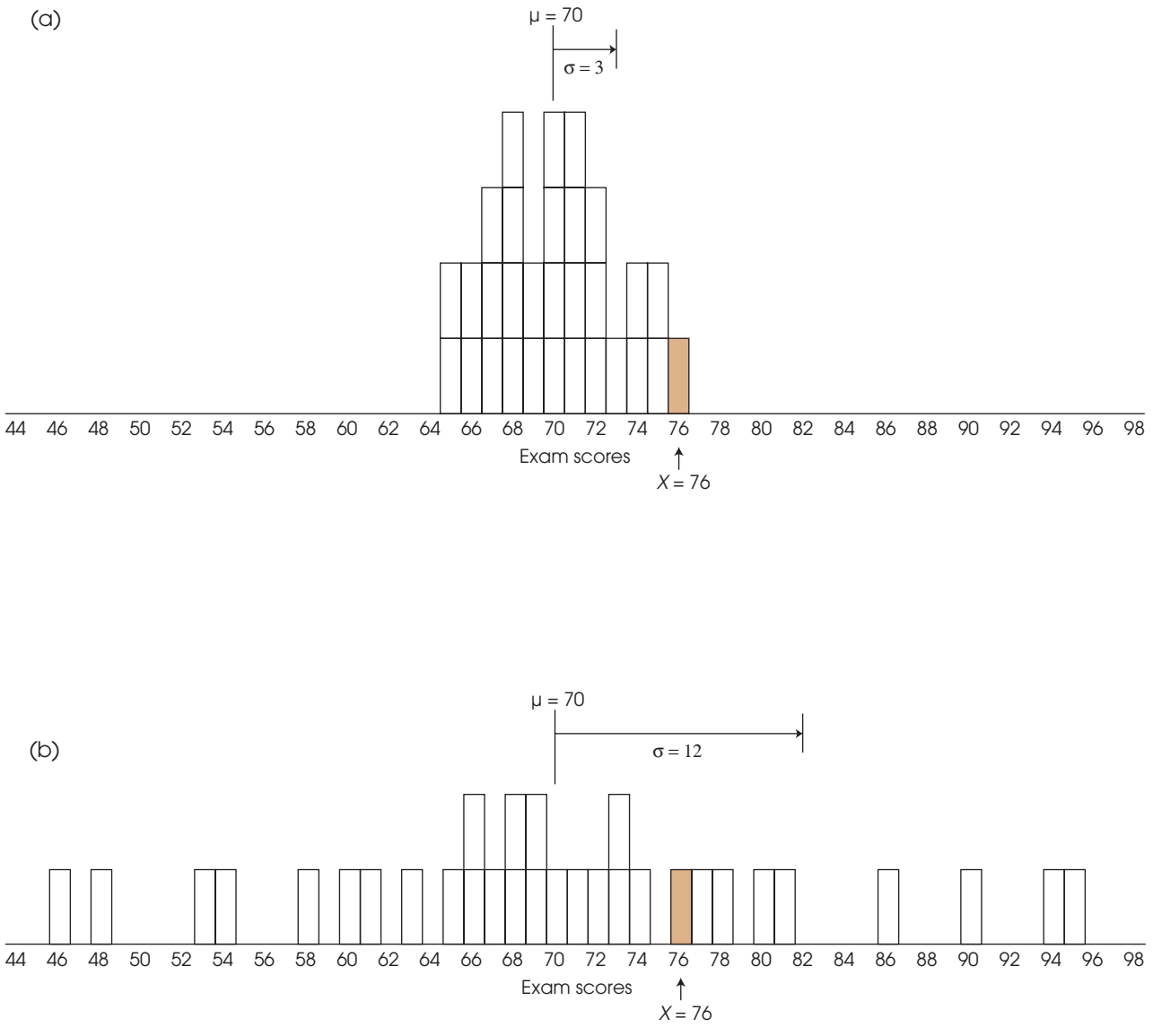


FIGURE 5.2

Two distributions of exam scores. For both distributions, $\mu = 70$, but for one distribution, $\sigma = 3$, and for the other, $\sigma = 12$. The relative position of $X = 76$ is very different for the two distributions.

standardizing takes different distributions and makes them equivalent. The advantage of this process is that it is possible to compare distributions even though they may have been quite different before standardization.

In summary, the process of transforming X values into z-scores serves two useful purposes:

1. Each z-score tells the exact location of the original X value within the distribution.

2. The z-scores form a standardized distribution that can be directly compared to other distributions that also have been transformed into z-scores.

Each of these purposes is discussed in the following sections.

5.2 z-SCORES AND LOCATION IN A DISTRIBUTION

One of the primary purposes of a z-score is to describe the exact location of a score within a distribution. The z-score accomplishes this goal by transforming each X value into a signed number (+ or –) so that

1. The *sign* tells whether the score is located above (+) or below (–) the mean, and
2. The *number* tells the distance between the score and the mean in terms of the number of standard deviations.

Thus, in a distribution of IQ scores with $\mu = 100$ and $\sigma = 15$, a score of $X = 130$ would be transformed into $z = +2.00$. The z value indicates that the score is located above the mean (+) by a distance of 2 standard deviations (30 points).

DEFINITION

A ***z*-score** specifies the precise location of each X value within a distribution. The sign of the z-score (+ or –) signifies whether the score is above the mean (positive) or below the mean (negative). The numerical value of the z-score specifies the distance from the mean by counting the number of standard deviations between X and μ.

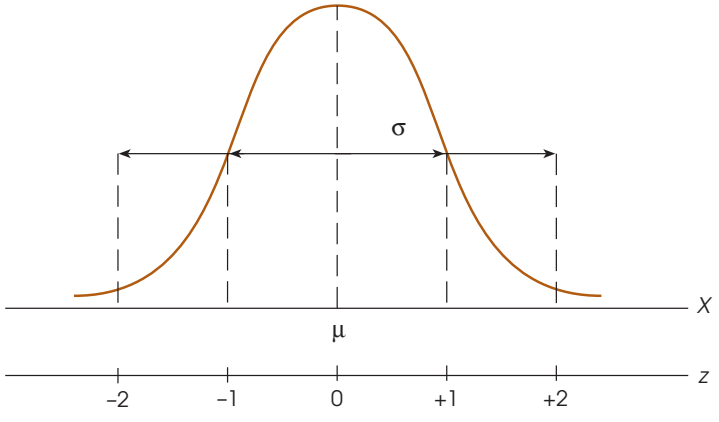
Whenever you are working with z-scores, you should imagine or draw a picture similar to Figure 5.3. Although you should realize that not all distributions are normal, we use the normal shape as an example when showing z-scores for populations.

Notice that a z-score always consists of two parts: a sign (+ or –) and a magnitude. Both parts are necessary to describe completely where a raw score is located within a distribution.

Figure 5.3 shows a population distribution with various positions identified by their z-score values. Notice that all z-scores above the mean are positive and all z-scores below the mean are negative. The sign of a z-score tells you immediately whether the score is located above or below the mean. Also, note that a z-score of $z = +1.00$ corresponds to a position exactly 1 standard deviation above the mean. A z-score of

FIGURE 5.3

The relationship between z-score values and locations in a population distribution.



$z = +2.00$ is always located exactly 2 standard deviations above the mean. The numerical value of the z-score tells you the number of standard deviations it is from the mean. Finally, you should notice that Figure 5.3 does not give any specific values for the population mean or the standard deviation. The locations identified by z-scores are the same for *all distributions,* no matter what mean or standard deviation the distributions may have.

Now we can return to the two distributions shown in Figure 5.2 and use a z-score to describe the position of $X = 76$ within each distribution as follows:

In Figure 5.2(a), with a standard deviation of $\sigma = 3$, the score $X = 76$ corresponds to a z-score of $z = +2.00$. That is, the score is located above the mean by exactly 2 standard deviations.

In Figure 5.2(b), with $\sigma = 12$, the score $X = 76$ corresponds to a z-score of $z = +0.50$. In this distribution, the score is located above the mean by exactly $\frac{1}{2}$ standard deviation.

LEARNING CHECK

1. Identify the z-score value corresponding to each of the following locations in a distribution.

a. Below the mean by 2 standard deviations.

b. Above the mean by $\frac{1}{2}$ standard deviation.

c. Below the mean by 1.50 standard deviations.

2. Describe the location in the distribution for each of the following z-scores. (For example, $z = +1.00$ is located above the mean by 1 standard deviation.)

a. $z = -1.50$ **b.** $z = 0.25$ **c.** $z = -2.50$ **d.** $z = 0.50$

3. For a population with $\mu = 30$ and $\sigma = 8$, find the z-score for each of the following scores:

a. $X = 32$ **b.** $X = 26$ **c.** $X = 42$

4. For a population with $\mu = 50$ and $\sigma = 12$, find the X value corresponding to each of the following z-scores:

a. $z = –0.25$ **b.** $z = 2.00$ **c.** $z = 0.50$

ANSWERS

1. a. $z = –2.00$ **b.** $z = +0.50$ **c.** $z = –1.50$

2. a. Below the mean by $1\frac{1}{2}$ standard deviations.

b. Above the mean by $\frac{1}{4}$ standard deviation.

c. Below the mean by $2\frac{1}{2}$ standard deviations.

d. Above the mean by $\frac{1}{2}$ standard deviation.

3. a. $z = +0.25$ **b.** $z = –0.50$ **c.** $z = +1.50$

4. a. $X = 47$ **b.** $X = 74$ **c.** $X = 56$

THE z-SCORE FORMULA

The z-score definition is adequate for transforming back and forth from X values to z-scores as long as the arithmetic is easy to do in your head. For more complicated values, it is best to have an equation to help structure the calculations. Fortunately, the relationship between X values and z-scores is easily expressed in a formula. The formula for transforming scores into z-scores is

$$z = \frac{X - \mu}{\sigma} \tag{5.1}$$

The numerator of the equation, $X - \mu$, is a *deviation score* (Chapter 4, page 107); it measures the distance in points between X and μ and indicates whether X is located above or below the mean. The deviation score is then divided by σ because we want the z-score to measure distance in terms of standard deviation units. The formula performs exactly the same arithmetic that is used with the z-score definition, and it provides a structured equation to organize the calculations when the numbers are more difficult. The following examples demonstrate the use of the z-score formula.

EXAMPLE 5.2

A distribution of scores has a mean of $\mu = 100$ and a standard deviation of $\sigma = 10$.

What z-score corresponds to a score of $X = 130$ in this distribution?

According to the definition, the z-score has a value of $+3$ because the score is located above the mean by exactly 3 standard deviations. Using the z-score formula, we obtain

$$z = \frac{X - \mu}{\sigma} = \frac{130 - 100}{10} = \frac{30}{10} = 3.00$$

The formula produces exactly the same result that is obtained using the z-score definition.

EXAMPLE 5.3

A distribution of scores has a mean of $\mu = 86$ and a standard deviation of $\sigma = 7$. What z-score corresponds to a score of $X = 95$ in this distribution?

Note that this problem is not particularly easy, especially if you try to use the z-score definition and perform the calculations in your head. However, the z-score formula organizes the numbers and allows you to finish the final arithmetic with your calculator. Using the formula, we obtain

$$z = \frac{X - \mu}{\sigma} = \frac{95 - 86}{7} = \frac{9}{7} = 1.29$$

According to the formula, a score of $X = 95$ corresponds to $z = 1.29$. The z-score indicates a location that is above the mean (positive) by slightly more than 1 standard deviation.

When you use the z-score formula, it can be useful to pay attention to the definition of a z-score as well. For example, we used the formula in Example 5.3 to calculate the z-score corresponding to $X = 95$, and obtained $z = 1.29$. Using the z-score definition, we note that $X = 95$ is located above the mean by 9 points, which is slightly more than one standard deviation ($\sigma = 7$). Therefore, the z-score should be positive and have a value slightly greater than 1.00. In this case, the answer predicted by the definition is in perfect agreement with the calculation. However, if the calculations produce a different value, for example $z = 0.78$, you should realize that this answer is not consistent with the definition of a z-score. In this case, an error has been made and you should double check the calculations.

DETERMINING A RAW SCORE (X) FROM A z-SCORE

Although the z-score equation (Formula 5.1) works well for transforming X values into z-scores, it can be awkward when you are trying to work in the opposite direction and change z-scores back into X values. In general it is easier to use the definition of a z-score, rather than a formula, when you are changing z-scores into X values. Remember, the z-score describes exactly where the score is located by identifying the direction and

distance from the mean. It is possible, however, to express this definition as a formula, and we use a sample problem to demonstrate how the formula can be created.

For a distribution with a mean of $\mu = 60$ and $\sigma = 5$, what X value corresponds to a z-score of $z = -3.00$?

To solve this problem, we use the z-score definition and carefully monitor the step-by-step process. The value of the z-score indicates that X is located below the mean by a distance equal to 3 standard deviations. Thus, the first step in the calculation is to determine the distance corresponding to 3 standard deviations. For this problem, the standard deviation is $\sigma = 5$ points, so 3 standard deviations is $3(5) = 15$ points. The next step is to find the value of X that is located below the mean by 15 points. With a mean of $\mu = 60$, the score is

$$X = \mu - 15 = 60 - 15 = 45$$

The two steps can be combined to form a single formula:

$$X = \mu + z\sigma \tag{5.2}$$

In the formula, the value of $z\sigma$ is the *deviation* of X and determines both the direction and the size of the distance from the mean. In this problem, $z\sigma = (-3)(5) = -15$, or 15 points below the mean. Formula 5.2 simply combines the mean and the deviation from the mean to determine the exact value of X.

Finally, you should realize that Formula 5.1 and Formula 5.2 are actually two different versions of the same equation. If you begin with either formula and use algebra to shuffle the terms around, you soon end up with the other formula. We leave this as an exercise for those who want to try it.

OTHER RELATIONSHIPS BETWEEN *z*, *X*, μ, AND σ

In most cases, we simply transform scores (X values) into z-scores, or change z-scores back into X values. However, you should realize that a z-score establishes a relationship between the score, the mean, and the standard deviation. This relationship can be used to answer a variety of different questions about scores and the distributions in which they are located. The following two examples demonstrate some possibilities.

EXAMPLE 5.4

In a population with a mean of $\mu = 65$, a score of $X = 59$ corresponds to $z = -2.00$. What is the standard deviation for the population?

To answer the question, we begin with the z-score value. A z-score of –2.00 indicates that the corresponding score is located below the mean by a distance of 2 standard deviations. You also can determine that the score ($X = 59$) is located below the mean ($\mu = 65$) by a distance of 6 points. Thus, 2 standard deviations correspond to a distance of 6 points, which means that 1 standard deviation must be $\sigma = 3$ points.

EXAMPLE 5.5

In a population with a standard deviation of $\sigma = 4$, a score of $X = 33$ corresponds to $z = +1.50$. What is the mean for the population?

Again, we begin with the z-score value. In this case, a z-score of +1.50 indicates that the score is located above the mean by a distance corresponding to 1.50 standard deviations. With a standard deviation of $\sigma = 4$, this distance is $(1.50)(4) = 6$ points.

Thus, the score is located 6 points above the mean. The score is $X = 33$, so the mean must be $\mu = 27$.

Many students find problems like those in Examples 5.4 and 5.5 easier to understand if they draw a picture showing all of the information presented in the problem. For the problem in Example 5.4, the picture would begin with a distribution that has a mean of $\mu = 65$ (we use a normal distribution, which is shown in Figure 5.4). The value of the standard deviation is unknown, but you can add arrows to the sketch pointing outward from the mean for a distance corresponding to 1 standard deviation. Finally, use standard deviation arrows to identify the location of $z = -2.00$ (2 standard deviations below the mean) and add $X = 59$ at that location. All of these factors are shown in Figure 5.4. In the figure, it is easy to see that $X = 59$ is located 6 points below the mean, and that the 6-point distance corresponds to exactly 2 standard deviations. Again, if 2 standard deviations equal 6 points, then 1 standard deviation must be $\sigma = 3$ points.

LEARNING CHECK

1. For a distribution with $\mu = 40$ and $\sigma = 12$, find the z-score for each of the following scores.
   a. $X = 36$ **b.** $X = 46$ **c.** $X = 56$
2. For a distribution with $\mu = 40$ and $\sigma = 12$, find the X value corresponding to each of the following z-scores.
   a. $z = 1.50$ **b.** $z = -1.25$ **c.** $z = \frac{1}{3}$
3. In a distribution with $\mu = 50$, a score of $X = 42$ corresponds to $z = -2.00$. What is the standard deviation for this distribution?
4. In a distribution with $\sigma = 12$, a score of $X = 56$ corresponds to $z = -0.25$. What is the mean for this distribution?

ANSWERS

1. **a.** $z = -0.33$ (or $-\frac{1}{3}$) **b.** $z = 0.50$ **c.** $z = 1.33$ ($+1\frac{1}{3}$)
2. **a.** $X = 58$ **b.** $X = 25$ **c.** $X = 44$
3. $\sigma = 4$
4. $\mu = 59$

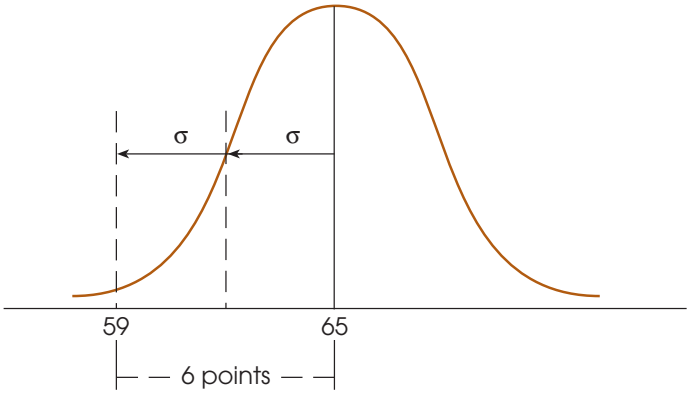


FIGURE 5.4
A visual presentation of the question in Example 5.4. If 2 standard deviations correspond to a 6-point distance, then 1 standard deviation must equal 3 points.

5.3 USING z-SCORES TO STANDARDIZE A DISTRIBUTION

It is possible to transform every X value in a distribution into a corresponding z-score. The result of this process is that the entire distribution of X values is transformed into a distribution of z-scores (Figure 5.5). The new distribution of z-scores has characteristics that make the *z-score transformation* a very useful tool. Specifically, if every X value is transformed into a z-score, then the distribution of z-scores will have the following properties:

1. Shape. The distribution of z-scores will have exactly the same shape as the original distribution of scores. If the original distribution is negatively skewed, for example, then the z-score distribution will also be negatively skewed. If the original distribution is normal, the distribution of z-scores will also be normal. Transforming raw scores into z-scores does not change anyone's position in the distribution. For example, any raw score that is above the mean by 1 standard deviation will be transformed to a z-score of $+1.00$, which is still above the mean by 1 standard deviation. Transforming a distribution from X values to z values does not move scores from one position to another; the procedure simply relabels each score (see Figure 5.5). Because each individual score stays in its same position within the distribution, the overall shape of the distribution does not change.

2. The mean. The z-score distribution will *always* have a mean of zero. In Figure 5.5, the original distribution of X values has a mean of $\mu = 100$. When this value, $X = 100$, is transformed into a z-score, the result is

$$z = \frac{X - \mu}{\sigma} = \frac{100 - 100}{10} = 0$$

Thus, the original population mean is transformed into a value of zero in the z-score distribution. The fact that the z-score distribution has a mean of zero makes the mean a convenient reference point. Recall from the definition of z-scores that all

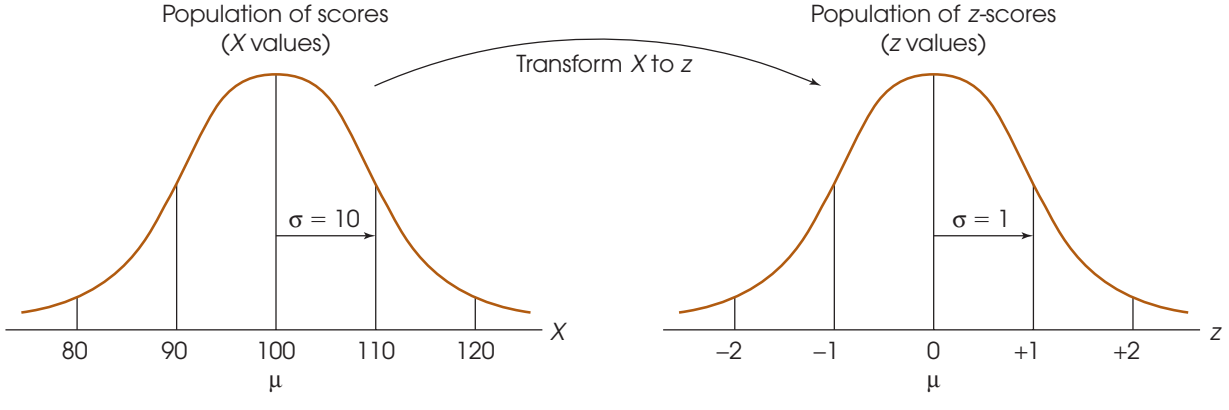


FIGURE 5.5

An entire population of scores is transformed into z-scores. The transformation does not change the shape of the population, but the mean is transformed into a value of 0 and the standard deviation is transformed to a value of 1.

positive z-scores are above the mean and all negative z-scores are below the mean. In other words, for z-scores, $\mu = 0$.

3. The standard deviation. The distribution of z-scores will *always* have a standard deviation of 1. In Figure 5.5, the original distribution of X values has $\mu = 100$ and $\sigma = 10$. In this distribution, a value of $X = 110$ is above the mean by exactly 10 points or 1 standard deviation. When $X = 110$ is transformed, it becomes $z = +1.00$, which is above the mean by exactly 1 point in the z-score distribution. Thus, the standard deviation corresponds to a 10-point distance in the X distribution and is transformed into a 1-point distance in the z-score distribution. The advantage of having a standard deviation of 1 is that the numerical value of a z-score is exactly the same as the number of standard deviations from the mean. For example, a z-score of $z = 1.50$ is exactly 1.50 standard deviations from the mean.

In Figure 5.5, we showed the z-score transformation as a process that changed a distribution of X values into a new distribution of z-scores. In fact, there is no need to create a whole new distribution. Instead, you can think of the z-score transformation as simply *relabeling* the values along the X-axis. That is, after a z-score transformation, you still have the same distribution, but now each individual is labeled with a z-score instead of an X value. Figure 5.6 demonstrates this concept with a single distribution that has two sets of labels: the X values along one line and the corresponding z-scores along another line. Notice that the mean for the distribution of z-scores is zero and the standard deviation is 1.

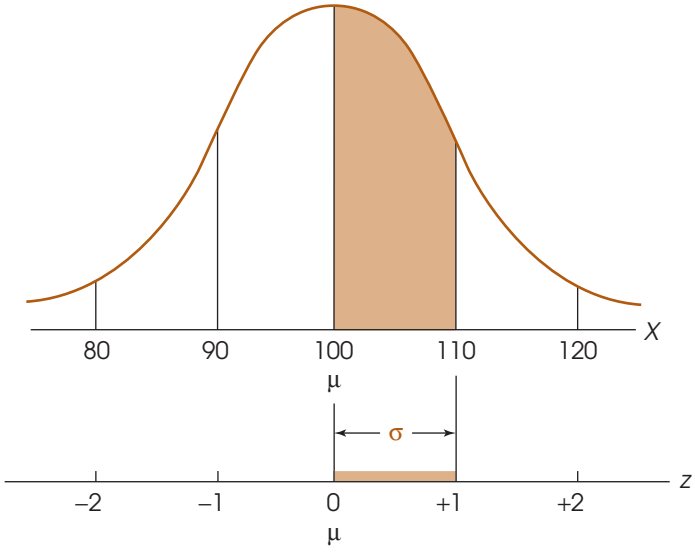
When *any* distribution (with any mean or standard deviation) is transformed into z-scores, the resulting distribution will always have a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. Because all z-score distributions have the same mean and the same standard deviation, the z-score distribution is called a *standardized distribution*.

DEFINITION

A **standardized distribution** is composed of scores that have been transformed to create predetermined values for μ and σ. Standardized distributions are used to make dissimilar distributions comparable.

FIGURE 5.6

Following a z-score transformation, the X-axis is relabeled in z-score units. The distance that is equivalent to 1 standard deviation on the X-axis ($\sigma = 10$ points in this example) corresponds to 1 point on the z-score scale.



A z-score distribution is an example of a standardized distribution with $\mu = 0$ and $\sigma = 1$. That is, when any distribution (with any mean or standard deviation) is transformed into z-scores, the transformed distribution will always have $\mu = 0$ and $\sigma = 1$.

DEMONSTRATION OF A z-SCORE TRANSFORMATION

Although the basic characteristics of a z-score distribution have been explained logically, the following example provides a concrete demonstration that a z-score transformation creates a new distribution with a mean of zero, a standard deviation of 1, and the same shape as the original population.

EXAMPLE 5.6

We begin with a population of $N = 6$ scores consisting of the following values: 0, 6, 5, 2, 3, 2. This population has a mean of $\mu = \frac{18}{6} = 3$ and a standard deviation of $\sigma = 2$ (check the calculations for yourself).

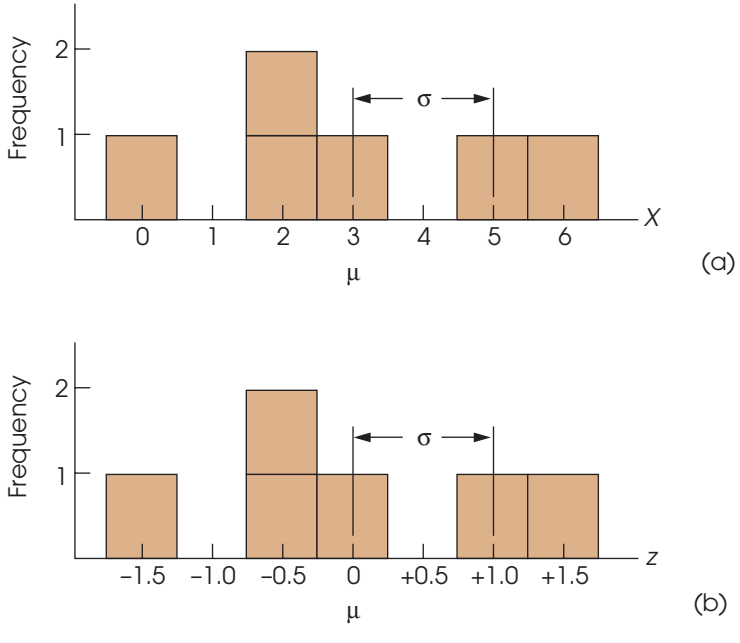
Each of the X values in the original population is then transformed into a z-score as summarized in the following table.

$X = 0$	Below the mean by $1\frac{1}{2}$ standard deviations	$z = -1.50$
$X = 6$	Above the mean by $1\frac{1}{2}$ standard deviations	$z = +1.50$
$X = 5$	Above the mean by 1 standard deviation	$z = +1.00$
$X = 2$	Below the mean by $\frac{1}{2}$ standard deviation	$z = -0.50$
$X = 3$	Exactly equal to the mean—zero deviation	$z = 0$
$X = 2$	Below the mean by $\frac{1}{2}$ standard deviation	$z = -0.50$

The frequency distribution for the original population of X values is shown in Figure 5.7(a) and the corresponding distribution for the z-scores is shown in Figure 5.7(b). A simple comparison of the two distributions demonstrates the results of a z-score transformation.

FIGURE 5.7

Transforming a distribution of raw scores (a) into z-scores (b) will not change the shape of the distribution.



1. The two distributions have exactly the same shape. Each individual has exactly the same relative position in the X distribution and in the z-score distribution.
2. After the transformation to z-scores, the mean of the distribution becomes $\mu = 0$. For these z-scores values, $N = 6$ and $\Sigma z = -1.50 + 1.50 + 1.00 + -0.50 + 0 + -0.50 = 0$. Thus, the mean for the z-scores is $\mu = \Sigma z/N = 0/6 = 0$.

   Note that the individual with a score of $X = 3$ is located exactly at the mean in the X distribution and this individual is transformed into $z = 0$, exactly at the mean in the z-distribution.
3. After the transformation, the standard deviation becomes $\sigma = 1$. For these z-scores, $\Sigma z = 0$ and

$$\begin{aligned}\Sigma z^2 &= (-1.50)^2 + (1.50)^2 + (1.00)^2 + (-0.50)^2 + (0)^2 + (-0.50)^2 \\ &= 2.25 + 2.25 + 1.00 + 0.25 + 0 + 0.25 \\ &= 6.00\end{aligned}$$

Using the computational formula for SS, substituting z in place of X, we obtain

$$SS = \Sigma z^2 - \frac{(\Sigma z)^2}{N} = 6 - \frac{(0)^2}{6} = 6.00$$

For these z-scores, the variance is $\sigma^2 = \frac{SS}{N} = \frac{6}{6} = 1.00$ and the standard deviation is

$$\sigma = \sqrt{1.00} = 1.00$$

Note that the individual with $X = 5$ is located above the mean by 2 points, which is exactly one standard deviation in the X distribution. After transformation, this individual has a z-score that is located above the mean by 1 point, which is exactly one standard deviation in the z-score distribution.

USING z-SCORES TO MAKE COMPARISONS

One advantage of standardizing distributions is that it makes it possible to compare different scores or different individuals even though they come from completely different distributions. Normally, if two scores come from different distributions, it is impossible to make any direct comparison between them. Suppose, for example, Dave received a score of $X = 60$ on a psychology exam and a score of $X = 56$ on a biology test. For which course should Dave expect the better grade?

Because the scores come from two different distributions, you cannot make any direct comparison. Without additional information, it is even impossible to determine whether Dave is above or below the mean in either distribution. Before you can begin to make comparisons, you must know the values for the mean and standard deviation for each distribution. Suppose the biology scores had $\mu = 48$ and $\sigma = 4$, and the psychology scores had $\mu = 50$ and $\sigma = 10$. With this new information, you could sketch the two distributions, locate Dave's score in each distribution, and compare the two locations.

Instead of drawing the two distributions to determine where Dave's two scores are located, we simply can compute the two z-scores to find the two locations. For psychology, Dave's z-score is

$$z = \frac{X - \mu}{\sigma} = \frac{60 - 50}{10} = \frac{10}{10} = +1.0$$

Be sure to use the μ and σ values for the distribution to which X belongs.

For biology, Dave's z-score is

$$z = \frac{56 - 48}{4} = \frac{8}{4} = +2.0$$

Note that Dave's z-score for biology is +2.0, which means that his test score is 2 standard deviations above the class mean. On the other hand, his z-score is +1.0 for psychology, or 1 standard deviation above the mean. In terms of relative class standing, Dave is doing much better in the biology class.

Notice that we cannot compare Dave's two exam scores ($X = 60$ and $X = 56$) because the scores come from different distributions with different means and standard deviations. However, we can compare the two z-scores because all distributions of z-scores have the same mean ($\mu = 0$) and the same standard deviation ($\sigma = 1$).

LEARNING CHECK

1. A normal-shaped distribution with $\mu = 40$ and $\sigma = 8$ is transformed into z-scores. Describe the shape, the mean, and the standard deviation for the resulting distribution of z-scores.
2. What is the advantage of having a mean of $\mu = 0$ for a distribution of z-scores?
3. A distribution of English exam scores has $\mu = 70$ and $\sigma = 4$. A distribution of history exam scores has $\mu = 60$ and $\sigma = 20$. For which exam would a score of $X = 78$ have a higher standing? Explain your answer.
4. A distribution of English exam scores has $\mu = 50$ and $\sigma = 12$. A distribution of history exam scores has $\mu = 58$ and $\sigma = 4$. For which exam would a score of $X = 62$ have a higher standing? Explain your answer.

ANSWERS

1. The z-score distribution would be normal with a mean of 0 and a standard deviation of 1.
2. With a mean of zero, all positive scores are above the mean and all negative scores are below the mean.
3. For the English exam, $X = 78$ corresponds to $z = 2.00$, which is a higher standing than $z = 0.90$ for the history exam.
4. The score $X = 62$ corresponds to $z = +1.00$ in both distributions. The score has exactly the same standing for both exams.

5.4 OTHER STANDARDIZED DISTRIBUTIONS BASED ON z-SCORES

TRANSFORMING z-SCORES TO A DISTRIBUTION WITH A PREDETERMINED μ AND σ

Although z-score distributions have distinct advantages, many people find them cumbersome because they contain negative values and decimals. For this reason, it is common to standardize a distribution by transforming the scores into a new distribution with a predetermined mean and standard deviation that are whole round numbers. The goal is to create a new (standardized) distribution that has "simple" values for the mean and standard deviation but does not change any individual's location within the distribution. Standardized scores of this type are frequently used in psychological or educational testing. For example, raw scores of the Scholastic Aptitude Test (SAT) are

transformed to a standardized distribution that has $\mu = 500$ and $\sigma = 100$. For intelligence tests, raw scores are frequently converted to standard scores that have a mean of 100 and a standard deviation of 15. Because most IQ tests are standardized so that they have the same mean and standard deviation, it is possible to compare IQ scores even though they may come from different tests.

The procedure for standardizing a distribution to create new values for μ and σ involves two-steps:

1. The original raw scores are transformed into *z*-scores.
2. The *z*-scores are then transformed into new *X* values so that the specific μ and σ are attained.

This procedure ensures that each individual has exactly the same *z*-score location in the new distribution as in the original distribution. The following example demonstrates the standardization procedure.

EXAMPLE 5.7

An instructor gives an exam to a psychology class. For this exam, the distribution of raw scores has a mean of $\mu = 57$ with $\sigma = 14$. The instructor would like to simplify the distribution by transforming all scores into a new, standardized distribution with $\mu = 50$ and $\sigma = 10$. To demonstrate this process, we consider what happens to two specific students: Maria, who has a raw score of $X = 64$ in the original distribution; and Joe, whose original raw score is $X = 43$.

STEP 1 Transform each of the original raw scores into *z*-scores. For Maria, $X = 64$, so her *z*-score is

$$z = \frac{X - \mu}{\sigma} = \frac{64 - 57}{14} = +0.5$$

For Joe, $X = 43$, and his *z*-score is

$$z = \frac{X - \mu}{\sigma} = \frac{43 - 57}{14} = -1.0$$

Remember: The values of μ and σ are for the distribution from which *X* was taken.

STEP 2 Change each *z*-score into an *X* value in the new standardized distribution that has a mean of $\mu = 50$ and a standard deviation of $\sigma = 10$.

Maria's *z*-score, $z = +0.50$, indicates that she is located above the mean by $\frac{1}{2}$ standard deviation. In the new, standardized distribution, this location corresponds to $X = 55$ (above the mean by 5 points).

Joe's *z*-score, $z = -1.00$, indicates that he is located below the mean by exactly 1 standard deviation. In the new distribution, this location corresponds to $X = 40$ (below the mean by 10 points).

The results of this two-step transformation process are summarized in Table 5.1. Note that Joe, for example, has exactly the same *z*-score ($z = -1.00$) in both the original distribution and the new standardized distribution. This means that Joe's position relative to the other students in the class has not changed.

Figure 5.8 provides another demonstration of the concept that standardizing a distribution does not change the individual positions within the distribution. The figure shows the original exam scores from Example 5.7, with a mean of $\mu = 57$ and a

TABLE 5.1

A demonstration of how two individual scores are changed when a distribution is standardized. See Example 5.7.

	Original Scores $\mu = 57$ and $\sigma = 14$		*z*-Score Location		Standardized Scores $\mu = 50$ and $\sigma = 10$
Maria	$X = 64$	⟶	$z = +0.50$	⟶	$X = 55$
Joe	$X = 43$	⟶	$z = -1.00$	⟶	$X = 40$

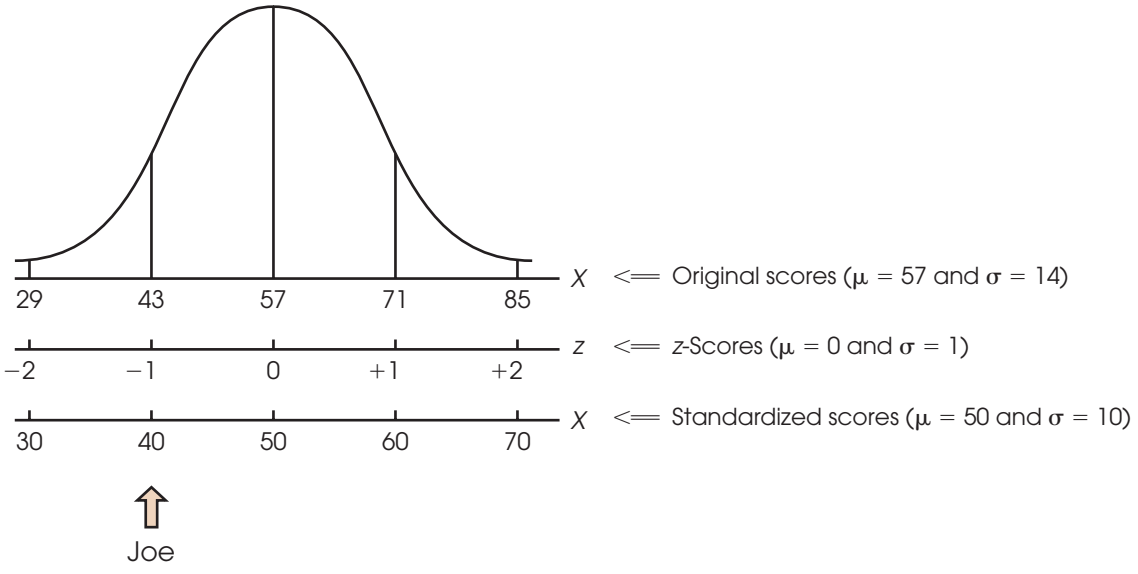


FIGURE 5.8

The distribution of exam scores from Example 5.7. The original distribution was standardized to produce a new distribution with $\mu = 50$ or $\sigma = 10$. Note that each individual is identified by an original score, a *z*-score, and a new, standardized score. For example, Joe has an original score of 43, a *z*-score of –1.00, and a standardized score of 40.

standard deviation of $\sigma = 14$. In the original distribution, Joe is located at a score of $X = 43$. In addition to the original scores, we have included a second scale showing the *z*-score value for each location in the distribution. In terms of *z*-scores, Joe is located at a value of $z = -1.00$. Finally, we have added a third scale showing the *standardized scores,* for which the mean is $\mu = 50$ and the standard deviation is $\sigma = 10$. For the standardized scores, Joe is located at $X = 40$. Note that Joe is always in the same place in the distribution. The only thing that changes is the number that is assigned to Joe: For the original scores, Joe is at 43; for the *z*-scores, Joe is at –1.00; and for the standardized scores, Joe is at 40.

LEARNING CHECK

1. A population of scores has $\mu = 73$ and $\sigma = 8$. If the distribution is standardized to create a new distribution with $\mu = 100$ and $\sigma = 20$, what are the new values for each of the following scores from the original distribution?

a. $X = 65$ **b.** $X = 71$ **c.** $X = 81$ **d.** $X = 83$

2. A population with a mean of $\mu = 44$ and a standard deviation of $\sigma = 6$ is standardized to create a new distribution with $\mu = 50$ and $\sigma = 10$.

a. What is the new standardized value for a score of $X = 47$ from the original distribution?

b. One individual has a new standardized score of $X = 65$. What was this person's score in the original distribution?

ANSWERS

1. a. $z = -1.00$, $X = 80$ **b.** $z = -0.25$, $X = 95$

c. $z = 1.00$, $X = 120$ **d.** $z = 1.25$, $X = 125$

2. a. $X = 47$ corresponds to z $= +0.50$ in the original distribution. In the new distribution, the corresponding score is $X = 55$.

b. In the new distribution, $X = 65$ corresponds to $z = +1.50$. The corresponding score in the original distribution is $X = 53$.

5.5 COMPUTING *z*-SCORES FOR A SAMPLE

Although z-scores are most commonly used in the context of a population, the same principles can be used to identify individual locations within a sample. The definition of a z-score is the same for a sample as for a population, provided that you use the sample mean and the sample standard deviation to specify each z-score location. Thus, for a sample, each X value is transformed into a z-score so that

1. The sign of the z-score indicates whether the X value is above (+) or below (–) the sample mean, and
2. The numerical value of the z-score identifies the distance from the sample mean by measuring the number of sample standard deviations between the score (X) and the sample mean (M).

See the population equations (5.1 and 5.2) on pages 142 and 144 for comparison.

Expressed as a formula, each X value in a sample can be transformed into a z-score as follows:

$$z = \frac{X - M}{s} \tag{5.3}$$

Similarly, each z-score can be transformed back into an X value, as follows:

$$X = M + zs \tag{5.4}$$

EXAMPLE 5.8

In a sample with a mean of $M = 40$ and a standard deviation of $s = 10$, what is the z-score corresponding to $X = 35$ and what is the X value corresponding to $z = +2.00$?

The score, $X = 35$, is located below the mean by 5 points, which is exactly half of the standard deviation. Therefore, the corresponding z-score is $z = -0.50$. The z-score, $z = +2.00$, corresponds to a location above the mean by 2 standard deviations. With a standard deviation of $s = 10$, this is distance of 20 points. The score that is located 20 points above the mean is $X = 60$. Note that it is possible to find these answers using either the z-score definition or one of the equations (5.3 or 5.4).

STANDARDIZING A SAMPLE DISTRIBUTION

If all the scores in a sample are transformed into z-scores, the result is a sample of z-scores. The transformed distribution of z-scores will have the same properties that exist when a population of X values is transformed into z-scores. Specifically,

1. The sample of z-scores will have the same shape as the original sample of scores.
2. The sample of z-scores will have a mean of $M_z = 0$.
3. The sample of z-scores will have a standard deviation of $s_z = 1$.

Note that the set of z-scores is still considered to be a sample (just like the set of X values) and the sample formulas must be used to compute variance and standard deviation. The following example demonstrates the process of transforming the scores from a sample into z-scores.

EXAMPLE 5.9

We begin with a sample of $n = 5$ scores: 0, 2, 4, 4, 5. With a few simple calculations, you should be able to verify that the sample mean is $M = 3$, the sample variance is $s^2 = 4$, and the sample standard deviation is $s = 2$. Using the sample mean and sample standard deviation, we can convert each X value into a z-score. For example, $X = 5$ is located above the mean by 2 points. Thus, $X = 5$ is above the mean by exactly 1 standard deviation and has a z-score of $z = +1.00$. The z-scores for the entire sample are shown in the following table.

X	z
0	−1.50
2	−0.50
4	+0.50
4	+0.50
5	+1.00

Again, a few simple calculations demonstrate that the sum of the z-score values is $\sigma z = 0$, so the mean is $M_z = 0$.

Because the mean is zero, each z-score value is its own deviation from the mean. Therefore, the sum of the squared deviations is simply the sum of the squared z-scores. For this sample of z-scores,

$$SS = \Sigma z^2 = (-1.50)^2 + (-0.50)^2 + (+0.50)^2 + (0.50)^2 + (+1.00)^2$$
$$= 2.25 + 0.25 + 0.25 + 0.25 + 1.00$$
$$= 4.00$$

The variance for the sample of z-scores is

$$s_z^2 = \frac{SS}{n-1} = \frac{4}{4} = 1.00$$

Finally, the standard deviation for the sample of z-scores is. $s_z^2 = \sqrt{1.00} = 1.00$. As always, the distribution of z-scores has a mean of 0 and a standard deviation of 1.

Notice that the set of z-scores is considered to be a sample and the variance is computed using the sample formula with $df = n - 1$.

5.6 LOOKING AHEAD TO INFERENTIAL STATISTICS

Recall that inferential statistics are techniques that use the information from samples to answer questions about populations. In later chapters, we use inferential statistics to help interpret the results from research studies. A typical research study begins with a question about how a treatment will affect the individuals in a population. Because it is usually impossible to study an entire population, the researcher selects a sample and administers the treatment to the individuals in the sample. This general research situation is shown in Figure 5.9. To evaluate the effect of the treatment, the researcher simply compares the treated sample with the original population. If the individuals in the sample are noticeably different from the individuals in the original population, the researcher has evidence that the treatment has had an effect. On the other hand, if the sample is not noticeably different from the original population, it would appear that the treatment has had no effect.

Notice that the interpretation of the research results depends on whether the sample is *noticeably different* from the population. One technique for deciding whether a sample is noticeably different is to use z-scores. For example, an individual with a z-score near 0 is located in the center of the population and would be considered to be a fairly typical or representative individual. However, an individual with an extreme z-score, beyond +2.00 or –2.00 for example, would be considered noticeably different from most of the individuals in the population. Thus, we can use z-scores to help decide whether the treatment has caused a change. Specifically, if the individuals who receive the treatment in a research study tend to have extreme z-scores, we can conclude that the treatment does appear to have an effect. The following example demonstrates this process.

EXAMPLE 5.10

A researcher is evaluating the effect of a new growth hormone. It is known that regular adult rats weigh an average of $\mu = 400$ grams. The weights vary from rat to rat, and the distribution of weights is normal with a standard deviation

FIGURE 5.9

A diagram of a research study. The goal of the study is to evaluate the effect of a treatment. A sample is selected from the population and the treatment is administered to the sample. If, after treatment, the individuals in the sample are noticeably different from the individuals in the original population, then we have evidence that the treatment does have an effect.

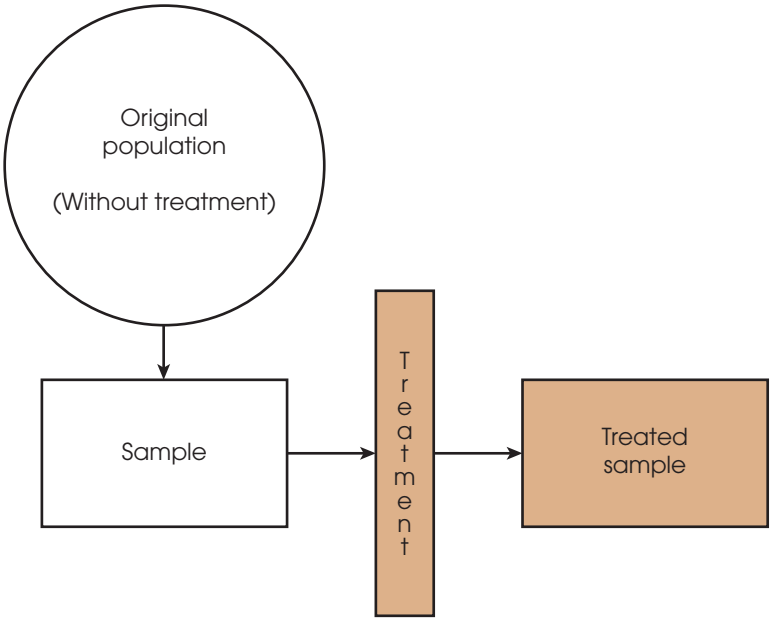
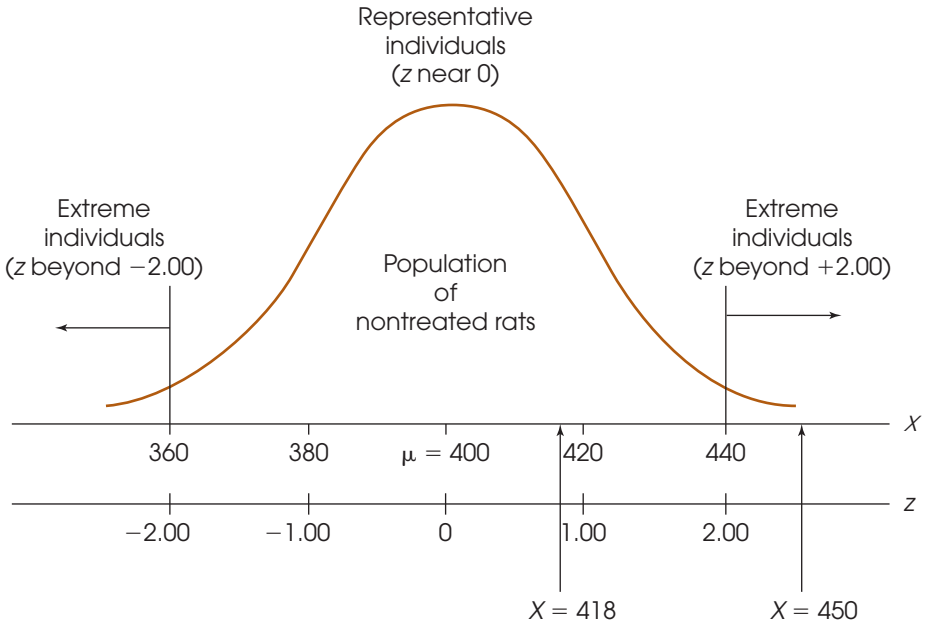


FIGURE 5.10

The distribution of weights for the population of adult rats. Note that individuals with z-scores near 0 are typical or representative. However, individuals with z-scores beyond +2.00 or −2.00 are extreme and noticeably different from most of the others in the distribution.



of $\sigma = 20$ grams. The population distribution is shown in Figure 5.10. The researcher selects one newborn rat and injects the rat with the growth hormone. When the rat reaches maturity, it is weighed to determine whether there is any evidence that the hormone has an effect.

First, assume that the hormone-injected rat weighs $X = 418$ grams. Although this is more than the average nontreated rat ($\mu = 400$ grams), is it convincing evidence that the hormone has an effect? If you look at the distribution in Figure 5.10, you should realize that a rat weighing 418 grams is not noticeably different from the regular rats that did not receive any hormone injection. Specifically, our injected rat would be located near the center of the distribution for regular rats with a z-score of

$$z = \frac{X - \mu}{\sigma} = \frac{418 - 400}{20} = \frac{18}{20} = 0.90$$

Because the injected rat still looks the same as a regular, nontreated rat, the conclusion is that the hormone does not appear to have an effect.

Now, assume that our injected rat weighs $X = 450$ grams. In the distribution of regular rats (see Figure 5.10), this animal would have a z-score of

$$z = \frac{X - \mu}{\sigma} = \frac{450 - 400}{20} = \frac{50}{20} = 2.50$$

In this case, the hormone-injected rat is substantially bigger than most ordinary rats, and it would be reasonable to conclude that the hormone does have an effect on weight.

In the preceding example, we used z-scores to help interpret the results obtained from a sample. Specifically, if the individuals who receive the treatment in a research

study have extreme z-scores compared to those who do not receive the treatment, we can conclude that the treatment does appear to have an effect. The example, however, used an arbitrary definition to determine which z-score values are noticeably different. Although it is reasonable to describe individuals with z-scores near 0 as "highly representative" of the population, and individuals with z-scores beyond ±2.00 as "extreme," you should realize that these z-score boundaries were not determined by any mathematical rule. In the following chapter we introduce *probability*, which gives us a rationale for deciding exactly where to set the boundaries.

LEARNING CHECK

1. For a sample with a mean of $M = 40$ and a standard deviation of $s = 12$, find the z-score corresponding to each of the following X values.

$X = 43$	$X = 58$	$X = 49$
$X = 34$	$X = 28$	$X = 16$

2. For a sample with a mean of $M = 80$ and a standard deviation of $s = 20$, find the X value corresponding to each of the following z-scores.

$z = –1.00$	$z = –0.50$	$z = –0.20$
$z = 1.50$	$z = 0.80$	$z = 1.40$

3. For a sample with a mean of $M = 85$, a score of $X = 80$ corresponds to $z = -0.50$. What is the standard deviation for the sample?

4. For a sample with a standard deviation of $s = 12$, a score of $X = 83$ corresponds to $z = 0.50$. What is the mean for the sample?

5. A sample has a mean of $M = 30$ and a standard deviation of $s = 8$.

a. Would a score of $X = 36$ be considered a central score or an extreme score in the sample?

b. If the standard deviation were $s = 2$, would $X = 36$ be central or extreme?

ANSWERS

1.

$z = 0.25$	$z = 1.50$	$z = 0.75$
$z = -0.50$	$z = -1.00$	$z = -2.00$

2.

$X = 60$	$X = 70$	$X = 76$
$X = 110$	$X = 96$	$X = 108$

3. $s = 10$

4. $M = 77$

5. **a.** $X = 36$ is a central score corresponding to $z = 0.75$.

b. $X = 36$ would be an extreme score corresponding to $z = 3.00$.

SUMMARY

1. Each X value can be transformed into a z-score that specifies the exact location of X within the distribution. The sign of the z-score indicates whether the location is above (positive) or below (negative) the mean. The numerical value of the z-score specifies the number of standard deviations between X and μ.
2. The z-score formula is used to transform X values into z-scores. For a population:

$$z = \frac{X - \mu}{\sigma}$$

   For a sample:

$$z = \frac{X - M}{s}$$

3. To transform z-scores back into X values, it usually is easier to use the z-score definition rather than a formula. However, the z-score formula can be transformed into a new equation. For a population:

$$X = \mu + z\sigma$$

   For a sample: $X = M + zs$
4. When an entire distribution of X values is transformed into z-scores, the result is a distribution of z-scores. The z-score distribution will have the same shape as the distribution of raw scores, and it always will have a mean of 0 and a standard deviation of 1.
5. When comparing raw scores from different distributions, it is necessary to standardize the distributions with a z-score transformation. The distributions will then be comparable because they will have the same parameters ($\mu = 0$, $\sigma = 1$). In practice, it is necessary to transform only those raw scores that are being compared.
6. In certain situations, such as psychological testing, a distribution may be standardized by converting the original X values into z-scores and then converting the z-scores into a new distribution of scores with predetermined values for the mean and the standard deviation.
7. In inferential statistics, z-scores provide an objective method for determining how well a specific score represents its population. A z-score near 0 indicates that the score is close to the population mean and, therefore, is representative. A z-score beyond +2.00 (or –2.00) indicates that the score is extreme and is noticeably different from the other scores in the distribution.

KEY TERMS

raw score (139)
z-score (139)
deviation score (143)
z-score transformation (146)
standardized distribution (147)
standardized score (152)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 5 on the book companion website. The website also includes a workshop entitled z-*Scores* that examines the basic concepts and calculations underlying z-scores.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to **Transform X Values into z-Scores for a Sample.**

Data Entry

1. Enter all of the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Descriptives.**
2. Highlight the column label for the set of scores (VAR0001) in the left box and click the arrow to move it into the **Variable** box.
3. Click the box to **Save standardized values as variables** at the bottom of the **Descriptives** screen.
4. Click **OK.**

SPSS Output

The program produces the usual output display listing the number of scores (N), the maximum and minimum scores, the mean, and the standard deviation. However, if you go back to the Data Editor (use the tool bar at the bottom of the screen), you can see that SPSS has produced a new column showing the z-score corresponding to each of the original X values.

Caution: The SPSS program computes the z-scores using the sample standard deviation instead of the population standard deviation. If your set of scores is intended to be a population, SPSS does not produce the correct z-score values. You can convert the SPSS values into population z-scores by multiplying each z-score value by the square root of $n/(n - 1)$.

FOCUS ON PROBLEM SOLVING

1. When you are converting an X value to a z-score (or vice versa), do not rely entirely on the formula. You can avoid careless mistakes if you use the definition of a z-score (sign and numerical value) to make a preliminary estimate of the answer before you begin computations. For example, a z-score of $z = –0.85$ identifies a score located below the mean by almost 1 standard deviation. When computing the X value for this z-score, be sure that your answer is smaller than the mean, and check that the distance between X and μ is slightly less than the standard deviation.

2. When comparing scores from distributions that have different standard deviations, it is important to be sure that you use the correct value for σ in the z-score formula. Use the σ value for the distribution from which the raw score in question was taken.
3. Remember that a z-score specifies a relative position within the context of a specific distribution. A z-score is a relative value, not an absolute value. For example, a z-score of $z = -2.0$ does not necessarily suggest a very low raw score—it simply means that the raw score is among the lowest within that specific group.

DEMONSTRATION 5.1

TRANSFORMING *X* VALUES INTO *z*-SCORES

A distribution of scores has a mean of $\mu = 60$ with $\sigma = 12$. Find the z-score for $X = 75$.

STEP 1 **Determine the sign of the *z*-score.** First, determine whether X is above or below the mean. This determines the sign of the z-score. For this demonstration, X is larger than (above) μ, so the z-score is positive.

STEP 2 **Convert the distance between *X* and μ into standard deviation units.** For $X = 75$ and $\mu = 60$, the distance between X and μ is 15 points. With $\sigma = 12$ points, this distance corresponds to $\frac{15}{12} = 1.25$ standard deviations.

STEP 3 **Combine the sign from step 1 with the numerical value from step 2.** The score is above the mean (+) by a distance of 1.25 standard deviations. Thus, $z = +1.25$.

STEP 4 **Confirm the answer using the *z*-score formula.** For this example, $X = 75$, $\mu = 60$, and $\sigma = 12$.

$$z = \frac{X - \mu}{\sigma} = \frac{75 - 60}{12} = \frac{+15}{12} = +1.25$$

DEMONSTRATION 5.2

CONVERTING *z*-SCORES TO *X* VALUES

For a population with $\mu = 60$ and $\sigma = 12$, what is the X value corresponding to $z = -0.50$?

STEP 1 **Locate *X* in relation to the mean.** A z-score of –0.50 indicates a location below the mean by half of a standard deviation.

STEP 2 **Convert the distance from standard deviation units to points.** With $\sigma = 12$, half of a standard deviation is 6 points.

STEP 3 **Identify the *X* value.** The value we want is located below the mean by 6 points. The mean is $\mu = 60$, so the score must be $X = 54$.

PROBLEMS

1. What information is provided by the sign (+/–) of a z-score? What information is provided by the numerical value of the z-score?

2. A distribution has a standard deviation of $\sigma = 12$. Find the z-score for each of the following locations in the distribution.
- **a.** Above the mean by 3 points.
- **b.** Above the mean by 12 points.
- **c.** Below the mean by 24 points.
- **d.** Below the mean by 18 points.

3. A distribution has a standard deviation of $\sigma = 6$. Describe the location of each of the following z-scores in terms of position relative to the mean. For example, $z = +1.00$ is a location that is 6 points above the mean.
- **a.** $z = +2.00$
- **b.** $z = +0.50$
- **c.** $z = -2.00$
- **d.** $z = -0.50$

4. For a population with $\mu = 50$ and $\sigma = 8$,
- **a.** Find the z-score for each of the following X values. (*Note:* You should be able to find these values using the definition of a z-score. You should not need to use a formula or do any serious calculations.)

$X = 54$	$X = 62$	$X = 52$
$X = 42$	$X = 48$	$X = 34$

- **b.** Find the score (X value) that corresponds to each of the following z-scores. (Again, you should be able to find these values without any formula or serious calculations.)

$z = 1.00$	$z = 0.75$	$z = 1.50$
$z = -0.50$	$z = -0.25$	$z = -1.50$

5. For a population with $\mu = 40$ and $\sigma = 7$, find the z-score for each of the following X values. (*Note:* You probably will need to use a formula and a calculator to find these values.)

$X = 45$	$X = 51$	$X = 41$
$X = 30$	$X = 25$	$X = 38$

6. For a population with a mean of $\mu = 100$ and a standard deviation of $\sigma = 12$,
- **a.** Find the z-score for each of the following X values.

$X = 106$	$X = 115$	$X = 130$
$X = 91$	$X = 88$	$X = 64$

- **b.** Find the score (X value) that corresponds to each of the following z-scores.

$z = -1.00$	$z = -0.50$	$z = 2.00$
$z = 0.75$	$z = 1.50$	$z = -1.25$

7. A population has a mean of $\mu = 40$ and a standard deviation of $\sigma = 8$.
- **a.** For this population, find the z-score for each of the following X values.

$X = 44$	$X = 50$	$X = 52$
$X = 34$	$X = 28$	$X = 64$

- **b.** For the same population, find the score (X value) that corresponds to each of the following z-scores.

$z = 0.75$	$z = 1.50$	$z = -2.00$
$z = -0.25$	$z = -0.50$	$z = 1.25$

8. A sample has a mean of $M = 40$ and a standard deviation of $s = 6$. Find the z-score for each of the following X values from this sample.

$X = 44$	$X = 42$	$X = 46$
$X = 28$	$X = 50$	$X = 37$

9. A sample has a mean of $M = 80$ and a standard deviation of $s = 10$. For this sample, find the X value corresponding to each of the following z-scores.

$z = 0.80$	$z = 1.20$	$z = 2.00$
$z = -0.40$	$z = -0.60$	$z = -1.80$

10. Find the z-score corresponding to a score of $X = 60$ for each of the following distributions.
- **a.** $\mu = 50$ and $\sigma = 20$
- **b.** $\mu = 50$ and $\sigma = 10$
- **c.** $\mu = 50$ and $\sigma = 5$
- **d.** $\mu = 50$ and $\sigma = 2$

11. Find the X value corresponding to $z = 0.25$ for each of the following distributions.
- **a.** $\mu = 40$ and $\sigma = 4$
- **b.** $\mu = 40$ and $\sigma = 8$
- **c.** $\mu = 40$ and $\sigma = 12$
- **d.** $\mu = 40$ and $\sigma = 20$

12. A score that is 6 points below the mean corresponds to a z-score of $z = -0.50$. What is the population standard deviation?

13. A score that is 12 points above the mean corresponds to a z-score of $z = 3.00$. What is the population standard deviation?

14. For a population with a standard deviation of $\sigma = 8$, a score of $X = 44$ corresponds to $z = -0.50$. What is the population mean?

15. For a sample with a standard deviation of $s = 10$, a score of $X = 65$ corresponds to $z = 1.50$. What is the sample mean?

16. For a sample with a mean of $\mu = 45$, a score of $X = 59$ corresponds to $z = 2.00$. What is the sample standard deviation?

17. For a population with a mean of $\mu = 70$, a score of $X = 62$ corresponds to $z = -2.00$. What is the population standard deviation?

18. In a population of exam scores, a score of $X = 48$ corresponds to $z = +1.00$ and a score of $X = 36$ corresponds to $z = -0.50$. Find the mean and standard deviation for the population. (*Hint:* Sketch the distribution and locate the two scores on your sketch.)

19. In a distribution of scores, $X = 64$ corresponds to $z = 1.00$, and $X = 67$ corresponds to $z = 2.00$. Find the mean and standard deviation for the distribution.

20. For each of the following populations, would a score of $X = 50$ be considered a central score (near the middle of the distribution) or an extreme score (far out in the tail of the distribution)?
- **a.** $\mu = 45$ and $\sigma = 10$
- **b.** $\mu = 45$ and $\sigma = 2$
- **c.** $\mu = 90$ and $\sigma = 20$
- **d.** $\mu = 60$ and $\sigma = 20$

21. A distribution of exam scores has a mean of $\mu = 80$.
- **a.** If your score is $X = 86$, which standard deviation would give you a better grade: $\sigma = 4$ $\sigma = 8$?
- **b.** If your score is $X = 74$, which standard deviation would give you a better grade: $\sigma = 4$ or $\sigma = 8$?

22. For each of the following, identify the exam score that should lead to the better grade. In each case, explain your answer.
- **a.** A score of $X = 56$, on an exam with $\mu = 50$ and $\sigma = 4$; or a score of $X = 60$ on an exam with $\mu = 50$ and $\sigma = 20$.
- **b.** A score of $X = 40$, on an exam with $\mu = 45$ and $\sigma = 2$; or a score of $X = 60$ on an exam with $\mu = 70$ and $\sigma = 20$.
- **c.** A score of $X = 62$, on an exam with $\mu = 50$ and $\sigma = 8$; or a score of $X = 23$ on an exam with $\mu = 20$ and $\sigma = 2$.

23. A distribution with a mean of $\mu = 62$ and a standard deviation of $\sigma = 8$ is transformed into a standardized distribution with $\mu = 100$ and $\sigma = 20$. Find the new, standardized score for each of the following values from the original population.
- **a.** $X = 60$
- **b.** $X = 54$
- **c.** $X = 72$
- **d.** $X = 66$

24. A distribution with a mean of $\mu = 56$ and a standard deviation of $\sigma = 20$ is transformed into a standardized distribution with $\mu = 50$ and $\sigma = 10$. Find the new, standardized score for each of the following values from the original population.
- **a.** $X = 46$
- **b.** $X = 76$
- **c.** $X = 40$
- **d.** $X = 80$

25. A population consists of the following $N = 5$ scores: 0, 6, 4, 3, and 12.
- **a.** Compute μ and σ for the population.
- **b.** Find the z-score for each score in the population.
- **c.** Transform the original population into a new population of $N = 5$ scores with a mean of $\mu = 100$ and a standard deviation of $\sigma = 20$.

26. A sample consists of the following $n = 6$ scores: 2, 7, 4, 6, 4, and 7.
- **a.** Compute the mean and standard deviation for the sample.
- **b.** Find the z-score for each score in the sample.
- **c.** Transform the original sample into a new sample with a mean of $M = 50$ and $s = 10$.

CHAPTER

6

Probability

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Proportions (math review, Appendix A)
  - Fractions
  - Decimals
  - Percentages
- Basic algebra (math review, Appendix A)
  - *z*-Scores (Chapter 5)

Preview

Background: If you open a dictionary and randomly pick one word, which are you more likely to select:

1. A word beginning with the letter K?
2. A word with a *K* as its third letter?

If you think about this question and answer honestly, you probably will decide that words beginning with a *K* are more probable.

A similar question was asked a group of participants in an experiment reported by Tversky and Kahneman (1973). Their participants estimated that words beginning with *K* are twice as likely as words with a *K* as the third letter. In truth, the relationship is just the opposite. There are more than twice as many words with a *K* in the third position as there are words beginning with a *K*. How can people be so wrong? Do they completely misunderstand probability?

When you were deciding which type of *K* words are more likely, you probably searched your memory and tried to estimate which words are more common. How many words can you think of that start with the letter *K*? How many words can you think of that have a *K* as the third letter? Because you have had years of practice alphabetizing words according to their first letter, you should find it much easier to search your memory for words beginning with a *K* than to search for words with a *K* in the third position. Consequently, you are likely to conclude that first-letter *K* words are more common.

If you had searched for words in a dictionary (instead of those in your memory), you would have found more third-letter *K* words, and you would have concluded (correctly) that these words are more common.

The Problem: If you open a dictionary and randomly pick one word, it is impossible to predict exactly which word you will get. In the same way, when researchers recruit people to participate in research studies, it is impossible to predict exactly which individuals will be obtained.

The Solution: Although it is impossible to predict exactly which word will be picked from a dictionary, or which person will participate in a research study, you can use *probability* to demonstrate that some outcomes are more likely than others. For example, it is more likely that you will pick a third-letter *K* word than a first-letter *K* word. Similarly, it is more likely that you will obtain a person with an IQ around 100 than a person with an IQ around 150.

6.1 INTRODUCTION TO PROBABILITY

In Chapter 1, we introduced the idea that research studies begin with a general question about an entire population, but the actual research is conducted using a sample. In this situation, the role of inferential statistics is to use the sample data as the basis for answering questions about the population. To accomplish this goal, inferential procedures are typically built around the concept of probability. Specifically, the relationships between samples and populations are usually defined in terms of probability.

Suppose, for example, that you are selecting a single marble from a jar that contains 50 black and 50 white marbles. (In this example, the jar of marbles is the *population* and the single marble to be selected is the *sample*.) Although you cannot guarantee the exact outcome of your sample, it is possible to talk about the potential outcomes in terms of probabilities. In this case, you have a 50-50 chance of getting either color. Now consider another jar (population) that has 90 black and only 10 white marbles. Again, you cannot predict the exact outcome of a sample, but now you know that the sample probably will be a black marble. By knowing the makeup of a population, we can determine the probability of obtaining specific samples. In this way, probability gives us a connection between populations and samples, and this connection is the foundation for the inferential statistics that are presented in the chapters that follow.

You may have noticed that the preceding examples begin with a population and then use probability to describe the samples that could be obtained. This is exactly

backward from what we want to do with inferential statistics. Remember that the goal of inferential statistics is to begin with a sample and then answer a general question about the population. We reach this goal in a two-stage process. In the first stage, we develop probability as a bridge from populations to samples. This stage involves identifying the types of samples that probably would be obtained from a specific population. Once this bridge is established, we simply reverse the probability rules to allow us to move from samples to populations (Figure 6.1). The process of reversing the probability relationship can be demonstrated by considering again the two jars of marbles we looked at earlier. (Jar 1 has 50 black and 50 white marbles; jar 2 has 90 black and only 10 white marbles.) This time, suppose you are blindfolded when the sample is selected, so you do not know which jar is being used. Your task is to look at the sample that you obtain and then decide which jar is most likely. If you select a sample of $n = 4$ marbles and all are black, which jar would you choose? It should be clear that it would be relatively unlikely (low probability) to obtain this sample from jar 1; in four draws, you almost certainly would get at least 1 white marble. On the other hand, this sample would have a high probability of coming from jar 2, where nearly all of the marbles are black. Your decision, therefore, is that the sample probably came from jar 2. Note that you now are using the sample to make an inference about the population.

DEFINING PROBABILITY

Probability is a huge topic that extends far beyond the limits of introductory statistics, and we do not attempt to examine it all here. Instead, we concentrate on the few concepts and definitions that are needed for an introduction to inferential statistics. We begin with a relatively simple definition of probability.

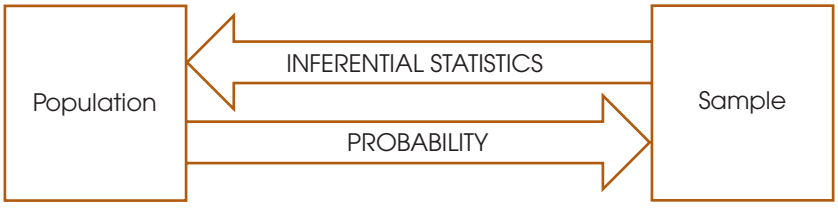
DEFINITION

For a situation in which several different outcomes are possible, the **probability** for any specific outcome is defined as a fraction or a proportion of all the possible outcomes. If the possible outcomes are identified as A, B, C, D, and so on, then

$$\text{probability of } A = \frac{\text{number of outcomes classified as } A}{\text{total number of possible outcomes}}$$

FIGURE 6.1

The role of probability in inferential statistics. Probability is used to predict what kind of samples are likely to be obtained from a population. Thus, probability establishes a connection between samples and populations. Inferential statistics rely on this connection when they use sample data as the basis for making conclusions about populations.



For example, if you are selecting a card from a complete deck, there are 52 possible outcomes. The probability of selecting the king of hearts is $p = \frac{1}{52}$. The probability of selecting an ace is $p = \frac{4}{52}$ because there are 4 aces in the deck.

To simplify the discussion of probability, we use a notation system that eliminates a lot of the words. The probability of a specific outcome is expressed with a p (for probability) followed by the specific outcome in parentheses. For example, the probability of selecting a king from a deck of cards is written as p(king). The probability of obtaining heads for a coin toss is written as p(heads).

Note that probability is defined as a proportion, or a part of the whole. This definition makes it possible to restate any probability problem as a proportion problem. For example, the probability problem "What is the probability of selecting a king from a deck of cards?" can be restated as "What proportion of the whole deck consists of kings?" In each case, the answer is $\frac{4}{52}$, or "4 out of 52." This translation from probability to proportion may seem trivial now, but it is a great aid when the probability problems become more complex. In most situations, we are concerned with the probability of obtaining a particular sample from a population. The terminology of *sample* and *population* do change the basic definition of probability. For example, the whole deck of cards can be considered as a population, and the single card we select is the sample.

Probability values The definition we are using identifies probability as a fraction or a proportion. If you work directly from this definition, the probability values you obtain are expressed as fractions. For example, if you are selecting a card at random,

$$p(\text{spade}) = \frac{13}{52} = \frac{1}{4}$$

Of if you are tossing a coin,

$$p(\text{heads}) = \frac{1}{2}$$

If you are unsure how to convert from fractions to decimals or percentages, you should review the section on proportions in the math review, Appendix A.

You should be aware that these fractions can be expressed equally well as either decimals or percentages:

$$p = \frac{1}{4} = 0.25 = 25\%$$

$$p = \frac{1}{2} = 0.50 = 50\%$$

By convention, probability values most often are expressed as decimal values. But you should realize that any of these three forms is acceptable.

You also should note that all of the possible probability values are contained in a limited range. At one extreme, when an event never occurs, the probability is zero, or 0% (Box 6.1). At the other extreme, when an event always occurs, the probability is 1, or 100%. Thus, all probability values are contained in a range from 0 to 1. For example, suppose that you have a jar containing 10 white marbles. The probability of randomly selecting a black marble is

$$p(\text{black}) = \frac{0}{10} = 0$$

The probability of selecting a white marble is

$$p(\text{white}) = \frac{10}{10} = 1$$

RANDOM SAMPLING

For the preceding definition of probability to be accurate, it is necessary that the outcomes be obtained by a process called *random sampling*.

DEFINITION

A **random sample** requires that each individual in the population has an *equal chance* of being selected.

A second requirement, necessary for many statistical formulas, states that if more than one individual is being selected, the probabilities must *stay constant* from one selection to the next. Adding this second requirement produces what is called *independent random sampling*. The term *independent* refers to the fact that the probability of selecting any particular individual is independent of those individuals who have already been selected for the sample. For example, the probability that you will be selected is constant and does not change even when other individuals are selected before you are.

DEFINITION

An **independent random sample** requires that each individual has an equal chance of being selected and that the probability of being selected stays constant from one selection to the next if more than one individual is selected.

Because independent random sample is a required component for most statistical applications, we always assume that this is the sampling method being used. To simplify discussion, we typically omit the word "independent" and simply refer to this sampling technique as *random sampling*. However, you should always assume that both requirements (equal chance and constant probability) are part of the process.

Each of the two requirements for random sampling has some interesting consequences. The first assures that there is no bias in the selection process. For a population with N individuals, each individual must have the same probability, $p = \frac{1}{N}$, of being selected. This means, for example, that you would not get a random sample of people in your city by selecting names from a yacht-club membership list. Similarly, you would not get a random sample of college students by selecting individuals from your psychology classes. You also should note that the first requirement of random sampling prohibits you from applying the definition of probability to situations in which the possible outcomes are not equally likely. Consider, for example, the question of whether you will win a million dollars in the lottery tomorrow. There are only two possible alternatives.

1. You will win.
2. You will not win.

According to our simple definition, the probability of winning would be one out of two, or $p = \frac{1}{2}$. However, the two alternatives are not equally likely, so the simple definition of probability does not apply.

The second requirement also is more interesting than may be apparent at first glance. Consider, for example, the selection of $n = 2$ cards from a complete deck. For the first draw, the probability of obtaining the jack of diamonds is

$$p(\text{jack of diamonds}) = \frac{1}{52}$$

After selecting one card for the sample, you are ready to draw the second card. What is the probability of obtaining the jack of diamonds this time? Assuming that you still are holding the first card, there are two possibilities:

$$p(\text{jack of diamonds}) = \frac{1}{51} \text{ if the first card was not the jack of diamonds}$$

or

$$p(\text{jack of diamonds}) = 0 \text{ if the first card was the jack of diamonds}$$

In either case, the probability is different from its value for the first draw. This contradicts the requirement for random sampling, which says that the probability must stay constant. To keep the probabilities from changing from one selection to the next, it is necessary to return each individual to the population before you make the next selection. This process is called *sampling with replacement*. The second requirement for random samples (constant probability) demands that you sample with replacement.

(*Note:* We are using a definition of random sampling that requires equal chance of selection and constant probabilities. This kind of sampling is also known as independent random sampling, and often is called *random sampling with replacement*. Many of the statistics we encounter later are founded on this kind of sampling. However, you should realize that other definitions exist for the concept of random sampling. In particular, it is very common to define random sampling without the requirement of constant probabilities—that is, *random sampling without replacement*. In addition, there are many different sampling techniques that are used when researchers are selecting individuals to participate in research studies.)

PROBABILITY AND FREQUENCY DISTRIBUTIONS

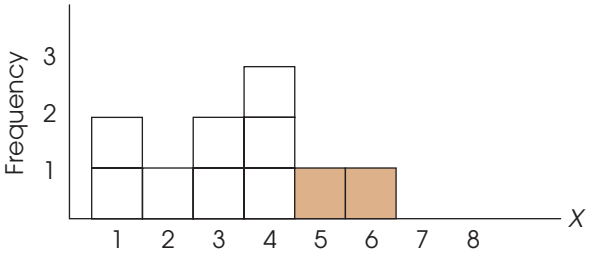
The situations in which we are concerned with probability usually involve a population of scores that can be displayed in a frequency distribution graph. If you think of the graph as representing the entire population, then different proportions of the graph represent different proportions of the population. Because probabilities and proportions are equivalent, a particular proportion of the graph corresponds to a particular probability in the population. Thus, whenever a population is presented in a frequency distribution graph, it is possible to represent probabilities as proportions of the graph. The relationship between graphs and probabilities is demonstrated in the following example.

EXAMPLE 6.1

We use a very simple population that contains only $N = 10$ scores with values 1, 1, 2, 3, 3, 4, 4, 4, 5, 6. This population is shown in the frequency distribution graph in Figure 6.2. If you are taking a random sample of $n = 1$ score from this population,

FIGURE 6.2

A frequency distribution histogram for a population that consists of $N = 10$ scores. The shaded part of the figure indicates the portion of the whole population that corresponds to scores greater than $X = 4$. The shaded portion is two-tenths ($p = \frac{2}{10}$) of the whole distribution.



what is the probability of obtaining an individual with a score greater than 4? In probability notation,

$$p(X > 4) = ?$$

Using the definition of probability, there are 2 scores that meet this criterion out of the total group of $N = 10$ scores, so the answer would be $p = \frac{2}{10}$. This answer can be obtained directly from the frequency distribution graph if you recall that probability and proportion measure the same thing. Looking at the graph (see Figure 6.2), what proportion of the population consists of scores greater than 4? The answer is the shaded part of the distribution—that is, 2 squares out of the total of 10 squares in the distribution. Notice that we now are defining probability as a proportion of *area* in the frequency distribution graph. This provides a very concrete and graphic way of representing probability.

Using the same population once again, what is the probability of selecting an individual with a score less than 5? In symbols,

$$p(X < 5) = ?$$

Going directly to the distribution in Figure 6.2, we now want to know what part of the graph is not shaded. The unshaded portion consists of 8 out of the 10 blocks (eight-tenths of the area of the graph), so the answer is $p = \frac{8}{10}$.

LEARNING CHECK

1. A survey of the students in a psychology class revealed that there were 19 females and 8 males. Of the 19 females, only 4 had no brothers or sisters, and 3 of the males were also the only child in the household. If a student is randomly selected from this class,

- **a.** What is the probability of obtaining a male?
- **b.** What is the probability of selecting a student who has at least one brother or sister?
- **c.** What is the probability of selecting a female who has no siblings?

2. A jar contains 10 red marbles and 30 blue marbles.

- **a.** If you randomly select 1 marble from the jar, what is the probability of obtaining a red marble?
- **b.** If you take a *random sample* of $n = 3$ marbles from the jar and the first two marbles are both blue, what is the probability that the third marble will be red?

3. Suppose that you are going to select a random sample of $n = 1$ score from the distribution in Figure 6.2. Find the following probabilities:

- **a.** $p(X > 2)$
- **b.** $p(X > 5)$
- **c.** $p(X < 3)$

ANSWERS

1. a. $p = \frac{8}{27}$

b. $p = \frac{20}{27}$

c. $p = \frac{4}{27}$

2. a. $p = \frac{10}{40} = 0.25$

b. $p = \frac{10}{40} = 0.25$. Remember that random sampling requires sampling with replacement.

3. a. $p = \frac{7}{10} = 0.70$

b. $p = \frac{1}{10} = 0.10$

c. $p = \frac{3}{10} = 0.30$

6.2 PROBABILITY AND THE NORMAL DISTRIBUTION

The normal distribution was first introduced in Chapter 2 as an example of a commonly occurring shape for population distributions. An example of a normal distribution is shown in Figure 6.3.

Note that the normal distribution is symmetrical, with the highest frequency in the middle and frequencies tapering off as you move toward either extreme. Although the exact shape for the normal distribution is defined by an equation (see Figure 6.3), the normal shape can also be described by the proportions of area contained in each section of the distribution. Statisticians often identify sections of a normal distribution by using z-scores. Figure 6.4 shows a normal distribution with several sections marked in z-score units. You should recall that z-scores measure positions in a distribution in terms of standard deviations from the mean. (Thus, $z = +1$ is 1 standard deviation above the mean, $z = +2$ is 2 standard deviations above the mean, and so on.) The graph shows the percentage of scores that fall in each of these sections. For example, the section between the mean ($z = 0$) and the point that is 1 standard deviation above the mean ($z = 1$) contains 34.13% of the scores. Similarly, 13.59% of the scores are located in the section between

FIGURE 6.3

The normal distribution. The exact shape of the normal distribution is specified by an equation relating each X value (score) with each Y value (frequency). The equation is

$$Y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$$

(π and e are mathematical constants). In simpler terms, the normal distribution is symmetrical with a single mode in the middle. The frequency tapers off as you move farther from the middle in either direction.

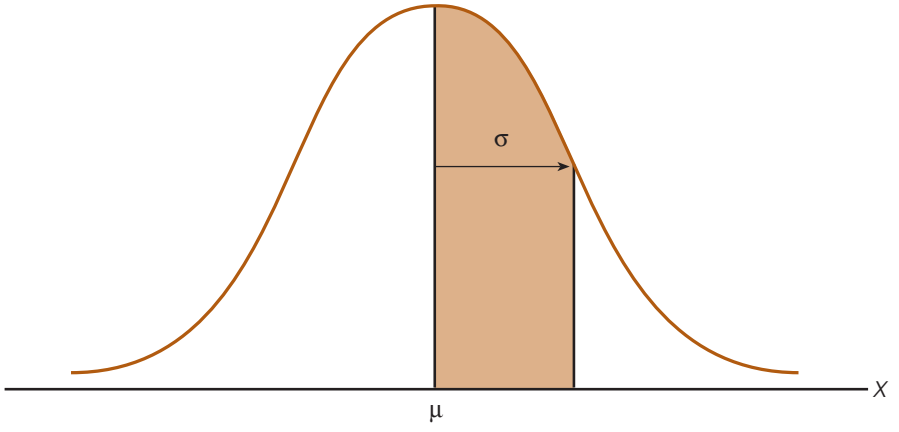
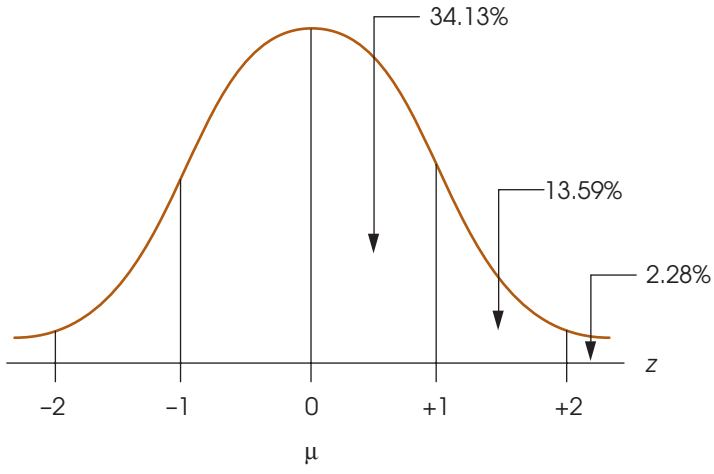


FIGURE 6.4
The normal distribution following a *z*-score transformation.



1 and 2 standard deviations above the mean. In this way it is possible to define a normal distribution in terms of its proportions; that is, a distribution is normal if and only if it has all the right proportions.

There are two additional points to be made about the distribution shown in Figure 6.4. First, you should realize that the sections on the left side of the distribution have exactly the same areas as the corresponding sections on the right side because the normal distribution is symmetrical. Second, because the locations in the distribution are identified by *z*-scores, the percentages shown in the figure apply to *any normal distribution* regardless of the values for the mean and the standard deviation. Remember: When any distribution is transformed into *z*-scores, the mean becomes zero and the standard deviation becomes one.

Because the normal distribution is a good model for many naturally occurring distributions and because this shape is guaranteed in some circumstances (as we see in Chapter 7), we devote considerable attention to this particular distribution. The process of answering probability questions about a normal distribution is introduced in the following example.

EXAMPLE 6.2

The population distribution of SAT scores is normal with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. Given this information about the population and the known proportions for a normal distribution (see Figure 6.4), we can determine the probabilities associated with specific samples. For example, what is the probability of randomly selecting an individual from this population who has an SAT score greater than 700?

Restating this question in probability notation, we get

$$p(X > 700) = ?$$

We follow a step-by-step process to find the answer to this question.

1. First, the probability question is translated into a proportion question: Out of all possible SAT scores, what proportion is greater than 700?

2. The set of "all possible SAT scores" is simply the population distribution. This population is shown in Figure 6.5. The mean is $\mu = 500$, so the score $X = 700$ is to the right of the mean. Because we are interested in all scores greater than 700, we shade in the area to the right of 700. This area represents the proportion we are trying to determine.
3. Identify the exact position of $X = 700$ by computing a z-score. For this example,

$$z = \frac{X - \mu}{\sigma} = \frac{700 - 500}{100} = \frac{200}{100} = 2.00$$

That is, an SAT score of $X = 700$ is exactly 2 standard deviations above the mean and corresponds to a z-score of $z = +2.00$. We have also located this z-score in Figure 6.5.

4. The proportion we are trying to determine may now be expressed in terms of its z-score:

$$p(z > 2.00) = ?$$

According to the proportions shown in Figure 6.4, all normal distributions, regardless of the values for μ and σ, have 2.28% of the scores in the tail beyond $z = +2.00$. Thus, for the population of SAT scores,

$$p(X > 700) = p(z > +2.00) = 2.28\%$$

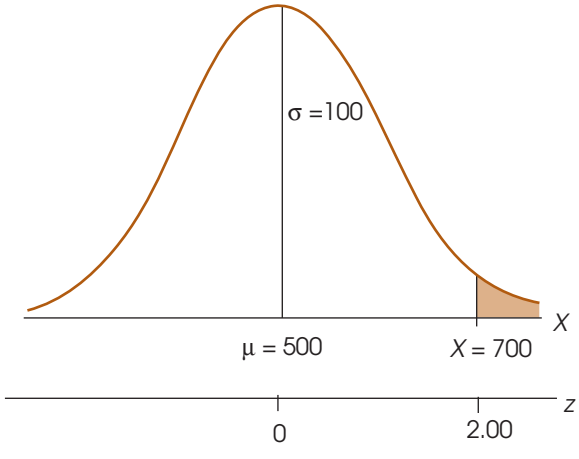
THE UNIT NORMAL TABLE

Before we attempt any more probability questions, we must introduce a more useful tool than the graph of the normal distribution shown in Figure 6.4. The graph shows proportions for only a few selected z-score values. A more complete listing of z-scores and proportions is provided in the *unit normal table*. This table lists proportions of the normal distribution for a full range of possible z-score values.

The complete unit normal table is provided in Appendix B Table B.1, and part of the table is reproduced in Figure 6.6. Notice that the table is structured in a four-column

FIGURE 6.5

The distribution of SAT scores described in Example 6.2.



(A) *z*	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and *z*
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331

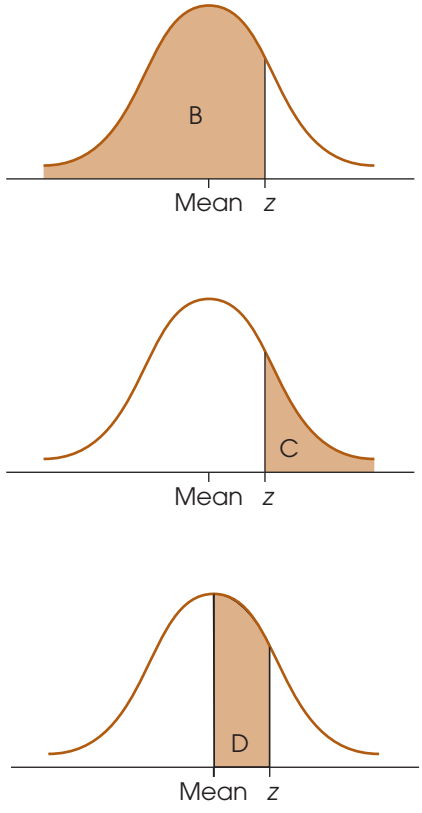


FIGURE 6.6

A portion of the unit normal table. This table lists proportions of the normal distribution corresponding to each *z*-score value. Column A of the table lists *z*-scores. Column B lists the proportion in the body of the normal distribution up to the *z*-score value. Column C lists the proportion of the normal distribution that is located in the tail of the distribution beyond the *z*-score value. Column D lists the proportion between the mean and the *z*-score value.

format. The first column (A) lists *z*-score values corresponding to different positions in a normal distribution. If you imagine a vertical line drawn through a normal distribution, then the exact location of the line can be described by one of the *z*-score values listed in column A. You should also realize that a vertical line separates the distribution into two sections: a larger section called the *body* and a smaller section called the *tail*. Columns B and C in the table identify the proportion of the distribution in each of the two sections. Column B presents the proportion in the body (the larger portion), and column C presents the proportion in the tail. Finally, we have added a fourth column, column D, that identifies the proportion of the distribution that is located *between* the mean and the *z*-score.

We use the distribution in Figure 6.7(a) to help introduce the unit normal table. The figure shows a normal distribution with a vertical line drawn at $z = +0.25$. Using the portion of the table shown in Figure 6.6, find the row in the table that contains $z = 0.25$ in column A. Reading across the row, you should find that the line drawn $z = +0.25$

separates the distribution into two sections with the larger section containing 0.5987 (59.87%) of the distribution and the smaller section containing 0.4013 (40.13%) of the distribution. Also, there is exactly 0.0987 (9.87%) of the distribution between the mean and $z = +0.25$.

To make full use of the unit normal table, there are a few facts to keep in mind:

1. The *body* always corresponds to the larger part of the distribution whether it is on the right-hand side or the left-hand side. Similarly, the *tail* is always the smaller section whether it is on the right or the left.
2. Because the normal distribution is symmetrical, the proportions on the right-hand side are exactly the same as the corresponding proportions on the left-hand side. Earlier, for example, we used the unit normal table to obtain proportions for $z = +0.25$. Figure 6.7(b) shows the same proportions for $z = -0.25$. For a negative z-score, however, notice that the tail of the distribution is on the left side and the body is on the right. For a positive z-score [Figure 6.7(a)], the positions are reversed. However, the proportions in each section are exactly the same, with 0.55987 in the body and 0.4013 in the tail. Once again, the table does not list negative z-score values. To find proportions for negative z-scores, you must look up the corresponding proportions for the positive value of z.
3. Although the z-score values change signs (+ and –) from one side to the other, the proportions are always positive. Thus, column C in the table always lists the proportion in the tail whether it is the right-hand tail or the left-hand tail.

PROBABILITIES, PROPORTIONS, AND *Z*-SCORES

The unit normal table lists relationships between z-score locations and proportions in a normal distribution. For any z-score location, you can use the table to look up the corresponding proportions. Similarly, if you know the proportions, you can use the table to find the specific z-score location. Because we have defined probability as equivalent to proportion, you can also use the unit normal table to look up probabilities for normal distributions. The following examples demonstrate a variety of different ways that the unit normal table can be used.

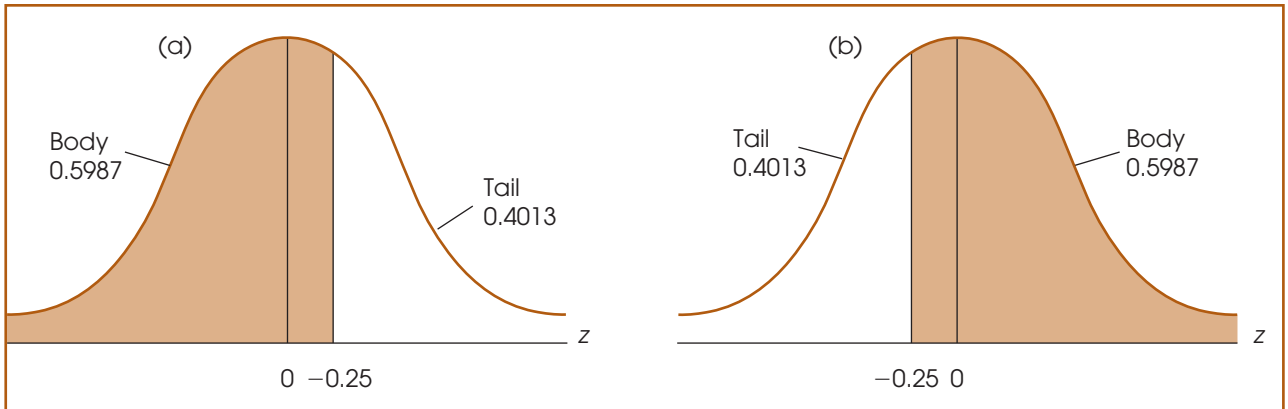


FIGURE 6.7

Proportions of a normal distribution corresponding to $z = +0.25$ (a) and –0.25 (b).

Finding proportions or probabilities for specific *z*-score values For each of the following examples, we begin with a specific *z*-score value and then use the unit normal table to find probabilities or proportions associated with the *z*-score.

EXAMPLE 6.3A

What proportion of the normal distribution corresponds to *z*-score values greater than $z = 1.00$? First, you should sketch the distribution and shade in the area you are trying to determine. This is shown in Figure 6.8(a). In this case, the shaded portion is the tail of the distribution beyond $z = 1.00$. To find this shaded area, you simply look for $z = 1.00$ in column A to find the appropriate row in the unit normal table. Then scan across the row to column C (tail) to find the proportion. Using the table in Appendix B, you should find that the answer is 0.1587.

You also should notice that this same problem could have been phrased as a probability question. Specifically, we could have asked, "For a normal distribution, what is the probability of selecting a *z*-score value greater than $z = +1.00$?" Again, the answer is $p(z > 1.00) = 0.1587$ (or 15.87%).

EXAMPLE 6.3B

For a normal distribution, what is the probability of selecting a *z*-score less than $z = 1.50$? In symbols, $p(z < 1.50) = ?$ Our goal is to determine what proportion of the normal distribution corresponds to *z*-scores less than 1.50. A normal distribution is shown in Figure 6.8(b) and $z = 1.50$ is marked in the distribution. Notice that we have shaded all the values to the left of (less than) $z = 1.50$. This is the portion we are trying to find. Clearly the shaded portion is more than 50%, so it corresponds to the body of the distribution. Therefore, find $z = 1.50$ in column A of the unit normal table and read across the row to obtain the proportion from column B. The answer is $p(z < 1.50) = 0.9332$ (or 93.32%).

EXAMPLE 6.3C

Moving to the left on the *X*-axis results in smaller *X* values and smaller *z*-scores. Thus, a *z*-score of –3.00 reflects a smaller value than a *z*-score of –1.

Many problems require that you find proportions for negative *z*-scores. For example, what proportion of the normal distribution is contained in the tail beyond $z = –0.50$? That is, $p(z < –0.50)$. This portion has been shaded in Figure 6.8(c). To answer questions with negative *z*-scores, simply remember that the normal distribution is symmetrical with a *z*-score of zero at the mean, positive values to the right, and negative values to the left. The proportion in the left tail beyond $z = –0.50$ is identical to the proportion

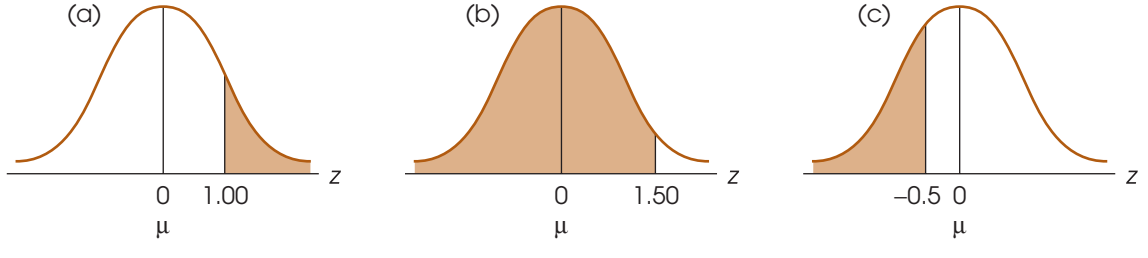


FIGURE 6.8
The distribution for Examples 6.3A to 6.3C.

in the right tail beyond $z = +0.50$. To find this proportion, look up $z = 0.50$ in column A, and read across the row to find the proportion in column C (tail). You should get an answer of 0.3085 (30.85%).

Finding the *z*-score location that corresponds to specific proportions The preceding examples all involved using a *z*-score value in column A to look up proportions in column B or C. You should realize, however, that the table also allows you to begin with a known proportion and then look up the corresponding *z*-score. The following examples demonstrate this process.

EXAMPLE 6.4A

For a normal distribution, what *z*-score separates the top 10% from the remainder of the distribution? To answer this question, we have sketched a normal distribution [Figure 6.9(a)] and drawn a vertical line that separates the highest 10% (approximately) from the rest. The problem is to locate the exact position of this line. For this distribution, we know that the tail contains 0.1000 (10%) and the body contains 0.9000 (90%). To find the *z*-score value, you simply locate the row in the unit normal table that has 0.1000 in column C or 0.9000 in column B. For example, you can scan down the values in column C (tail) until you find a proportion of 0.1000. Note that you probably will not find the exact proportion, but you can use the closest value listed in the table. For this example, a proportion of 0.1000 is not listed in column C but you can use 0.1003, which is listed. Once you have found the correct proportion in the table, simply read across the row to find the corresponding z-score value in column A.

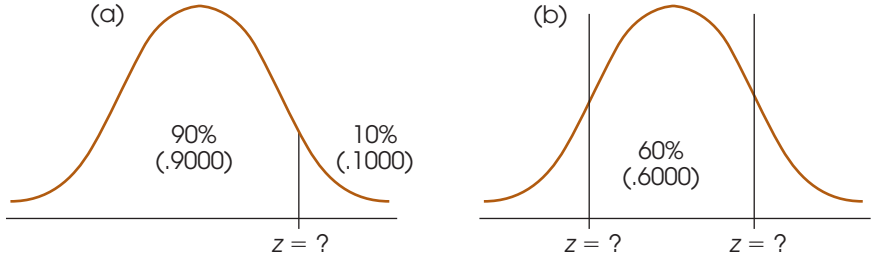
For this example, the *z*-score that separates the extreme 10% in the tail is $z = 1.28$. At this point you must be careful because the table does not differentiate between the right-hand tail and the left-hand tail of the distribution. Specifically, the final answer could be either $z = +1.28$, which separates 10% in the right-hand tail, or $z = -1.28$, which separates 10% in the left-hand tail. For this problem we want the right-hand tail (the highest 10%), so the *z*-score value is $z = +1.28$.

EXAMPLE 6.4B

For a normal distribution, what *z*-score values form the boundaries that separate the middle 60% of the distribution from the rest of the scores?

Again, we have sketched a normal distribution [Figure 6.9(b)] and drawn vertical lines so that roughly 60% of the distribution in the central section, with the remainder

FIGURE 6.9
The distribution for Examples 6.4A and 6.4B.



split equally between the two tails. The problem is to find the z-score values that define the exact locations for the lines. To find the z-score values, we begin with the known proportions: 0.6000 in the center and 0.4000 divided equally between the two tails. Although these proportions can be used in several different ways, this example provides an opportunity to demonstrate how column D in the table can be used to solve problems. For this problem, the 0.6000 in the center can be divided in half with exactly 0.3000 to the right of the mean and exactly 0.3000 to the left. Each of these sections corresponds to the proportion listed in column D. Begin by scanning down column D, looking for a value of 0.3000. Again, this exact proportion is not in the table, but the closest value is 0.2995. Reading across the row to column A, you should find a z-score value of $z = 0.84$. Looking again at the sketch [Figure 6.9(b)], the right-hand line is located at $z = +0.84$ and the left-hand line is located at $z = -0.84$.

You may have noticed that we have sketched distributions for each of the preceding problems. As a general rule, you should always sketch a distribution, locate the mean with a vertical line, and shade in the portion that you are trying to determine. Look at your sketch. It will help you to determine which columns to use in the unit normal table. If you make a habit of drawing sketches, you will avoid careless errors when using the table.

LEARNING CHECK

1. Find the proportion of a normal distribution that corresponds to each of the following sections:

a. $z < 0.25$

b. $z > 0.80$

c. $z < -1.50$

d. $z > -0.75$

2. For a normal distribution, find the z-score location that divides the distribution as follows:

a. Separate the top 20% from the rest.

b. Separate the top 60% from the rest.

c. Separate the middle 70% from the rest.

3. The tail will be on the right-hand side of a normal distribution for any positive z-score. (True or false?)

ANSWERS

1. **a.** $p = 0.5987$

b. $p = 0.2119$

c. $p = 0.0668$

d. $p = 0.7734$

2. **a.** $z = 0.84$

b. $z = -0.25$

c. $z = -1.04$ and $+1.04$

3. True

6.3 PROBABILITIES AND PROPORTIONS FOR SCORES FROM A NORMAL DISTRIBUTION

In the preceding section, we used the unit normal table to find probabilities and proportions corresponding to specific z-score values. In most situations, however, it is necessary to find probabilities for specific X values. Consider the following example:

> It is known that IQ scores form a normal distribution with $\mu = 100$ and $\sigma = 15$. Given this information, what is the probability of randomly selecting an individual with an IQ score less than 120?

This problem is asking for a specific probability or proportion of a normal distribution. However, before we can look up the answer in the unit normal table, we must first transform the IQ scores (X values) into z-scores. Thus, to solve this new kind of probability problem, we must add one new step to the process. Specifically, to answer probability questions about scores (X values) from a normal distribution, you must use the following two-step procedure:

Caution: The unit normal table can be used only with normal-shaped distributions. If a distribution is not normal, transforming to z-scores does not make it normal.

1. Transform the X values into z-scores.
2. Use the unit normal table to look up the proportions corresponding to the z-score values.

This process is demonstrated in the following examples. Once again, we suggest that you sketch the distribution and shade the portion you are trying to find to avoid careless mistakes.

EXAMPLE 6.5

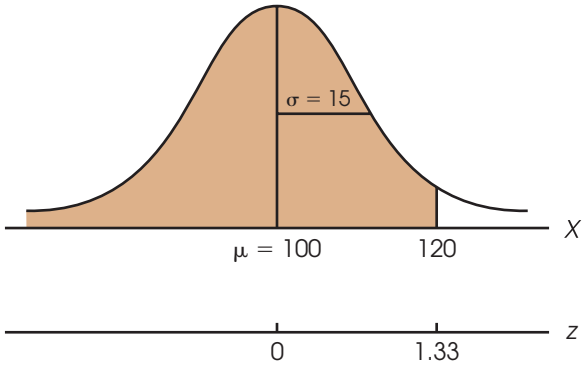
We now answer the probability question about IQ scores that we presented earlier. Specifically, what is the probability of randomly selecting an individual with an IQ score less than 120? Restated in terms of proportions, we want to find the proportion of the IQ distribution that corresponds to scores less than 120. The distribution is drawn in Figure 6.10, and the portion we want has been shaded.

The first step is to change the X values into z-scores. In particular, the score of $X = 120$ is changed to

$$z = \frac{X - \mu}{\sigma} = \frac{120 - 100}{15} = \frac{20}{15} = 1.33$$

FIGURE 6.10

The distribution of IQ scores. The problem is to find the probability or proportion of the distribution corresponding to scores less than 120.



Thus, an IQ score of $X = 120$ corresponds to a z-score of $z = 1.33$, and IQ scores less than 120 correspond to z-scores less than 1.33.

Next, look up the z-score value in the unit normal table. Because we want the proportion of the distribution in the body to the left of $X = 120$ (see Figure 6.10), the answer is in column B. Consulting the table, we see that a z-score of 1.33 corresponds to a proportion of 0.9082. The probability of randomly selecting an individual with an IQ less than 120 is $p = 0.9082$. In symbols,

$$p(X < 120) = p(z < 1.33) = 0.9082 \text{ (or } 90.82\%)$$

Finally, notice that we phrased this question in terms of a *probability*. Specifically, we asked, "What is the probability of selecting an individual with an IQ less than 120?" However, the same question can be phrased in terms of a *proportion:* "What proportion of all of the individuals in the population have IQ scores less than 120?" Both versions ask exactly the same question and produce exactly the same answer. A third alternative for presenting the same question is introduced in Box 6.1.

Finding proportions/probabilities located between two scores The next example demonstrates the process of finding the probability of selecting a score that is located *between* two specific values. Although these problems can be solved using the proportions of columns B and C (body and tail), they are often easier to solve with the proportions listed in column D.

EXAMPLE 6.6 The highway department conducted a study measuring driving speeds on a local section of interstate highway. They found an average speed of $\mu = 58$ miles per hour with a standard deviation of $\sigma = 10$. The distribution was approximately normal.

BOX 6.1 PROBABILITIES, PROPORTIONS, AND PERCENTILE RANKS

Thus far we have discussed parts of distributions in terms of proportions and probabilities. However, there is another set of terminology that deals with many of the same concepts. Specifically, in Chapter 2 we defined the *percentile rank* for a specific score as the percentage of the individuals in the distribution who have scores that are less than or equal to the specific score. For example, if 70% of the individuals have scores of $X = 45$ or lower, then $X = 45$ has a percentile rank of 70%. When a score is referred to by its percentile rank, the score is called a *percentile.* For example, a score with a percentile rank of 70% is called the 70th percentile.

Using this terminology, it is possible to rephrase some of the probability problems that we have been working. In Example 6.5, the problem is presented as "What is the probability of randomly selecting an individual with an IQ of less than 120?" Exactly the same question could be phrased as "What is the percentile rank for an IQ score of 120?" In each case, we are drawing a line at $X = 120$ and looking for the proportion of the distribution on the left-hand side of the line. Similarly, Example 6.8 asks "How much time do you have to spend commuting each day to be in the highest 10% nationwide?" Because this score separates the top 10% from the bottom 90%, the same question could be rephrased as "What is the 90th percentile for the distribution of commuting times?"

Given this information, what proportion of the cars are traveling between 55 and 65 miles per hour? Using probability notation, we can express the problem as

$$p(55 < X < 65) = ?$$

The distribution of driving speeds is shown in Figure 6.11 with the appropriate area shaded. The first step is to determine the z-score corresponding to the X value at each end of the interval.

$$\text{For } X = 55\text{: } z = \frac{X - \mu}{\sigma} = \frac{55 - 58}{10} = \frac{-3}{10} = -0.30$$

$$\text{For } X = 65\text{: } z = \frac{X - \mu}{\sigma} = \frac{65 - 58}{10} = \frac{7}{10} = 0.70$$

Looking again at Figure 6.11, we see that the proportion we are seeking can be divided into two sections: (1) the area left of the mean, and (2) the area right of the mean. The first area is the proportion between the mean and $z = -0.30$, and the second is the proportion between the mean and $z = +0.70$. Using column D of the unit normal table, these two proportions are 0.1179 and 0.2580. The total proportion is obtained by adding these two sections:

$$p(55 < X < 65) = p(-0.30 < z < +0.70) = 0.1179 + 0.2580 = 0.3759$$

EXAMPLE 6.7 Using the same distribution of driving speeds from the previous example, what proportion of cars are traveling between 65 and 75 miles per hour?

$$p(65 < X < 75) = ?$$

The distribution is shown in Figure 6.12 with the appropriate area shaded. Again, we start by determining the z-score corresponding to each end of the interval.

$$\text{For } X = 75\text{: } \quad z = \frac{X - \mu}{\sigma} = \frac{75 - 58}{10} = \frac{17}{10} = 1.70$$

$$\text{For } X = 65\text{: } \quad z = \frac{X - \mu}{\sigma} = \frac{65 - 58}{10} = \frac{7}{10} = 0.70$$

FIGURE 6.11
The distribution for Example 6.6.

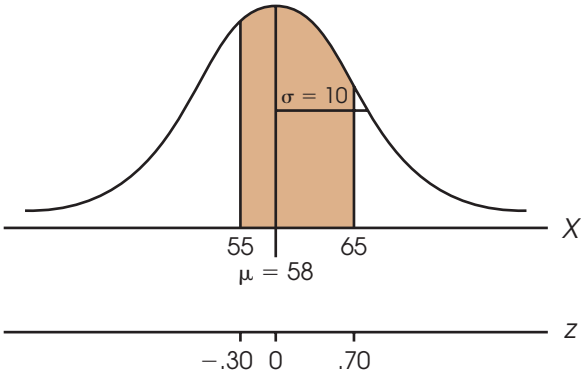
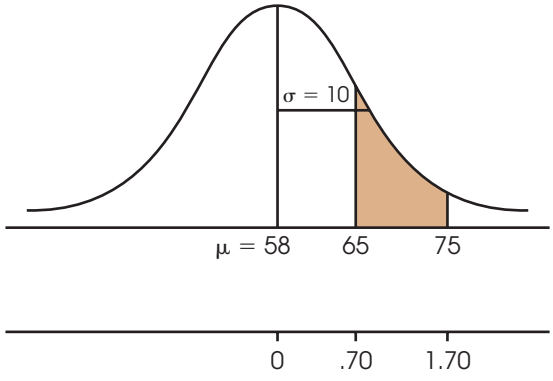


FIGURE 6.12
The distribution for Example 6.7.

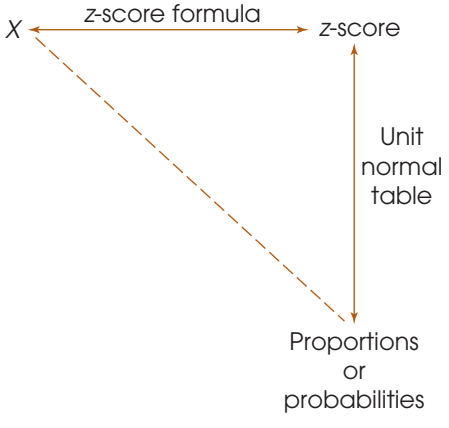


There are several different ways to use the unit normal table to find the proportion between these two z-scores. For this example, we use the proportions in the tail of the distribution (column C). According to column C in the unit normal table, the proportion in the tail beyond $z = 0.70$ is $p = 0.2420$. Note that this proportion includes the section that we want, but it also includes an extra, unwanted section located in the tail beyond $z = 1.70$. Locating $z = 1.70$ in the table, and reading across the row to column C, we see that the unwanted section is $p = 0.0446$. To obtain the correct answer, we subtract the unwanted portion from the total proportion in the tail beyond $z = 0.70$.

$$p(65 < X < 75) = p(0.70 < z < 1.70) = 0.2420 - 0.0446 = 0.1974$$

Finding scores corresponding to specific proportions or probabilities In the previous three examples, the problem was to find the proportion or probability corresponding to specific X values. The two-step process for finding these proportions is shown in Figure 6.13. Thus far, we have only considered examples that move in a clockwise direction around the triangle shown in the figure; that is, we start with an X value that is transformed into a z-score, and then we use the unit normal table to look up the

FIGURE 6.13
Determining probabilities or proportions for a normal distribution is shown as a two-step process with z-scores as an intermediate stop along the way. Note that you cannot move directly along the dashed line between X values and probabilities and proportions. Instead, you must follow the solid lines around the corner.



appropriate proportion. You should realize, however, that it is possible to reverse this two-step process so that we move backward, or counterclockwise, around the triangle. This reverse process allows us to find the score (X value) corresponding to a specific proportion in the distribution. Following the lines in Figure 6.13, we begin with a specific proportion, use the unit normal table to look up the corresponding z-score, and then transform the z-score into an X value. The following example demonstrates this process.

EXAMPLE 6.8

The U.S. Census Bureau (2005) reports that Americans spend an average of $\mu = 24.3$ minutes commuting to work each day. Assuming that the distribution of commuting times is normal with a standard deviation of $\sigma = 10$ minutes, how much time do you have to spend commuting each day to be in the highest 10% nationwide? (An alternative form of the same question is presented in Box 6.1.) The distribution is shown in Figure 6.14 with a portion representing approximately 10% shaded in the right-hand tail.

In this problem, we begin with a proportion (10% or 0.10), and we are looking for a score. According to the map in Figure 6.13, we can move from p (proportion) to X (score) via z-scores. The first step is to use the unit normal table to find the z-score that corresponds to a proportion of 0.10 in the tail. First, scan the values in column C to locate the row that has a proportion of 0.10 in the tail of the distribution. Note that you will not find 0.1000 exactly, but locate the closest value possible. In this case, the closest value is 0.1003. Reading across the row, we find $z = 1.28$ in column A.

The next step is to determine whether the z-score is positive or negative. Remember that the table does not specify the sign of the z-score. Looking at the distribution in Figure 6.14, you should realize that the score we want is above the mean, so the z-score is positive, $z = +1.28$.

The final step is to transform the z-score into an X value. By definition, a z-score of +1.28 corresponds to a score that is located above the mean by 1.28 standard deviations. One standard deviation is equal to 10 points ($\sigma = 10$), so 1.28 standard deviations is

$$1.28\sigma = 1.28(10) = 12.8 \text{ points}$$

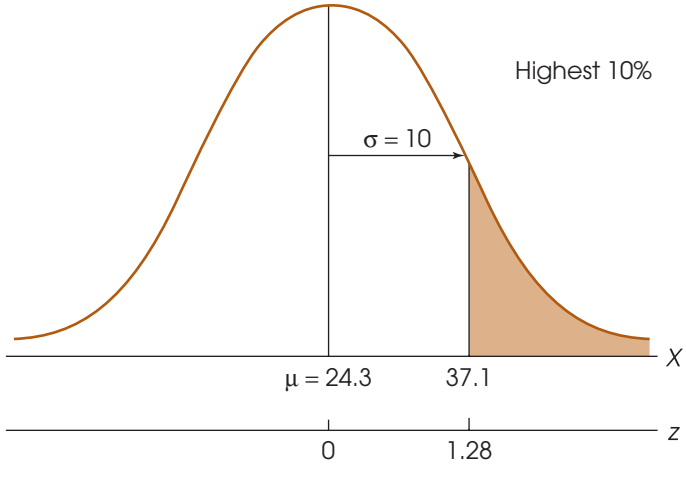


FIGURE 6.14

The distribution of commuting times for American workers. The problem is to find the score that separates the highest 10% of commuting times from the rest.

Thus, our score is located above the mean ($\mu = 24.3$) by a distance of 12.8 points. Therefore,

$$X = 24.3 + 12.8 = 37.1$$

The answer for our original question is that you must commute at least 37.1 minutes a day to be in the top 10% of American commuters.

EXAMPLE 6.9

Again, the distribution of commuting time for American workers is normal with a mean of $\mu = 24.3$ minutes and a standard deviation of $\sigma = 10$ minutes. For this example, we find the range of values that defines the middle 90% of the distribution. The entire distribution is shown in Figure 6.15 with the middle portion shaded.

The 90% (0.9000) in the middle of the distribution can be split in half with 45% (0.4500) on each side of the mean. Looking up 0.4500, in column D of the unit normal table, you will find that the exact proportion is not listed. However, you will find 0.4495 and 0.4505, which are equally close. Technically, either value is acceptable, but we use 0.4505 so that the total area in the middle is at least 90%. Reading across the row, you should find a z-score of $z = 1.65$ in column A. Thus, the z-score at the right boundary is $z = +1.65$ and the z-score at the left boundary is $z = -1.65$. In either case, a z-score of 1.65 indicates a location that is 1.65 standard deviations away from the mean. For the distribution of commuting times, one standard deviation is $\sigma = 10$, so 1.65 standard deviations is a distance of

$$1.65\sigma = 1.65(10) = 16.5 \text{ points}$$

Therefore, the score at the right-hand boundary is located above the mean by 16.5 points and corresponds to $X = 24.3 + 16.5 = 40.8$. Similarly, the score at the left-hand boundary is below the mean by 16.5 points and corresponds to $X = 24.3 - 16.5 = 7.8$. The middle 90% of the distribution corresponds to values between 7.8 and 40.8. Thus, 90% of American commuters spend between 7.8 and 40.8 minutes commuting to work each day. Only 10% of commuters spend either more time or less time.

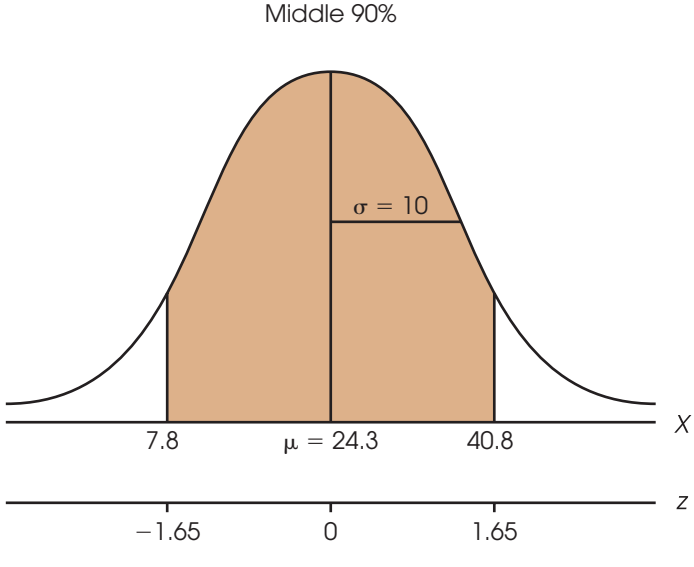


FIGURE 6.15
The distribution of commuting times for American workers. The problem is to find the middle 90% of the distribution.

LEARNING CHECK

1. For a normal distribution with a mean of $\mu = 60$ and a standard deviation of $\sigma = 12$, find each probability value requested.
   - **a.** $p(X > 66)$
   - **b.** $p(X < 75)$
   - **c.** $p(X < 57)$
   - **d.** $p(48 < X < 72)$
2. Scores on the Mathematics section of the SAT Reasoning Test form a normal distribution with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$.
   - **a.** If the state college only accepts students who score in the top 60% on this test, what is the minimum score needed for admission?
   - **b.** What is the minimum score necessary to be in the top 10% of the distribution?
   - **c.** What scores form the boundaries for the middle 50% of the distribution?
3. What is the probability of selecting a score greater than 45 from a positively skewed distribution with $\mu = 40$ and $\sigma = 10$? (Be careful.)

ANSWERS

1. **a.** $p = 0.3085$
   b. $p = 0.8944$
   c. $p = 0.4013$
   d. $p = 0.6826$
2. **a.** $z = -0.25$; $X = 475$
   b. $z = 1.28$; $X = 628$
   c. $z = \pm 0.67$; $X = 433$ and $X = 567$
3. You cannot obtain the answer. The unit normal table cannot be used to answer this question because the distribution is not normal.

6.4 PROBABILITY AND THE BINOMIAL DISTRIBUTION

When a variable is measured on a scale consisting of exactly two categories, the resulting data are called binomial. The term *binomial* can be loosely translated as "two names," referring to the two categories on the measurement scale.

Binomial data can occur when a variable naturally exists with only two categories. For example, people can be classified as male or female, and a coin toss results in either heads or tails. It also is common for a researcher to simplify data by collapsing the scores into two categories. For example, a psychologist may use personality scores to classify people as either high or low in aggression.

In binomial situations, the researcher often knows the probabilities associated with each of the two categories. With a balanced coin, for example, $p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$. The question of interest is the number of times each category occurs in a series of trials or in a sample of individuals. For example:

> What is the probability of obtaining 15 heads in 20 tosses of a balanced coin?
> What is the probability of obtaining more than 40 introverts in a sampling of 50 college freshmen?

As we shall see, the normal distribution serves as an excellent model for computing probabilities with binomial data.

THE BINOMIAL DISTRIBUTION

To answer probability questions about binomial data, we must examine the binomial distribution. To define and describe this distribution, we first introduce some notation.

1. The two categories are identified as A and B.
2. The probabilities (or proportions) associated with each category are identified as

   $p = p(A) =$ the probability of A
   $q = p(B) =$ the probability of B

   Notice that $p + q = 1.00$ because A and B are the only two possible outcomes.
3. The number of individuals or observations in the sample is identified by n.
4. The variable X refers to the number of times category A occurs in the sample.

Notice that X can have any value from 0 (none of the sample is in category A) to n (all of the sample is in category A).

DEFINITION

Using the notation presented here, the **binomial distribution** shows the probability associated with each value of X from $X = 0$ to $X = n$.

A simple example of a binomial distribution is presented next.

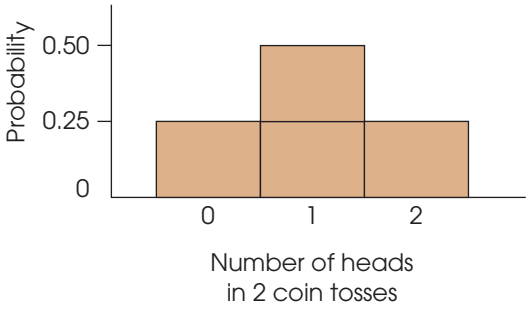
EXAMPLE 6.10

Figure 6.16 shows the binomial distribution for the number of heads obtained in 2 tosses of a balanced coin. This distribution shows that it is possible to obtain as many as 2 heads or as few as 0 heads in 2 tosses. The most likely outcome (highest probability) is to obtain exactly 1 head in 2 tosses. The construction of this binomial distribution is discussed in detail next.

For this example, the event we are considering is a coin toss. There are two possible outcomes, heads and tails. We assume the coin is balanced, so

$$p = p(\text{heads}) = \frac{1}{2}$$

FIGURE 6.16
The binomial distribution showing the probability for the number of heads in 2 tosses of a balanced coin.



$$q = p(\text{tails}) = \frac{1}{2}$$

We are looking at a sample of $n = 2$ tosses, and the variable of interest is

$X =$ the number of heads

To construct the binomial distribution, we look at all of the possible outcomes from tossing a coin 2 times. The complete set of 4 outcomes is listed in the following table.

1st Toss	2nd Toss	
Heads	Heads	(Both heads)
Heads	Tails	(Each sequence has exactly 1 head)
Tails	Heads	
Tails	Tails	(No heads)

Notice that there are 4 possible outcomes when you toss a coin 2 times. Only 1 of the 4 outcomes has 2 heads, so the probability of obtaining 2 heads is $p = \frac{1}{4}$. Similarly, 2 of the 4 outcomes have exactly 1 head, so the probability of 1 head is $p = \frac{2}{4} = \frac{1}{2}$. Finally, the probability of no heads ($X = 0$) is $p = \frac{1}{4}$. These are the probabilities shown in Figure 6.16.

Note that this binomial distribution can be used to answer probability questions. For example, what is the probability of obtaining at least 1 head in 2 tosses? According to the distribution shown in Figure 6.16, the answer is $\frac{3}{4}$.

Similar binomial distributions have been constructed for the number of heads in 4 tosses of a balanced coin and in 6 tosses of a coin (Figure 6.17). It should be obvious from the binomial distributions shown in Figures 6.16 and 6.17 that the binomial

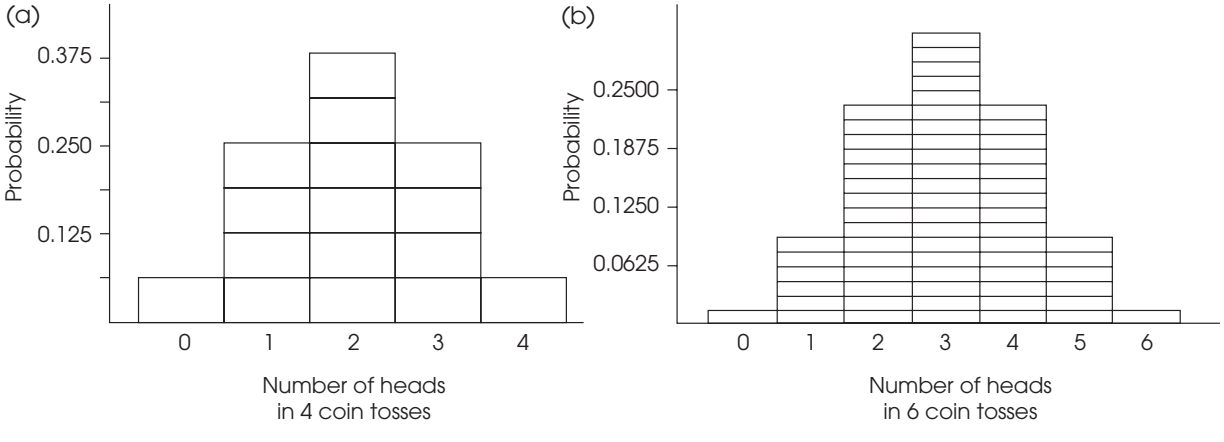


FIGURE 6.17

Binomial distributions showing probabilities for the number of heads (a) in 4 tosses of a balanced coin and (b) in 6 tosses of a balanced coin.

distribution tends toward a normal shape, especially when the sample size (n) is relatively large.

It should not be surprising that the binomial distribution tends to be normal. With $n = 10$ coin tosses, for example, the most likely outcome would be to obtain around $X = 5$ heads. On the other hand, values far from 5 would be very unlikely—you would not expect to get all 10 heads or all 10 tails (0 heads) in 10 tosses. Notice that we have described a normal-shaped distribution: The probabilities are highest in the middle (around $X = 5$), and they taper off as you move toward either extreme.

THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

We have stated that the binomial distribution tends to approximate a normal distribution, particularly when n is large. To be more specific, the binomial distribution is a nearly perfect normal distribution when pn and qn are both equal to or greater than 10. Under these circumstances, the binomial distribution approximates a normal distribution with the following parameters:

The value of 10 for pn or qn is a general guide, not an absolute cutoff. Values slightly less than 10 still provide a good approximation. However, with smaller values the normal approximation becomes less accurate as a substitute for the binomial distribution.

$$\text{Mean: } \mu = pn \tag{6.1}$$

$$\text{standard deviation: } \sigma = \sqrt{npq} \tag{6.2}$$

Within this normal distribution, each value of X has a corresponding z-score,

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}} \tag{6.3}$$

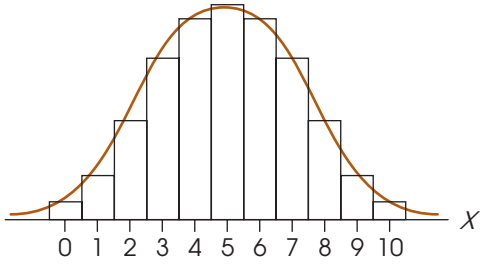
The fact that the binomial distribution tends to be normal in shape means that we can compute probability values directly from z-scores and the unit normal table.

Coin tosses produce discrete events. In a series of coin tosses, you may observe 1 head, 2 heads, 3 heads, and so on, but no values between them are possible (p. 21).

It is important to remember that the normal distribution is only an approximation of a true binomial distribution. Binomial values, such as the number of heads in a series of coin tosses, are *discrete*. The normal distribution is *continuous*. However, the *normal approximation* provides an extremely accurate model for computing binomial probabilities in many situations. Figure 6.18 shows the difference between the true binomial distribution, the discrete histogram, and the normal curve that approximates the binomial distribution. Although the two distributions are slightly different, the area under the distributions is nearly equivalent. *Remember, it is the area under the distribution that is used to find probabilities.*

FIGURE 6.18

The relationship between the binomial distribution and the normal distribution. The binomial distribution is always a discrete histogram, and the normal distribution is a continuous, smooth curve. Each X value is represented by a bar in the histogram or a section of the normal distribution.



To gain maximum accuracy when using the normal approximation, you must remember that each X value in the binomial distribution actually corresponds to a bar in the histogram. In the histogram in Figure 6.18, for example, the score $X = 6$ is represented by a bar that is bounded by real limits of 5.5 and 6.5. The actual probability of $X = 6$ is determined by the area contained in this bar. To approximate this probability using the normal distribution, you should find the area that is contained between the two real limits. Similarly, if you are using the normal approximation to find the probability of obtaining a score greater than $X = 6$, you should use the area beyond the real limit boundary of 6.5. The following example demonstrates how the normal approximation to the binomial distribution is used to compute probability values.

EXAMPLE 6.11

Suppose that you plan to test for ESP (extra-sensory perception) by asking people to predict the suit of a card that is randomly selected from a complete deck. Before you begin your test, however, you need to know what kind of performance is expected from people who do not have ESP and are simply guessing. For these people, there are two possible outcomes, correct or incorrect, on each trial. Because there are four different suits, the probability of a correct prediction (assuming that there is no ESP) is $p = \frac{1}{4}$ and the probability of an incorrect prediction is $q = \frac{3}{4}$. With a series of $n = 48$ trials, this situation meets the criteria for the normal approximation to the binomial distribution:

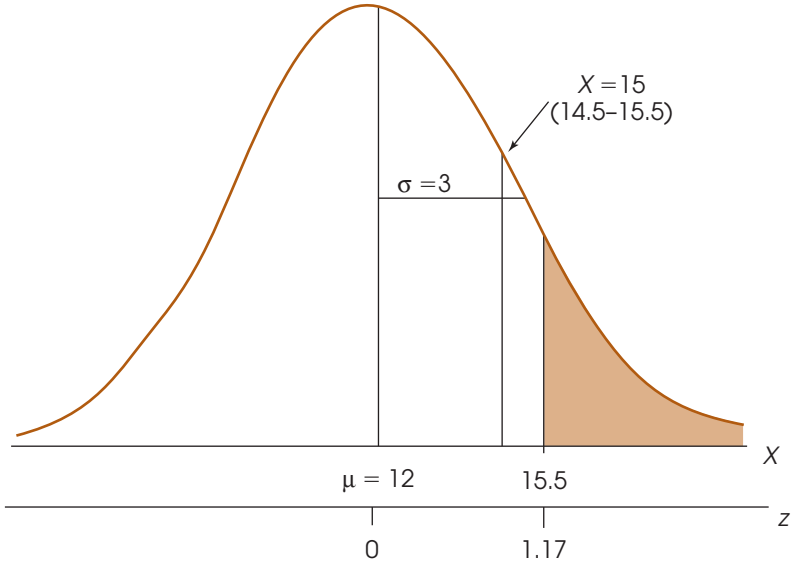
$$pn = \tfrac{1}{4}(48) = 12 \quad qn = \tfrac{3}{4}(48) = 36 \text{ Both are greater than 10.}$$

Thus, the distribution of correct predictions forms a normal-shaped distribution with a mean of $\mu = pn = 12$ and a standard deviation of $\sigma = \sqrt{npq} = \sqrt{9} = 3$. We can use this distribution to determine probabilities for different levels of performance. For example, we can calculate the probability that a person without ESP would guess correctly more than 15 times in a series of 48 trials.

Figure 6.19 shows the binomial distribution that we are considering. Because we want the probability of obtaining *more than* 15 correct predictions, we must find the

FIGURE 6.19

The normal approximation of the binomial distribution discussed in Example 6.11.



Caution: If the question had asked for the probability of *15 or more* correct predictions, we would find the area beyond $X = 14.5$. Read the question carefully.

shaded area in the tail of the distribution beyond $X = 15.5$. (Remember that a score of 15 corresponds to an interval from 14.5 to 15.5. We want scores beyond this interval.) The first step is to find the z-score corresponding to $X = 15.5$.

$$z = \frac{X - \mu}{\sigma} = \frac{15.5 - 12}{3} = 1.17$$

Next, look up the probability in the unit normal table. In this case, we want the proportion in the tail beyond $z = 1.17$. The value from the table is $p = 0.1210$. This is the answer we want. The probability of correctly predicting the suit of a card more than 15 times in a series of 48 trials is only $p = 0.1210$ or 12.10%. Thus, it is very unlikely for an individual without ESP to guess correctly more than 15 out of 48 trials.

LEARNING CHECK

1. Under what circumstances is the normal distribution an accurate approximation of the binomial distribution?
2. In the game Rock-Paper-Scissors, the probability that both players will select the same response and tie is $p = \frac{1}{3}$, and the probability that they will pick different responses is $p = \frac{2}{3}$. If two people play 72 rounds of the game and choose their responses randomly, what is the probability that they will choose the same response (tie) more than 28 times?
3. If you toss a balanced coin 36 times, you would expect, on the average, to get 18 heads and 18 tails. What is the probability of obtaining exactly 18 heads in 36 tosses?

ANSWERS

1. When pn and qn are both greater than 10
2. With $p = \frac{1}{3}$ and $q = \frac{2}{3}$, the binomial distribution is normal with $\mu = 24$ and $\sigma = 4$; $p(X > 28.5) = p(z > 1.13) = 0.1292$.
3. $X = 18$ is an interval with real limits of 17.5 and 18.5. The real limits correspond to $z = \pm 0.17$, and a probability of $p = 0.1350$.

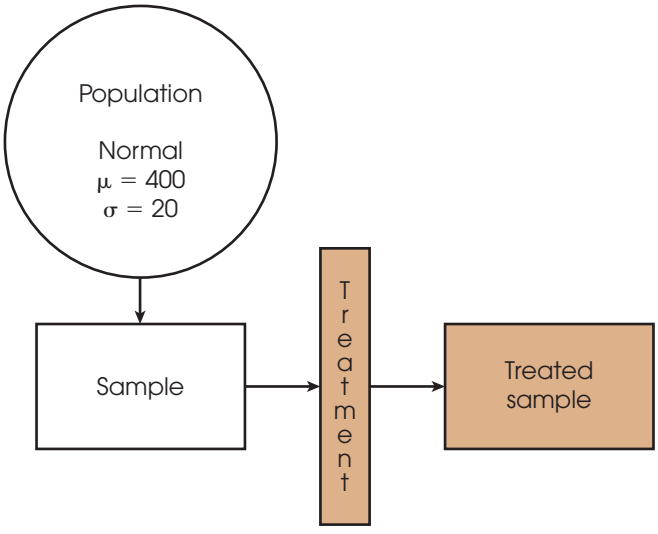
6.5 LOOKING AHEAD TO INFERENTIAL STATISTICS

Probability forms a direct link between samples and the populations from which they come. As we noted at the beginning of this chapter, this link is the foundation for the inferential statistics in future chapters. The following example provides a brief preview of how probability is used in the context of inferential statistics.

We ended Chapter 5 with a demonstration of how inferential statistics are used to help interpret the results of a research study. A general research situation was shown in Figure 5.9 and is repeated here in Figure 6.20. The research begins with a population that forms a normal distribution with a mean of $\mu = 400$ and a standard deviation of $\sigma = 20$. A sample is selected from the population and a treatment is administered to the sample. The goal for the study is to evaluate the effect of the treatment.

FIGURE 6.20

A diagram of a research study. A sample is selected from the population and receives a treatment. The goal is to determine whether the treatment has an effect.



To determine whether the treatment has an effect, the researcher simply compares the treated sample with the original population. If the individuals in the sample have scores around 400 (the original population mean), then we must conclude that the treatment appears to have no effect. On the other hand, if the treated individuals have scores that are noticeably different from 400, then the researcher has evidence that the treatment does have an effect. Notice that the study is using a sample to help answer a question about a population; this is the essence of inferential statistics.

The problem for the researcher is determining exactly what is meant by "noticeably different" from 400. If a treated individual has a score of $X = 415$, is that enough to say that the treatment has an effect? What about $X = 420$ or $X = 450$? In Chapter 5, we suggested that z-scores provide one method for solving this problem. Specifically, we suggested that a z-score value beyond $z = 2.00$ (or –2.00) was an extreme value and, therefore, noticeably different. However, the choice of $z = \pm 2.00$ was purely arbitrary. Now we have another tool, *probability,* to help us decide exactly where to set the boundaries.

Figure 6.21 shows the original population from our hypothetical research study. Note that most of the scores are located close to $\mu = 400$. Also note that we have added boundaries separating the middle 95% of the distribution from the extreme 5%, or 0.0500, in the two tails. Dividing the 0.0500 in half produces proportions of 0.0250 in the right-hand tail and 0.0250 in the left-hand tail. Using column C of the unit normal table, the z-score boundaries for the right and left tails are $z = +1.96$ and $z = -1.96$, respectively.

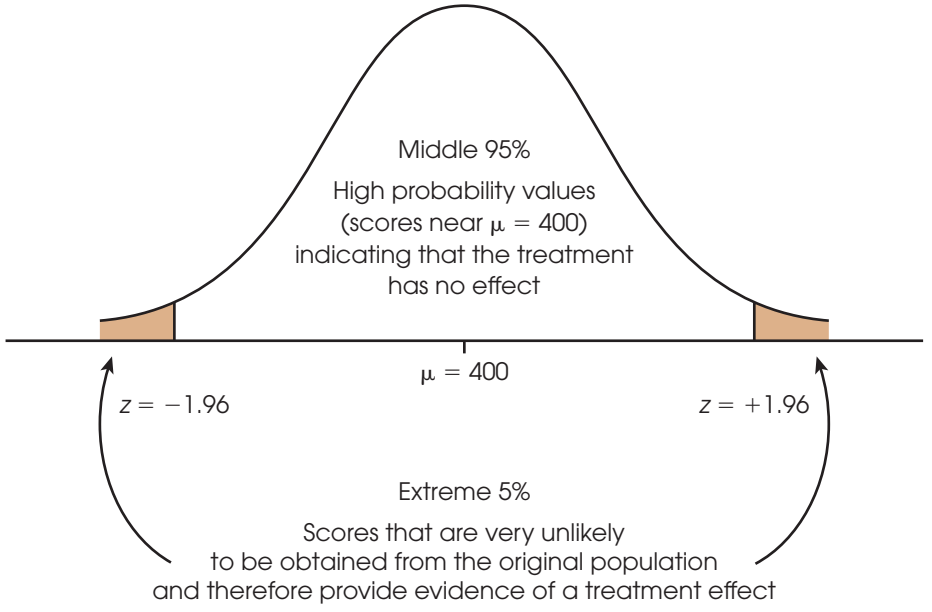
The boundaries set at $z = \pm 1.96$ provide objective criteria for deciding whether our sample provides evidence that the treatment has an effect. Specifically, we use the sample data to help decide between the following two alternatives:

1. The treatment has no effect. After treatment, the scores still average $\mu = 400$.
2. The treatment does have an effect. The treatment changes the scores so that, after treatment, they no longer average $\mu = 400$.

As a starting point, we assume that the first alternative is true and the treatment has no effect. In this case, treated individuals should be no different from the individuals in

FIGURE 6.21

Using probability to evaluate a treatment effect. Values that are extremely unlikely to be obtained from the original population are viewed as evidence of a treatment effect.



the original population, which is shown in Figure 6.21. Notice that, if our assumption is correct, it is extremely unlikely (probability less than 5%) for a treated individual to be outside the ±1.96 boundaries. Therefore, if we obtain a treated individual who is outside the boundaries, we must conclude that the assumption is probably not correct. In this case, we are left with the second alternative (the treatment does have an effect) as the more likely explanation.

Notice that we are comparing the treated sample with the original population to see if the sample is noticeably different. If it is different, we can conclude that the treatment seems to have an effect. Now we are defining "noticeably different" as meaning "very unlikely." Specifically, if the sample is very unlikely to have come from a population of untreated individuals, then we must conclude that the treatment has an effect and has caused the sample to be different.

We are using the sample data and the ±1.96 boundaries, which were determined by probabilities, to make a general decision about the treatment. If the sample falls outside the boundaries we make the following logical conclusion:

a. This kind of sample is very unlikely to occur if the treatment has no effect.

b. Therefore, the treatment must have an effect that changed the sample.

On the other hand, if the sample falls between the ±1.96 boundaries, we conclude:

a. This is the kind of sample that is likely to occur if the treatment has no effect.

b. Therefore, the treatment does not appear to have had an effect.

SUMMARY

1. The probability of a particular event A is defined as a fraction or proportion:

$$p(A) = \frac{\text{number of outcomes classified as } A}{\text{total number of possible outcomes}}$$

2. Our definition of probability is accurate only for random samples. There are two requirements that must be satisfied for a random sample:
   - **a.** Every individual in the population has an equal chance of being selected.
   - **b.** When more than one individual is being selected, the probabilities must stay constant. This means that there must be sampling with replacement.

3. All probability problems can be restated as proportion problems. The "probability of selecting a king from a deck of cards" is equivalent to the "proportion of the deck that consists of kings." For frequency distributions, probability questions can be answered by determining proportions of area. The "probability of selecting an individual with an IQ greater than 108" is equivalent to the "proportion of the whole population that consists of IQs greater than 108."

4. For normal distributions, probabilities (proportions) can be found in the unit normal table. The table provides a listing of the proportions of a normal distribution that correspond to each z-score value. With the table, it is possible to move between X values and probabilities using a two-step procedure:
   - **a.** The z-score formula (Chapter 5) allows you to transform X to z or to change z back to X.
   - **b.** The unit normal table allows you to look up the probability (proportion) corresponding to each z-score or the z-score corresponding to each probability.

5. Percentiles and percentile ranks measure the relative standing of a score within a distribution (see Box 6.1). Percentile rank is the percentage of individuals with scores at or below a particular X value. A percentile is an X value that is identified by its rank. The percentile rank always corresponds to the proportion to the left of the score in question.

6. The binomial distribution is used whenever the measurement procedure classifies individuals into exactly two categories. The two categories are identified as A and B, with probabilities of

$$p(A) = p \quad \text{and} \quad p(B) = q$$

7. The binomial distribution gives the probability for each value of X, where X equals the number of occurrences of category A in a series of n events. For example, X equals the number of heads in $n = 10$ tosses of a coin.

   When pn and qn are both at least 10, the binomial distribution is closely approximated by a normal distribution with

$$\mu = pn \quad \sigma = \sqrt{npq}$$

8. In the normal approximation to the binomial distribution, each value of X has a corresponding z-score:

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}}$$

   With the z-score and the unit normal table, you can find probability values associated with any value of X. For maximum accuracy, you should use the appropriate real limits for X when computing z-scores and probabilities.

KEY TERMS

probability (165)
random sample (167)
independent random sample (167)
sampling with replacement (168)
unit normal table (172)
percentile rank (179)
percentile (179)
binomial distribution (185)
normal approximation (binomial) (187)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 6 on the book companion website.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

The statistics computer package SPSS is not structured to compute probabilities. However, the program does report probability values as part of the inferential statistics that we examine later in this book. In the context of inferential statistics, the probabilities are called *significance levels,* and they warn researchers about the probability of misinterpreting their research results.

FOCUS ON PROBLEM SOLVING

1. We have defined probability as being equivalent to a proportion, which means that you can restate every probability problem as a proportion problem. This definition is particularly useful when you are working with frequency distribution graphs in which the population is represented by the whole graph and probabilities (proportions) are represented by portions of the graph. When working problems with the normal distribution, you always should start with a sketch of the distribution. You should shade the portion of the graph that reflects the proportion you are looking for.
2. Remember that the unit normal table shows only positive z-scores in column A. However, because the normal distribution is symmetrical, the proportions in the table apply to both positive and negative z-score values.

3. A common error for students is to use negative values for proportions on the left-hand side of the normal distribution. Proportions (or probabilities) are always positive: 10% is 10% whether it is in the left or right tail of the distribution.
4. The proportions in the unit normal table are accurate only for normal distributions. If a distribution is not normal, you cannot use the table.
5. For maximum accuracy when using the normal approximation to the binomial distribution, you must remember that each X value is an interval bounded by real limits. For example, a score of $X = 10$ is actually an interval from 9.5 to 10.5. To find the probability of obtaining an X value greater than 10, you should use the real limit 10.5 in the z-score formula. Similarly, to find the probability of obtaining an X value less than 10, you should use the real limit 9.5.

DEMONSTRATION 6.1

FINDING PROBABILITY FROM THE UNIT NORMAL TABLE

A population is normally distributed with a mean of $\mu = 45$ and a standard deviation of $\sigma = 4$. What is the probability of randomly selecting a score that is greater than 43? In other words, what proportion of the distribution consists of scores greater than 43?

STEP 1 **Sketch the distribution.** For this demonstration, the distribution is normal with $\mu = 45$ and $\sigma = 4$. The score of $X = 43$ is lower than the mean and therefore is placed to the left of the mean. The question asks for the proportion corresponding to scores greater than 43, so shade in the area to the right of this score. Figure 6.22 shows the sketch.

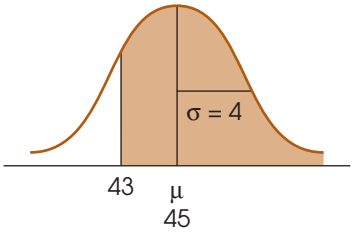
STEP 2 **Transform the X value to a z-score.**

$$z = \frac{X - \mu}{\sigma} = \frac{43 - 45}{4} = \frac{-2}{4} = -0.5$$

STEP 3 **Find the appropriate proportion in the unit normal table.** Ignoring the negative size, locate $z = -0.50$ in column A. In this case, the proportion we want corresponds to the body of the distribution and the value is found in column B. For this example,

$$p(X > 43) = p(z > -0.50) = 0.6915$$

FIGURE 6.22
A sketch of the distribution for Demonstration 6.1.



DEMONSTRATION 6.2

PROBABILITY AND THE BINOMIAL DISTRIBUTION

Suppose that you completely forgot to study for a quiz and now must guess on every question. It is a true/false quiz with $n = 40$ questions. What is the probability that you will get at least 26 questions correct just by chance? Stated in symbols,

$$p(X \geq 26) = ?$$

STEP 1 **Identify *p* and *q*.** This problem is a binomial situation, in which

p = probability of guessing correctly = 0.50

q = probability of guessing incorrectly = 0.50

With $n = 40$ quiz items, both pn and qn are greater than 10. Thus, the criteria for the normal approximation to the binomial distribution are satisfied:

$$pn = 0.50(40) = 20$$

$$qn = 0.50(40) = 20$$

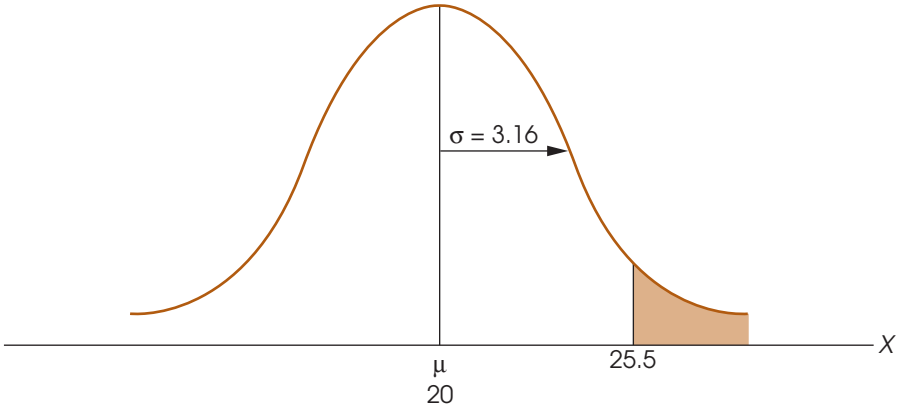
STEP 2 **Identify the parameters, and sketch the binomial distribution.** For a true/false quiz, correct and incorrect guesses are equally likely, $p = q = \frac{1}{2}$. With pn and qn both greater than 10, the normal approximation is appropriate and has a mean and a standard deviation as follows:

$$\mu = pn = 0.5(40) = 20$$
$$\sigma = \sqrt{npq} = \sqrt{10} = 3.16$$

Figure 6.23 shows the distribution. We are looking for the probability of getting $X = 26$ or more questions correct, so we use the lower real limit for 26, which is 25.5.

FIGURE 6.23

The normal approximation of a binomial distribution with $\mu = 20$ and $\sigma = 3.16$. The proportion of all scores equal to or greater than 26 is shaded. Notice that the real lower limit (25.5) for $X = 26$ is used.



STEP 3 The z-score for $X = 25.5$ is calculated as follows:

$$z = \frac{X - pn}{\sqrt{npq}} = \frac{25.5 - 20}{3.16} = +1.74$$

According to the unit normal table, the proportion we want is 0.0409. Thus, the probability of getting at least 26 questions right just by guessing is

$$p(X \geq 26) = 0.0409 \text{ (or } 4.09\%)$$

PROBLEMS

1. A local hardware store has a "Savings Wheel" at the checkout. Customers get to spin the wheel and, when the wheel stops, a pointer indicates how much they will save. The wheel can stop in any one of 50 sections. Of the sections, 10 produce 0% off, 20 sections are for 10% off, 10 sections for 20%, 5 for 30%, 3 for 40%, 1 for 50%, and 1 for 100% off. Assuming that all 50 sections are equally likely,
- **a.** What is the probability that a customer's purchase will be free (100% off)?
- **b.** What is the probability that a customer will get no savings from the wheel (0% off)?
- **c.** What is the probability that a customer will get at least 20% off?

2. A psychology class consists of 14 males and 36 females. If the professor selects names from the class list using *random sampling,*
- **a.** What is the probability that the first student selected will be a female?
- **b.** If a random sample of $n = 3$ students is selected and the first two are both females, what is the probability that the third student selected will be a male?

3. What are the two requirements that must be satisfied for a random sample?

4. What is sampling with replacement, and why is it used?

5. Draw a vertical line through a normal distribution for each of the following z-score locations. Determine whether the tail is on the right or left side of the line and find the proportion in the tail.
- **a.** $z = 2.00$
- **b.** $z = 0.60$
- **c.** $z = -1.30$
- **d.** $z = -0.30$

6. Draw a vertical line through a normal distribution for each of the following z-score locations. Determine whether the body is on the right or left side of the line and find the proportion in the body.
- **a.** $z = 2.20$
- **b.** $z = 1.60$
- **c.** $z = -1.50$
- **d.** $z = -0.70$

7. Find each of the following probabilities for a normal distribution.
- **a.** $p(z > 0.25)$
- **b.** $p(z > -0.75)$
- **c.** $p(z < 1.20)$
- **d.** $p(z < -1.20)$

8. What proportion of a normal distribution is located between each of the following z-score boundaries?
- **a.** $z = -0.50$ and $z = +0.50$
- **b.** $z = -0.90$ and $z = +0.90$
- **c.** $z = -1.50$ and $z = +1.50$

9. Find each of the following probabilities for a normal distribution.
- **a.** $p(-0.25 < z < 0.25)$
- **b.** $p(-2.00 < z < 2.00)$
- **c.** $p(-0.30 < z < 1.00)$
- **d.** $p(-1.25 < z < 0.25)$

10. Find the z-score location of a vertical line that separates a normal distribution as described in each of the following.
- **a.** 20% in the tail on the left
- **b.** 40% in the tail on the right
- **c.** 75% in the body on the left
- **d.** 99% in the body on the right

11. Find the z-score boundaries that separate a normal distribution as described in each of the following.
- **a.** The middle 20% from the 80% in the tails.
- **b.** The middle 50% from the 50% in the tails.
- **c.** The middle 95% from the 5% in the tails.
- **d.** The middle 99% from the 1% in the tails.

12. For a normal distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$, find the proportion of the population corresponding to each of the following scores.
- **a.** Scores greater than 85.
- **b.** Scores less than 100.
- **c.** Scores between 70 and 90.

13. A normal distribution has a mean of $\mu = 50$ and a standard deviation of $\sigma = 12$. For each of the following scores, indicate whether the tail is to the right or left of the score and find the proportion of the distribution located in the tail.
- **a.** $X = 53$
- **b.** $X = 44$
- **c.** $X = 68$
- **d.** $X = 38$

14. IQ test scores are standardized to produce a normal distribution with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. Find the proportion of the population in each of the following IQ categories.
- **a.** Genius or near genius: IQ greater than 140
- **b.** Very superior intelligence: IQ between 120 and 140
- **c.** Average or normal intelligence: IQ between 90 and 109

15. The distribution of scores on the SAT is approximately normal with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. For the population of students who have taken the SAT,
- **a.** What proportion have SAT scores greater than 700?
- **b.** What proportion have SAT scores greater than 550?
- **c.** What is the minimum SAT score needed to be in the highest 10% of the population?
- **d.** If the state college only accepts students from the top 60% of the SAT distribution, what is the minimum SAT score needed to be accepted?

16. The distribution of SAT scores is normal with $\mu = 500$ and $\sigma = 100$.
- **a.** What SAT score, X value, separates the top 15% of the distribution from the rest?
- **b.** What SAT score, X value, separates the top 10% of the distribution from the rest?
- **c.** What SAT score, X value, separates the top 2% of the distribution from the rest?

17. A recent newspaper article reported the results of a survey of well-educated suburban parents. The responses to one question indicated that by age 2, children were watching an average of $\mu = 60$ minutes of television each day. Assuming that the distribution of television-watching times is normal with a standard deviation of $\sigma = 20$ minutes, find each of the following proportions.
- **a.** What proportion of 2-year-old children watch more than 90 minutes of television each day?
- **b.** What proportion of 2-year-old children watch less than 20 minutes a day?

18. Information from the Department of Motor Vehicles indicates that the average age of licensed drivers is $\mu = 45.7$ years with a standard deviation of $\sigma = 12.5$ years. Assuming that the distribution of drivers' ages is approximately normal,
- **a.** What proportion of licensed drivers are older than 50 years old?
- **b.** What proportion of licensed drivers are younger than 30 years old?

19. A consumer survey indicates that the average household spends $\mu = \$185$ on groceries each week. The distribution of spending amounts is approximately normal with a standard deviation of $\sigma = \$25$. Based on this distribution,
- **a.** What proportion of the population spends more than \$200 per week on groceries?
- **b.** What is the probability of randomly selecting a family that spends less than \$150 per week on groceries?
- **c.** How much money do you need to spend on groceries each week to be in the top 20% of the distribution?

20. Over the past 10 years, the local school district has measured physical fitness for all high school freshmen. During that time, the average score on a treadmill endurance task has been $\mu = 19.8$ minutes with a standard deviation of $\sigma = 7.2$ minutes. Assuming that the distribution is approximately normal, find each of the following probabilities.
- **a.** What is the probability of randomly selecting a student with a treadmill time greater than 25 minutes? In symbols, $p(X > 25) = ?$
- **b.** What is the probability of randomly selecting a student with a time greater than 30 minutes? In symbols, $p(X > 30) = ?$
- **c.** If the school required a minimum time of 10 minutes for students to pass the physical education course, what proportion of the freshmen would fail?

21. Rochester, New York, averages $\mu = 21.9$ inches of snow for the month of December. The distribution of snowfall amounts is approximately normal with a standard deviation of $\sigma = 6.5$ inches. This year, a local jewelry store is advertising a refund of 50% off of all purchases made in December, if Rochester finishes the month with more than 3 feet (36 inches)

of total snowfall. What is the probability that the jewelry store will have to pay off on its promise?

22. A multiple-choice test has 48 questions, each with four response choices. If a student is simply guessing at the answers,
- **a.** What is the probability of guessing correctly for any question?
- **b.** On average, how many questions would a student get correct for the entire test?
- **c.** What is the probability that a student would get more than 15 answers correct simply by guessing?
- **d.** What is the probability that a student would get 15 or more answers correct simply by guessing?

23. A true/false test has 40 questions. If a students is simply guessing at the answers,
- **a.** What is the probability of guessing correctly for any one question?
- **b.** On average, how many questions would the student get correct for the entire test?
- **c.** What is the probability that the student would get more than 25 answers correct simply by guessing?
- **d.** What is the probability that the student would get 25 or more answers correct simply by guessing?

24. A roulette wheel has alternating red and black numbered slots into one of which the ball finally stops to determine the winner. If a gambler always bets on black to win, what is the probability of winning at least 24 times in a series of 36 spins? (Note that *at least* 24 wins means 24 or more.)

25. One test for ESP involves using Zener cards. Each card shows one of five different symbols (square, circle, star, cross, wavy lines), and the person being tested has to predict the shape on each card before it is selected. Find each of the probabilities requested for a person who has no ESP and is just guessing.
- **a.** What is the probability of correctly predicting 20 cards in a series of 100 trials?
- **b.** What is the probability of correctly predicting more than 30 cards in a series of 100 trials?
- **c.** What is the probability of correctly predicting 50 or more cards in a series of 200 trials?

26. A trick coin has been weighted so that heads occurs with a probability of $p = \frac{2}{3}$, and $p(\text{tails}) = \frac{1}{3}$. If you toss this coin 72 times,
- **a.** How many heads would you expect to get on average?
- **b.** What is the probability of getting more than 50 heads?
- **c.** What is the probability of getting exactly 50 heads?

27. For a balanced coin:
- **a.** What is the probability of getting more than 30 heads in 50 tosses?
- **b.** What is the probability of getting more than 60 heads in 100 tosses?
- **c.** Parts a and b both asked for the probability of getting more than 60% heads in a series of coin tosses ($\frac{30}{50} = \frac{60}{100} = 60\%$). Why do you think the two probabilities are different?

28. A national health organization predicts that 20% of American adults will get the flu this season. If a sample of 100 adults is selected from the population,
- **a.** What is the probability that at least 25 of the people will be diagnosed with the flu? (Be careful: "at least 25" means "25 or more.")
- **b.** What is the probability that fewer than 15 of the people will be diagnosed with the flu? (Be careful: "fewer than 15" means "14 or less.")

CHAPTER

7

Probability and Samples: The Distribution of Sample Means

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter and section before proceeding.

- Random sampling (Chapter 6)
- Probability and the normal distribution (Chapter 6)
- *z*-Scores (Chapter 5)

Preview

In this chapter we extend the topic of probability to cover larger samples; specifically, samples that have more than one score. Fortunately, you already know the one basic fact governing probability for samples:

Samples tend to be similar to the populations from which they are taken.

For example, if you take a sample from a population that consists of 75% females and only 25% males, you probably will get a sample that has more females than males. Or, if you select a sample from a population for which the average age is $\mu = 21$ years, you probably will get a sample with an average age around 21 years. We are confident that you already know this basic fact because research indicates that even 8-month-old infants understand this basic law of sampling.

Xu and Garcia (2008) began one experiment by showing 8-month-old infants a large box filled with ping-pong balls. The box was brought onto a puppet stage and the front panel was opened to reveal the balls inside. The box contained either mostly red with a few white balls or mostly white with a few red balls. The experimenter alternated between the two boxes until the infants had seen both displays several times. After the infants were familiar with the boxes, the researchers began a series of test trials. On each trial, the box was brought on stage with the front panel closed. The researcher reached in the box and, one at a time, drew out a sample of five balls. The balls were placed in a transparent container next to the box. On half of the trials, the sample was rigged to have 1 red ball and 4 white balls. For the other half, the sample had 1 white ball and 4 red balls. The researchers then removed the front panel to reveal the contents of the box and recorded how long the infants continued to look at the box. The contents of the box were either consistent with the sample, and, therefore, expected, or inconsistent with the sample, and, therefore, unexpected. An *expected* outcome, for example, means that a sample with 4 red balls and 1 white ball should come from a box with mostly red balls. This same sample is *unexpected* from a box with mostly white balls. The results showed that the infants consistently looked longer at the unexpected outcome ($M = 9.9$ seconds) than at the expected outcome ($M = 7.5$ seconds), indicating that the infants considered the unexpected outcome surprising and more interesting than the expected outcome.

The Problem: Xu and Garcia's results strongly suggest that even 8-month-old infants understand the basic principles that determine which samples have high probability and which have low probability. Nevertheless, whenever you are picking ping pong balls from a box or recruiting people to participate in a research study, it usually is possible to obtain thousands or even millions of different samples from the same population. Under these circumstances, how can we determine the probability for obtaining any specific sample?

The Solution: In this chapter we introduce the *distribution of sample means,* which allows us to find the exact probability of obtaining a specific sample mean from a specific population. This distribution describes the entire set of all the possible sample means for any sized sample. Because we can describe the entire set, we can find probabilities associated with specific sample means. (Recall from Chapter 6 that probabilities are equivalent to proportions of the entire distribution.) Also, because the distribution of sample means tends to be normal, it is possible to find probabilities using z-scores and the unit normal table. Although it is impossible to predict exactly which sample will be obtained, the probabilities allow researchers to determine which samples are likely (and which are very unlikely).

7.1 SAMPLES AND POPULATIONS

The preceding two chapters presented the topics of z-scores and probability. Whenever a score is selected from a population, you should be able to compute a z-score that describes exactly where the score is located in the distribution. If the population is normal, you also should be able to determine the probability value for obtaining any individual score. In a normal distribution, for example, any score located in the tail of the distribution beyond $z = +2.00$ is an extreme value, and a score this large has a probability of only $p = 0.0228$.

However, the *z*-scores and probabilities that we have considered so far are limited to situations in which the sample consists of a single score. Most research studies involve much larger samples, such as $n = 25$ preschool children or $n = 100$ American Idol contestants. In these situations, the sample mean, rather than a single score, is used to answer questions about the population. In this chapter we extend the concepts of *z*-scores and probability to cover situations with larger samples. In particular, we introduce a procedure for transforming a sample mean into a *z*-score. Thus, a researcher is able to compute a *z*-score that describes an entire sample. As always, a *z*-score near zero indicates a central, representative sample; a *z*-score beyond +2.00 or –2.00 indicates an extreme sample. Thus, it is possible to describe how any specific sample is related to all the other possible samples. In addition, we can use the *z*-scores to look up probabilities for obtaining certain samples, no matter how many scores the sample contains.

In general, the difficulty of working with samples is that a sample provides an incomplete picture of the population. Suppose, for example, a researcher randomly selects a sample of $n = 25$ students from the state college. Although the sample should be representative of the entire student population, there are almost certainly some segments of the population that are not included in the sample. In addition, any statistics that are computed for the sample are not identical to the corresponding parameters for the entire population. For example, the average IQ for the sample of 25 students is not the same as the overall mean IQ for the entire population. This difference, or *error,* between sample statistics and the corresponding population parameters is called *sampling error* and was illustrated in Figure 1.2 (p. 201).

DEFINITION

Sampling error is the natural discrepancy, or amount of error, between a sample statistic and its corresponding population parameter.

Furthermore, samples are variable; they are not all the same. If you take two separate samples from the same population, the samples are different. They contain different individuals, they have different scores, and they have different sample means. How can you tell which sample gives the best description of the population? Can you even predict how well a sample describes its population? What is the probability of selecting a sample with specific characteristics? These questions can be answered once we establish the rules that relate samples and populations.

7.2 THE DISTRIBUTION OF SAMPLE MEANS

As noted, two separate samples probably are different even though they are taken from the same population. The samples have different individuals, different scores, different means, and so on. In most cases, it is possible to obtain thousands of different samples from one population. With all these different samples coming from the same population, it may seem hopeless to try to establish some simple rules for the relationships between samples and populations. Fortunately, however, the huge set of possible samples forms a relatively simple and orderly pattern that makes it possible to predict the characteristics of a sample with some accuracy. The ability to predict sample characteristics is based on the *distribution of sample means*.

DEFINITION

The **distribution of sample means** is the collection of sample means for all of the possible random samples of a particular size (n) that can be obtained from a population.

Notice that the distribution of sample means contains *all of the possible samples*. It is necessary to have all of the possible values to compute probabilities. For example, if the entire set contains exactly 100 samples, then the probability of obtaining any specific sample is 1 out of 100: $p = \frac{1}{100}$ (Box 7.1).

Also, you should notice that the distribution of sample means is different from the distributions that we have considered before. Until now we always have discussed distributions of scores; now the values in the distribution are not scores, but statistics (sample means). Because statistics are obtained from samples, a distribution of statistics is referred to as a *sampling distribution*.

DEFINITION

A **sampling distribution** is a distribution of statistics obtained by selecting all of the possible samples of a specific size from a population.

Thus, the distribution of sample means is an example of a sampling distribution. In fact, it often is called the sampling distribution of M.

If you actually wanted to construct the distribution of sample means, you would first select a random sample of a specific size (n) from a population, calculate the sample mean, and place the sample mean in a frequency distribution. Then you would select another random sample with the same number of scores. Again, you would calculate the sample mean and add it to your distribution. You would continue selecting samples and calculating means, over and over, until you had the complete set of all the possible random samples. At this point, your frequency distribution would show the distribution of sample means.

We demonstrate the process of constructing a distribution of sample means in Example 7.1, but first we use common sense and a little logic to predict the general characteristics of the distribution.

1. The sample means should pile up around the population mean. Samples are not expected to be perfect but they are representative of the population. As a result, most of the sample means should be relatively close to the population mean.

BOX 7.1 PROBABILITY AND THE DISTRIBUTION OF SAMPLE MEANS

I have a bad habit of losing playing cards. This habit is compounded by the fact that I always save the old deck in the hope that someday I will find the missing cards. As a result, I have a drawer filled with partial decks of playing cards. Suppose that I take one of these almost-complete decks, shuffle the cards carefully, and then randomly select one card. What is the probability that I will draw a king?

You should realize that it is impossible to answer this probability question. To find the probability of selecting a king, you must know how many cards are in the deck and exactly which cards are missing. (It is crucial that you know whether any kings are missing.) The point of this simple example is that any probability question requires that you have complete information about the population from which the sample is being selected. In this case, you must know all of the possible cards in the deck before you can find the probability for selecting any specific card.

In this chapter, we are examining probability and sample means. To find the probability for any specific sample mean, you first must know *all of the possible sample means*. Therefore, we begin by defining and describing the set of all possible sample means that can be obtained from a particular population. Once we have specified the complete set of all possible sample means (i.e., the distribution of sample means), we can find the probability of selecting any specific sample means.

2. The pile of sample means should tend to form a normal-shaped distribution. Logically, most of the samples should have means close to μ, and it should be relatively rare to find sample means that are substantially different from μ. As a result, the sample means should pile up in the center of the distribution (around μ) and the frequencies should taper off as the distance between M and μ increases. This describes a normal-shaped distribution.
3. In general, the larger the sample size, the closer the sample means should be to the population mean, μ. Logically, a large sample should be a better representative than a small sample. Thus, the sample means obtained with a large sample size should cluster relatively close to the population mean; the means obtained from small samples should be more widely scattered.

As you will see, each of these three commonsense characteristics is an accurate description of the distribution of sample means. The following example demonstrates the process of constructing the distribution of sample means by repeatedly selecting samples from a population.

EXAMPLE 7.1

Consider a population that consists of only 4 scores: 2, 4, 6, 8. This population is pictured in the frequency distribution histogram in Figure 7.1.

Remember that random sampling requires sampling with replacement.

We are going to use this population as the basis for constructing the distribution of sample means for $n = 2$. Remember: This distribution is the collection of sample means from all of the possible random samples of $n = 2$ from this population. We begin by looking at all of the possible samples. For this example, there are 16 different samples, and they are all listed in Table 7.1. Notice that the samples are listed systematically. First, we list all of the possible samples with $X = 2$ as the first score, then all of the possible samples with $X = 4$ as the first score, and so on. In this way, we can be sure that we have all of the possible random samples.

Next, we compute the mean, M, for each of the 16 samples (see the last column of Table 7.1). The 16 means are then placed in a frequency distribution histogram in Figure 7.2. This is the distribution of sample means. Note that the distribution in Figure 7.2 demonstrates two of the characteristics that we predicted for the distribution of sample means.

1. The sample means pile up around the population mean. For this example, the population mean is μ = 5, and the sample means are clustered around a value of 5. It should not surprise you that the sample means tend to approximate the

FIGURE 7.1

Frequency distribution histogram for a population of 4 scores: 2, 4, 6, 8.

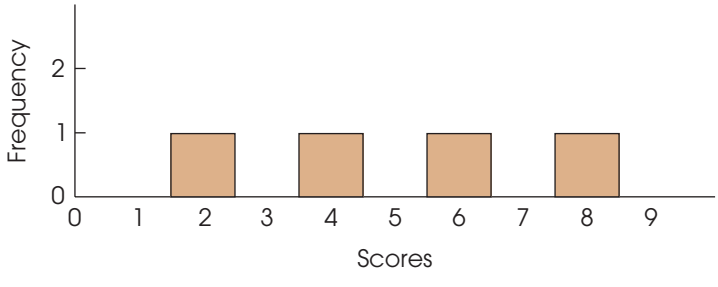


TABLE 7.1

All the possible samples of $n = 2$ scores that can be obtained from the population presented in Figure 7.1. Notice that the table lists *random samples*. This requires sampling with replacement, so it is possible to select the same score twice.

	Scores		Sample Mean
Sample	First	Second	(M)
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

population mean. After all, samples are supposed to be representative of the population.

2. The distribution of sample means is approximately normal in shape. This is a characteristic that is discussed in detail later and is extremely useful because we already know a great deal about probabilities and the normal distribution (Chapter 6).

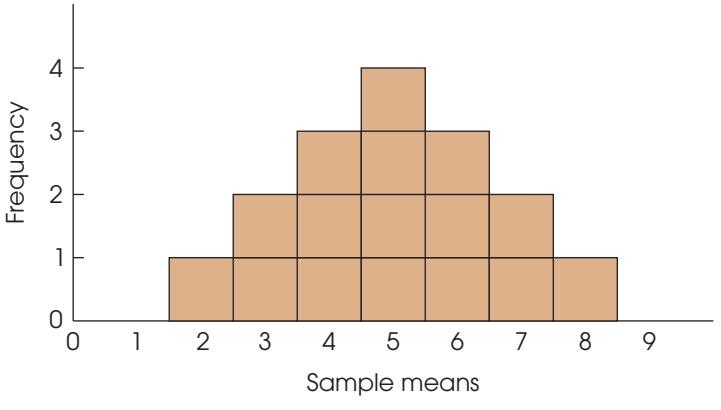
Remember that our goal in this chapter is to answer probability questions about samples with $n > 1$.

Finally, you should notice that we can use the distribution of sample means to answer probability questions about sample means. For example, if you take a sample of $n = 2$ scores from the original population, what is the probability of obtaining a sample mean greater than 7? In symbols,

$$p(M > 7) = ?$$

FIGURE 7.2

The distribution of sample means for $n = 2$. The distribution shows the 16 sample means from Table 7.1.



Because probability is equivalent to proportion, the probability question can be restated as follows: Of all of the possible sample means, what proportion have values greater than 7? In this form, the question is easily answered by looking at the distribution of sample means. All of the possible sample means are pictured (see Figure 7.2), and only 1 out of the 16 means has a value greater than 7. The answer, therefore, is 1 out of 16, or $p = \frac{1}{16}$.

THE CENTRAL LIMIT THEOREM

Example 7.1 demonstrated the construction of the distribution of sample means for an overly simplified situation with a very small population and samples that each contain only $n = 2$ scores. In more realistic circumstances, with larger populations and larger samples, the number of possible samples increases dramatically and it is virtually impossible to actually obtain every possible random sample. Fortunately, it is possible to determine exactly what the distribution of sample means looks like without taking hundreds or thousands of samples. Specifically, a mathematical proposition known as the *central limit theorem* provides a precise description of the distribution that would be obtained if you selected every possible sample, calculated every sample mean, and constructed the distribution of the sample mean. This important and useful theorem serves as a cornerstone for much of inferential statistics. Following is the essence of the theorem.

> **Central limit theorem:** For any population with mean μ and standard deviation σ, the distribution of sample means for sample size n will have a mean of μ and a standard deviation of $\sigma/\sqrt{n}$ and will approach a normal distribution as n approaches infinity.

The value of this theorem comes from two simple facts. First, it describes the distribution of sample means for *any population,* no matter what shape, mean, or standard deviation. Second, the distribution of sample means "approaches" a normal distribution very rapidly. By the time the sample size reaches $n = 30$, the distribution is almost perfectly normal.

Note that the central limit theorem describes the distribution of sample means by identifying the three basic characteristics that describe any distribution: shape, central tendency, and variability. We examine each of these.

THE SHAPE OF THE DISTRIBUTION OF SAMPLE MEANS

It has been observed that the distribution of sample means tends to be a normal distribution. In fact, this distribution is almost perfectly normal if either of the following two conditions is satisfied:

1. The population from which the samples are selected is a normal distribution.
2. The number of scores (n) in each sample is relatively large, around 30 or more.

(As n gets larger, the distribution of sample means more closely approximates a normal distribution. When $n > 30$, the distribution is almost normal, regardless of the shape of the original population.)

As we noted earlier, the fact that the distribution of sample means tends to be normal is not surprising. Whenever you take a sample from a population, you expect the sample mean to be near to the population mean. When you take lots of different samples, you expect the sample means to "pile up" around μ, resulting in a normal-shaped distribution. You can see this tendency emerging (although it is not yet normal) in Figure 7.2.

THE MEAN OF THE DISTRIBUTION OF SAMPLE MEANS: THE EXPECTED VALUE OF *M*

In Example 7.1, the distribution of sample means is centered around the mean of the population from which the samples were obtained. In fact, the average value of all the sample means is exactly equal to the value of the population mean. This fact should be intuitively reasonable; the sample means are expected to be close to the population mean, and they do tend to pile up around μ. The formal statement of this phenomenon is that the mean of the distribution of sample means always is identical to the population mean. This mean value is called the *expected value of* M.

In commonsense terms, a sample mean is "expected" to be near its population mean. When all of the possible sample means are obtained, the average value is identical to μ.

The fact that the average value of *M* is equal to μ was first introduced in Chapter 4 (p. 121) in the context of *biased* versus *unbiased* statistics. The sample mean is an example of an unbiased statistic, which means that, on average, the sample statistic produces a value that is exactly equal to the corresponding population parameter. In this case, the average value of all of the sample means is exactly equal to μ.

DEFINITION

The mean of the distribution of sample means is equal to the mean of the population of scores, μ, and is called the **expected value of *M***.

THE STANDARD ERROR OF *M*

So far, we have considered the shape and the central tendency of the distribution of sample means. To completely describe this distribution, we need one more characteristic, variability. The value we will be working with is the standard deviation for the distribution of sample means. This standard deviation is identified by the symbol σ_M and is called the *standard error of* M.

When the standard deviation was first introduced in Chapter 4, we noted that this measure of variability serves two general purposes. First, the standard deviation describes the distribution by telling whether the individual scores are clustered close together or scattered over a wide range. Second, the standard deviation measures how well any individual score represents the population by providing a measure of how much distance is reasonable to expect between a score and the population mean. The standard error serves the same two purposes for the distribution of sample means.

1. The standard error describes the distribution of sample means. It provides a measure of how much difference is expected from one sample to another. When the standard error is small, then all of the sample means are close together and have similar values. If the standard error is large, then the sample means are scattered over a wide range and there are big differences from one sample to another.
2. Standard error measures how well an individual sample mean represents the entire distribution. Specifically, it provides a measure of how much distance is reasonable to expect between a sample mean and the overall mean for the distribution of sample means. However, because the overall mean is equal to μ, the standard error also provides a measure of how much distance to expect between a sample mean (*M*) and the population mean (μ).

Remember that a sample is not expected to provide a perfectly accurate reflection of its population. Although a sample mean should be representative of the population mean, there typically is some error between the sample and the population. The standard

error measures exactly how much difference is expected on average between a sample mean, M, and the population mean, μ.

DEFINITION

The standard deviation of the distribution of sample means, σ_M, is called the **standard error of M**. The standard error provides a measure of how much distance is expected on average between a sample mean (M) and the population mean (μ).

Once again, the symbol for the standard error is σ_M. The σ indicates that this value is a standard deviation, and the subscript M indicates that it is the standard deviation for the distribution of sample means. Similarly, it is common to use the symbol μ_M to represent the mean of the distribution of sample means. However, μ_M is always equal to μ and our primary interest in inferential statistics is to compare sample means (M) with their population means (μ). Therefore, we simply use the symbol μ to refer to the mean of the distribution of sample means.

The standard error is an extremely valuable measure because it specifies precisely how well a sample mean estimates its population mean—that is, how much error you should expect, on the average, between M and μ. Remember that one basic reason for taking samples is to use the sample data to answer questions about the population. However, you do not expect a sample to provide a perfectly accurate picture of the population. There always is some discrepancy, or error, between a sample statistic and the corresponding population parameter. Now we are able to calculate exactly how much error to expect. For any sample size (n), we can compute the standard error, which measures the average distance between a sample mean and the population mean.

The magnitude of the standard error is determined by two factors: (1) the size of the sample and (2) the standard deviation of the population from which the sample is selected. We examine each of these factors.

The sample size Earlier we predicted, based on common sense, that the size of a sample should influence how accurately the sample represents its population. Specifically, a large sample should be more accurate than a small sample. In general, as the sample size increases, the error between the sample mean and the population mean should decrease. This rule is also known as the *law of large numbers*.

DEFINITION

The **law of large numbers** states that the larger the sample size (n), the more probable it is that the sample mean is close to the population mean.

The population standard deviation As we noted earlier, there is an inverse relationship between the sample size and the standard error: bigger samples have smaller error, and smaller samples have bigger error. At the extreme, the smallest possible sample (and the largest standard error) occurs when the sample consists of $n = 1$ score. At this extreme, each sample is a single score and the distribution of sample means is identical to the original distribution of scores. In this case, the standard deviation for the distribution of sample means, which is the standard error, is identical to the standard deviation for the distribution of scores. In other words, when $n = 1$, the standard error $= \sigma_M$ is identical to the standard deviation $= \sigma$.

When $n = 1$, $\sigma_M = \sigma$ (standard error = standard deviation).

You can think of the standard deviation as the "starting point" for standard error. When $n = 1$, the standard error and the standard deviation are the same: $\sigma_M = \sigma$.

As sample size increases beyond $n = 1$, the sample becomes a more accurate representative of the population, and the standard error decreases. The formula for standard error expresses this relationship between standard deviation and sample size (n).

This formula is contained in the central limit theorem.

$$\text{standard error} = \sigma_M = \frac{\sigma}{\sqrt{n}} \tag{7.1}$$

Note that the formula satisfies all of the requirements for the standard error. Specifically,

a. As sample size (n) increases, the size of the standard error decreases. (Larger samples are more accurate.)

b. When the sample consists of a single score ($n = 1$), the standard error is the same as the standard deviation ($\sigma_M = \sigma$).

In Equation 7.1 and in most of the preceding discussion, we defined standard error in terms of the population standard deviation. However, the population standard deviation (σ) and the population variance (σ^2) are directly related, and it is easy to substitute variance into the equation for standard error. Using the simple equality $\sigma = \sqrt{\sigma^2}$, the equation for standard error can be rewritten as follows:

$$\text{standard error} = \sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} \tag{7.2}$$

Throughout the rest of this chapter (and in Chapter 8), we continue to define standard error in terms of the standard deviation (Equation 7.1). However, in later chapters (starting in Chapter 9) the formula based on variance (Equation 7.2) will become more useful.

Figure 7.3 illustrates the general relationship between standard error and sample size. (The calculations for the data points in Figure 7.3 are presented in Table 7.2.) Again, the basic concept is that the larger a sample is, the more accurately it represents its population. Also note that the standard error decreases in relation to the *square root* of the sample size. As a result, researchers can substantially reduce error by increasing

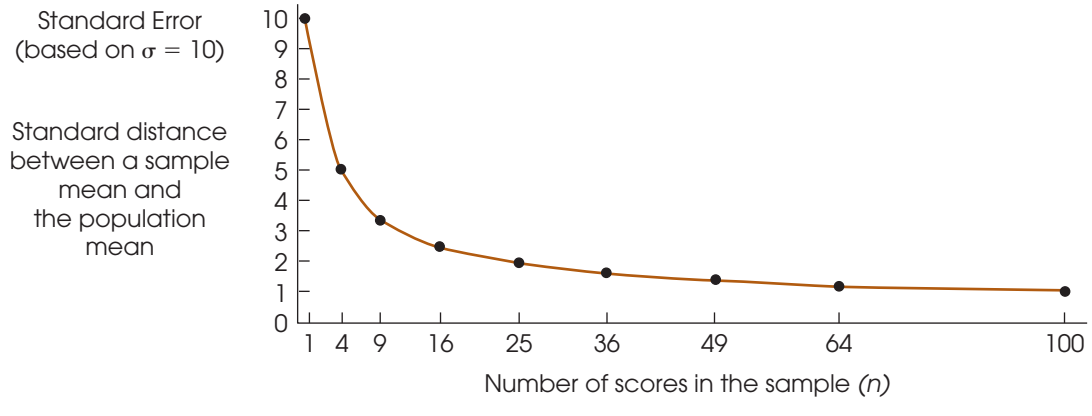


FIGURE 7.3

The relationship between standard error and sample size. As the sample size is increased, there is less error between the sample mean and the population mean.

TABLE 7.2

Calculations for the points shown in Figure 7.3. Again, notice that the size of the standard error decreases as the size of the sample increases.

Sample Size (n)	Standard Error	
1	$\sigma_M = \frac{10}{\sqrt{1}}$	$= 10.00$
4	$\sigma_M = \frac{10}{\sqrt{4}}$	$= 5.00$
9	$\sigma_M = \frac{10}{\sqrt{9}}$	$= 3.33$
16	$\sigma_M = \frac{10}{\sqrt{16}}$	$= 2.50$
25	$\sigma_M = \frac{10}{\sqrt{25}}$	$= 2.00$
49	$\sigma_M = \frac{10}{\sqrt{49}}$	$= 1.43$
64	$\sigma_M = \frac{10}{\sqrt{64}}$	$= 1.25$
100	$\sigma_M = \frac{10}{\sqrt{100}}$	$= 1.00$

sample size up to around $n = 30$. However, increasing sample size beyond $n = 30$ does not produce much additional improvement in how well the sample represents the population.

THREE DIFFERENT DISTRIBUTIONS

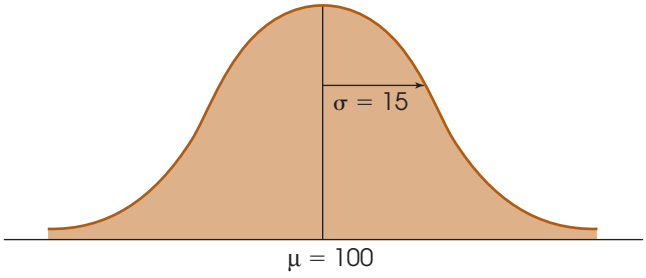
Before we move forward with our discussion of the distribution of sample means, we pause for a moment to emphasize the idea that we are now dealing with three different but interrelated distributions.

1. First, we have the original population of scores. This population contains the scores for thousands or millions of individual people, and it has its own shape, mean, and standard deviation. For example, the population of IQ scores consists of millions of individual IQ scores that form a normal distribution with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. An example of a population is shown in Figure 7.4(a).
2. Next, we have a sample that is selected from the population. The sample consists of a small set of scores for a few people who have been selected to represent the entire population. For example, we could select a sample of $n = 25$ people and measure each individual's IQ score. The 25 scores could be organized in a frequency distribution and we could calculate the sample mean and the sample standard deviation. Note that the sample also has its own shape, mean, and standard deviation. An example of a sample is shown in Figure 7.4(b).
3. The third distribution is the distribution of sample means. This is a theoretical distribution consisting of the sample means obtained from all of the possible random samples of a specific size. For example, the distribution of sample means for samples of $n = 25$ IQ scores would be normal with a mean (expected value) of $\mu = 100$ and a standard deviation (standard error) of $\sigma_M = \frac{15}{\sqrt{25}} = 3$. This

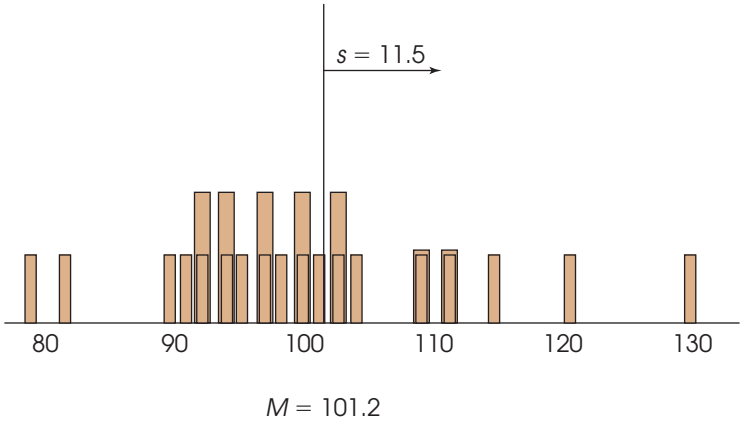
FIGURE 7.4

The distribution. Part (a) shows the population of IQ scores. Part (b) shows a sample of $n = 25$ IQ scores. Part (c) shows the distribution of sample means for samples of $n = 25$ scores. Note that the sample mean from part (b) is one of the thousands of sample means in the part (c) distribution.

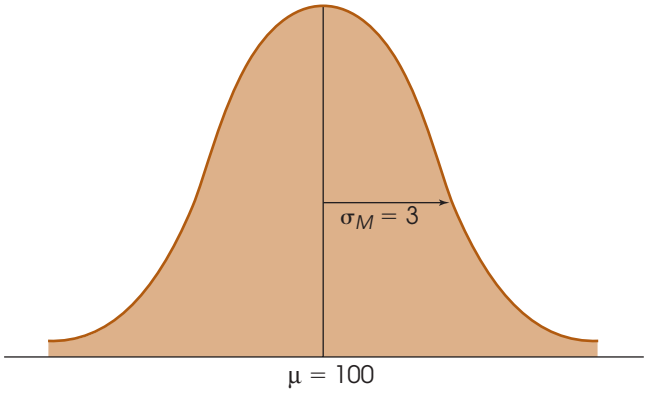
(a) Original population of IQ scores.



(b) A sample of $n = 25$ IQ scores.



(c) The distribution of sample means. Sample means for all the possible random samples of $n = 25$ IQ scores.



distribution, shown in Figure 7.4(c), also has its own shape, mean, and standard deviation.

Note that the scores for the sample [Figure 7.4(b)] were taken from the original population [Figure 7.4(a)] and that the mean for the sample is one of the values contained in the distribution of sample means [Figure 7.4(c)]. Thus, the three distributions are all connected, but they are all distinct.

LEARNING CHECK

1. A population has a mean of $\mu = 50$ and a standard deviation of $\sigma = 12$.
   - **a.** For samples of size $n = 4$, what is the mean (expected value) and the standard deviation (standard error) for the distribution of sample means?
   - **b.** If the population distribution is not normal, describe the shape of the distribution of sample means based on $n = 4$.
   - **c.** For samples of size $n = 36$, what is the mean (expected value) and the standard deviation (standard error) for the distribution of sample means?
   - **d.** If the population distribution is not normal, describe the shape of the distribution of sample means based on $n = 36$.
2. As sample size increases, the value of expected value also increases. (True or false?)
3. As sample size increases, the value of the standard error also increases. (True or false?)

ANSWERS

1. **a.** The distribution of sample means would have a mean of $\mu = 50$ and a standard error of $\sigma_M = 12/\sqrt{4} = 6$.
   b. The distribution of sample means does not satisfy either criterion to be normal. It would not be a normal distribution.
   c. The distribution of sample means is normal and would have a mean of $\mu = 50$ and a standard error of $\sigma_M = 12/\sqrt{36} = 2$.
   d. Because the sample size is greater than 30, the distribution of sample means is a normal distribution.
2. False. The expected value does not depend on sample size.
3. False. The standard error decreases as sample size increases.

7.3 PROBABILITY AND THE DISTRIBUTION OF SAMPLE MEANS

The primary use of the distribution of sample means is to find the probability associated with any specific sample. Recall that probability is equivalent to proportion. Because the distribution of sample means presents the entire set of all possible sample means, we can use proportions of this distribution to determine probabilities. The following example demonstrates this process.

EXAMPLE 7.2

The population of scores on the SAT forms a normal distribution with $\mu = 500$ and $\sigma = 100$. If you take a random sample of $n = 25$ students, what is the probability that the sample mean will be greater than $M = 540$?

Caution: Whenever you have a probability question about a sample mean, you must use the distribution of sample means.

First, you can restate this probability question as a proportion question: Out of all of the possible sample means, what proportion have values greater than 540? You know about "all of the possible sample means"; this is the distribution of sample means. The problem is to find a specific portion of this distribution.

Although we cannot construct the distribution of sample means by repeatedly taking samples and calculating means (as in Example 7.1), we know exactly what the

distribution looks like based on the information from the central limit theorem. Specifically, the distribution of sample means has the following characteristics:

a. The distribution is normal because the population of SAT scores is normal.

b. The distribution has a mean of 500 because the population mean is $\mu = 500$.

c. For $n = 25$, the distribution has a standard error of $\sigma_M = 20$:

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{25}} = \frac{100}{5} = 20$$

This distribution of sample means is shown in Figure 7.5.

We are interested in sample means greater than 540 (the shaded area in Figure 7.5), so the next step is to use a z-score to locate the exact position of $M = 540$ in the distribution. The value 540 is located above the mean by 40 points, which is exactly 2 standard deviations (in this case, exactly 2 standard errors). Thus, the z-score for $M = 540$ is $z = +2.00$.

Because this distribution of sample means is normal, you can use the unit normal table to find the probability associated with $z = +2.00$. The table indicates that 0.0228 of the distribution is located in the tail of the distribution beyond $z = +2.00$. Our conclusion is that it is very unlikely, $p = 0.0228$ (2.28%), to obtain a random sample of $n = 25$ students with an average SAT score greater than 540.

A *z*-SCORE FOR SAMPLE MEANS

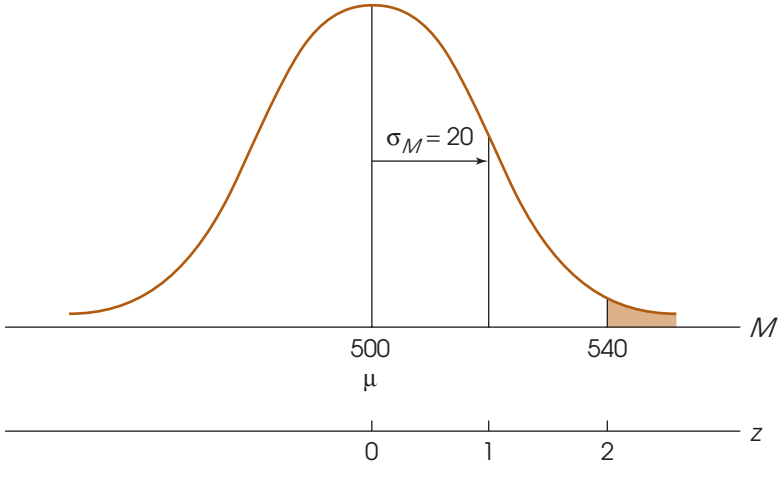
As demonstrated in Example 7.2, it is possible to use a z-score to describe the exact location of any specific sample mean within the distribution of sample means. The z-score tells exactly where the sample mean is located in relation to all of the other possible sample means that could have been obtained. As defined in Chapter 5, a z-score identifies the location with a signed number so that

1. The sign tells whether the location is above (+) or below (−) the mean.

2. The number tells the distance between the location and the mean in terms of the number of standard deviations.

FIGURE 7.5

The distribution of sample means for $n = 25$. Samples were selected from a normal population with $\mu = 500$ and $\sigma = 100$.



However, we are now finding a location within the distribution of sample means. Therefore, we must use the notation and terminology appropriate for this distribution. First, we are finding the location for a sample mean (M) rather than a score (X). Second, the standard deviation for the distribution of sample means is the standard error, σ_M. With these changes, the z-score formula for locating a sample mean is

$$z = \frac{M - \mu}{\sigma_M} \tag{7.3}$$

Caution: When computing z for a single score, use the standard deviation, σ. When computing z for a sample mean, you must use the standard error, σ_M (see Box 7.2).

Just as every score (X) has a z-score that describes its position in the distribution of scores, every sample mean (M) has a z-score that describes its position in the distribution of sample means. When the distribution of sample means is normal, it is possible to use z-scores and the unit normal table to find the probability associated with any specific sample mean (as in Example 7.2). The following example demonstrates that it also is possible to make quantitative predictions about the kinds of samples that should be obtained from any population.

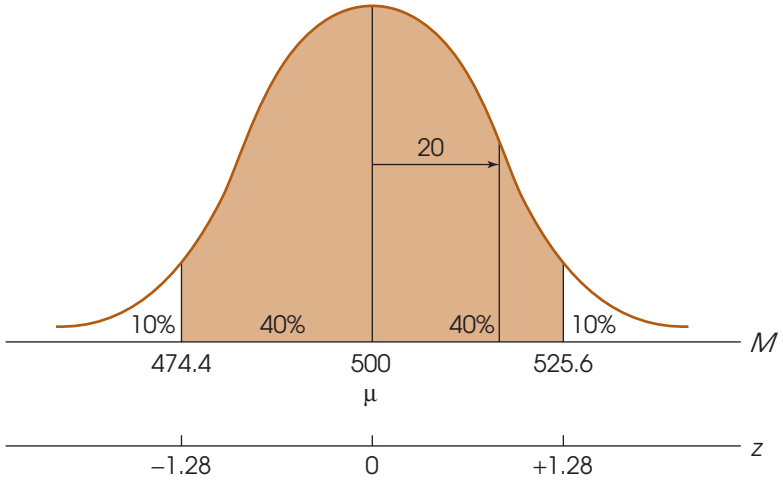
EXAMPLE 7.3

Once again, the distribution of SAT scores forms a normal distribution with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. For this example, we are going to determine what kind of sample mean is likely to be obtained as the average SAT score for a random sample of $n = 25$ students. Specifically, we determine the exact range of values that is expected for the sample mean 80% of the time.

We begin with the distribution of sample means for $n = 25$. As demonstrated in Example 7.2, this distribution is normal with an expected value of $\mu = 500$ and a standard error of $\sigma_M = 20$ (Figure 7.6). Our goal is to find the range of values that make up the middle 80% of the distribution. Because the distribution is normal, we can use the unit normal table. First, the 80% in the middle is split in half, with 40% (0.4000) on each side of the mean. Looking up 0.4000 in column D (the proportion between the mean and z), we find a corresponding z-score of $z = 1.28$. Thus, the z-score boundaries for the middle 80% are $z = +1.28$ and $z = -1.28$. By definition, a z-score of 1.28 represents a location that is 1.28 standard deviations (or standard

FIGURE 7.6

The middle 80% of the distribution of sample means for $n = 25$. Sample were selected from a normal population with $\mu = 500$ and $\sigma = 100$.



BOX 7.2 THE DIFFERENCE BETWEEN STANDARD DEVIATION AND STANDARD ERROR

A constant source of confusion for many students is the difference between standard deviation and standard error. Remember that standard deviation measures the standard distance between a *score* and the population mean, $X - \mu$. If you are working with a distribution of scores, the standard deviation is the appropriate measure of variability. Standard error, on the other hand, measures the standard distance between a *sample mean* and the population mean, $M - \mu$. Whenever you have a question concerning a sample, the standard error is the appropriate measure of variability.

If you still find the distinction confusing, there is a simple solution. Namely, if you always use standard error, you always will be right. Consider the formula for standard error:

$$\text{standard error} = \sigma_M = \frac{\sigma}{\sqrt{n}}$$

If you are working with a single score, then $n = 1$, and the standard error becomes

$$\text{standard error} = \sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{1}}$$

$$= \sigma = \text{standard deviation}$$

Thus, standard error always measures the standard distance from the population mean for any sample size, including $n = 1$.

errors) from the mean. With a standard error of 20 points, the distance from the mean is 1.28(20) = 25.6 points. The mean is $\mu = 500$, so a distance of 25.6 in both directions produces a range of values from 474.4 to 525.6.

Thus, 80% of all the possible sample means are contained in a range between 474.4 and 525.6. If we select a sample of $n = 25$ students, we can be 80% confident that the mean SAT score for the sample will be in this range.

The point of Example 7.3 is that the distribution of sample means makes it possible to predict the value that ought to be obtained for a sample mean. We know, for example, that a sample of $n = 25$ students ought to have a mean SAT score around 500. More specifically, we are 80% confident that the value of the sample mean will be between 474.4 and 525.6. The ability to predict sample means in this way is a valuable tool for the inferential statistics that follow.

LEARNING CHECK

1. For a population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 8$, find the z-score corresponding to a sample mean of $M = 44$ for each of the following sample sizes.
   - **a.** $n = 4$
   - **b.** $n = 16$
2. What is the probability of obtaining a sample mean greater than $M = 60$ for a random sample of $n = 16$ scores selected from a normal population with a mean of $\mu = 65$ and a standard deviation of $\sigma = 20$?
3. A positively skewed distribution has $\mu = 60$ and $\sigma = 8$.
   - **a.** What is the probability of obtaining a sample mean greater than $M = 62$ for a sample of $n = 4$ scores? (Be careful. This is a trick question.)

b. What is the probability of obtaining a sample mean greater than $M = 62$ for a sample of $n = 64$ scores?

ANSWERS

1. a. The standard error is $\sigma_M = 4$, and $z = 1.00$.

b. The standard error is $\sigma_M = 2$, and $z = 2.00$.

2. The standard error is $\sigma_M = 5$, and $M = 60$ corresponds to $z = -1.00$, $p(M > 60) = p(z > -1.00) = 0.8413$ (or 84.13%).

3. a. The distribution of sample means does not satisfy either of the criteria for being normal. Therefore, you cannot use the unit normal table, and it is impossible to find the probability.

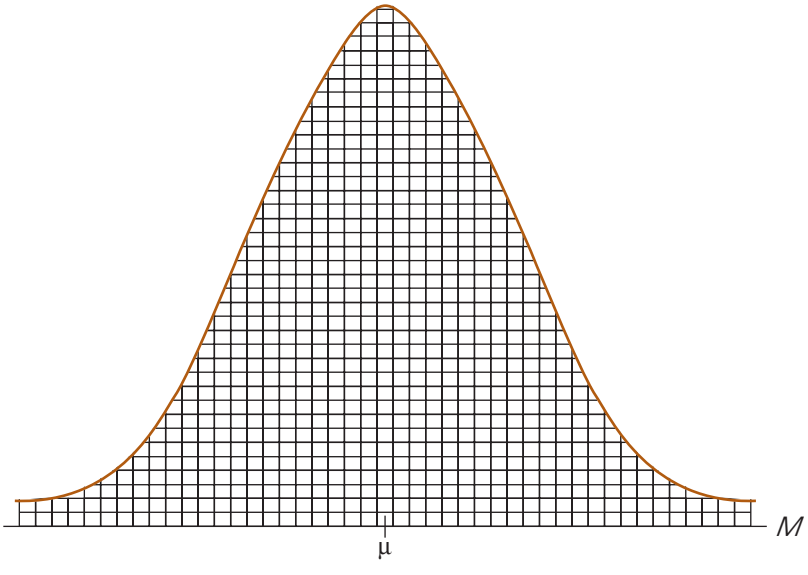
b. With $n = 64$, the distribution of sample means is nearly normal. The standard error is $8/\sqrt{64} = 1$, the z-score is $+2.00$, and the probability is 0.0228.

7.4 MORE ABOUT STANDARD ERROR

At the beginning of this chapter, we introduced the idea that it is possible to obtain thousands of different samples from a single population. Each sample has its own individuals, its own scores, and its own sample mean. The distribution of sample means provides a method for organizing all of the different sample means into a single picture. Figure 7.7 shows a prototypical distribution of sample means. To emphasize the fact that the distribution contains many different samples, we have constructed this figure so that the distribution is made up of hundreds of small boxes, each box representing a single sample mean. Also notice that the sample means tend to pile up around the population mean (μ), forming a normal-shaped distribution as predicted by the central limit theorem.

FIGURE 7.7

An example of a typical distribution of sample means. Each of the small boxes represents the mean obtained for one sample.



The distribution shown in Figure 7.7 provides a concrete example for reviewing the general concepts of sampling error and standard error. Although the following points may seem obvious, they are intended to provide you with a better understanding of these two statistical concepts.

1. Sampling Error. The general concept of sampling error is that a sample typically does not provide a perfectly accurate representation of its population. More specifically, there typically is some discrepancy (or error) between a statistic computed for a sample and the corresponding parameter for the population. As you look at Figure 7.7, notice that the individual sample means are not exactly equal to the population mean. In fact, 50% of the samples have means that are smaller than μ (the entire left-hand side of the distribution). Similarly, 50% of the samples produce means that overestimate the true population mean. In general, there is some discrepancy, or *sampling error,* between the mean for a sample and the mean for the population from which the sample was obtained.
2. Standard Error. Again looking at Figure 7.7, notice that most of the sample means are relatively close to the population mean (those in the center of the distribution). These samples provide a fairly accurate representation of the population. On the other hand, some samples produce means that are out in the tails of the distribution, relatively far from the population mean. These extreme sample means do not accurately represent the population. For each individual sample, you can measure the error (or distance) between the sample mean and the population mean. For some samples, the error is relatively small, but for other samples, the error is relatively large. The *standard error* provides a way to measure the "average," or standard, distance between a sample mean and the population mean.

Thus, the standard error provides a method for defining and measuring sampling error. Knowing the standard error gives researchers a good indication of how accurately their sample data represent the populations that they are studying. In most research situations, for example, the population mean is unknown, and the researcher selects a sample to help obtain information about the unknown population. Specifically, the sample mean provides information about the value of the unknown population mean. The sample mean is not expected to give a perfectly accurate representation of the population mean; there will be some error, and the standard error tells *exactly how much error,* on average, should exist between the sample mean and the unknown population mean. The following example demonstrates the use of standard error and provides additional information about the relationship between standard error and standard deviation.

EXAMPLE 7.4

A recent survey of students at a local college included the following question: How many minutes do you spend each day watching electronic video (e.g., online, TV, cell phone, iPod, etc.). The average response was $\mu = 80$ minutes, and the distribution of viewing times was approximately normal with a standard deviation of $\sigma = 20$ minutes. Next, we take a sample from this population and examine how accurately the sample mean represents the population mean. More specifically, we will examine how sample size affects accuracy by considering three different samples: one with $n = 1$ student, one with $n = 4$ students, and one with $n = 100$ students.

Figure 7.8 shows the distributions of sample means based on samples of $n = 1$, $n = 4$, and $n = 100$. Each distribution shows the collection of all possible sample means that could be obtained for that particular sample size. Notice that all three sampling distributions are normal (because the original population is normal), and

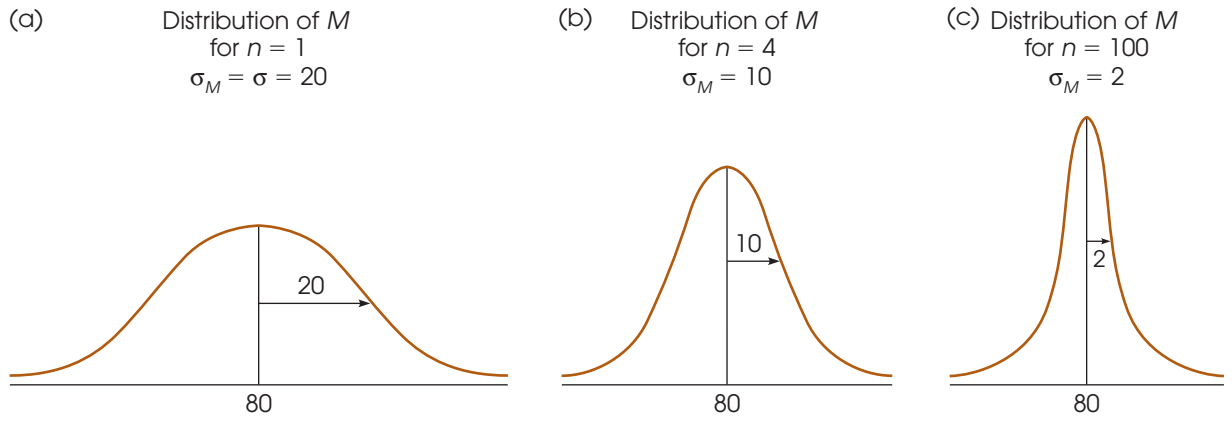


FIGURE 7.8

The distribution of sample means for random samples of size (a) $n = 1$, (b) $n = 4$, and (c) $n = 100$ obtained from a normal population with ($\mu = 80$) and ($\sigma = 20$). Notice that the size of the standard error decreases as the sample size increases.

all three have the same mean, $\mu = 80$, which is the expected value of M. However, the three distributions differ greatly with respect to variability. We will consider each one separately.

The smallest sample size is $n = 1$. When a sample consists of a single student, the mean for the sample equals the score for the student, $M = X$. Thus, when $n = 1$, the distribution of sample means is identical to the original population of scores. In this case, the standard error for the distribution of sample means is equal to the standard deviation for the original population. Equation 7.1 confirms this observation.

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{1}} = 20$$

When the sample consists of a single student, you expect, on average, a 20-point difference between the sample mean and the mean for the population. As we noted earlier, the population standard deviation is the "starting point" for the standard error. With the smallest possible sample, $n = 1$, the standard error is equal to the standard deviation [see Figure 7.8(a)].

As the sample size increases, however, the standard error gets smaller. For a sample of $n = 4$ students, the standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{4}} = \frac{20}{2} = 10$$

That is, the typical (or standard) distance between M and μ is 10 points. Figure 7.8(b) illustrates this distribution. Notice that the sample means in this distribution approximate the population mean more closely than in the previous distribution where $n = 1$.

With a sample of $n = 100$, the standard error is still smaller.

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = \frac{20}{10} = 2$$

A sample of $n = 100$ students should produce a sample mean that represents the population much more accurately than a sample of $n = 4$ or $n = 1$. As shown in Figure 7.8(c), there is very little error between M and μ when $n = 100$. Specifically, you would expect, on average, only a 2-point difference between the population mean and the sample mean.

In summary, this example illustrates that with the smallest possible sample ($n = 1$), the standard error and the population standard deviation are the same. When sample size is increased, the standard error gets smaller, and the sample means tend to approximate μ more closely. Thus, standard error defines the relationship between sample size and the accuracy with which M represents μ.

IN THE LITERATURE

REPORTING STANDARD ERROR

As we will see later, standard error plays a very important role in inferential statistics. Because of its crucial role, the standard error for a sample mean, rather than the sample standard deviation, is often reported in scientific papers. Scientific journals vary in how they refer to the standard error, but frequently the symbols *SE* and *SEM* (for standard error of the mean) are used. The standard error is reported in two ways. Much like the standard deviation, it may be reported in a table along with the sample means (Table 7.3). Alternatively, the standard error may be reported in graphs.

Figure 7.9 illustrates the use of a bar graph to display information about the sample mean and the standard error. In this experiment, two samples (groups A

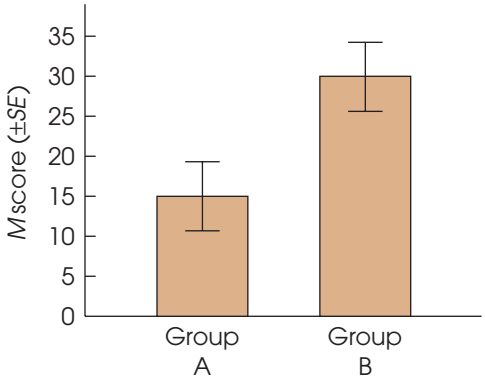
TABLE 7.3

The mean self-consciousness scores for participants who were working in front of a video camera and those who were not (controls).

	n	Mean	*SE*
Control	17	32.23	2.31
Camera	15	45.17	2.78

FIGURE 7.9

The mean (±*SE*) score for treatment groups A and B.

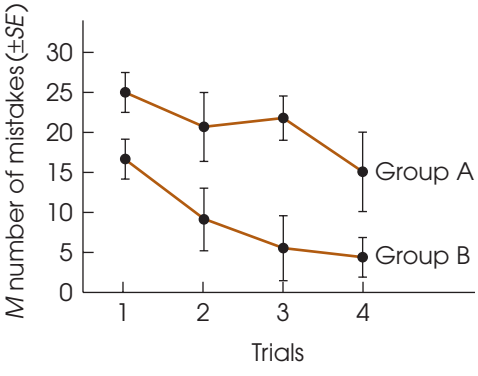


and B) are given different treatments, and then the subjects' scores on a dependent variable are recorded. The mean for group A is $M = 15$, and for group B, it is $M = 30$. For both samples, the standard error of M is $\sigma_M = 4$. Note that the mean is represented by the height of the bar, and the standard error is depicted by brackets at the top of each bar. Each bracket extends 1 standard error above and 1 standard error below the sample mean. Thus, the graph illustrates the mean for each group plus or minus 1 standard error ($M \pm SE$). When you glance at Figure 7.9, not only do you get a "picture" of the sample means, but also you get an idea of how much error you should expect for those means.

Figure 7.10 shows how sample means and standard error are displayed in a line graph. In this study, two samples representing different age groups are tested on a task for four trials. The number of errors committed on each trial is recorded for all participants. The graph shows the mean (M) number of errors committed for each group on each trial. The brackets show the size of the standard error for each sample mean. Again, the brackets extend 1 standard error above and below the value of the mean.

FIGURE 7.10

The mean ($\pm SE$) number of mistakes made for groups A and B on each trial.



LEARNING CHECK

1. A population has a standard deviation of $\sigma = 10$.
   a. On average, how much difference should there be between the population mean and a single score selected from this population?
   b. On average, how much difference should there be between the population mean and the sample mean for $n = 4$ scores selected from this population?
   c. On average, how much difference should there be between the population mean and the sample mean for $n = 25$ scores selected from this population?
2. Can the value of the standard error ever be larger than the value of the population standard deviation? Explain your answer.
3. A researcher plans to select a random sample from a population with a standard deviation of $\sigma = 12$.
   a. How large a sample is needed to have a standard error of 6 points or less?
   b. How large a sample is needed to have a standard error of 4 points or less?

ANSWERS

1. **a.** $\sigma = 10$ points
   b. $\sigma_M = 5$ points
   c. $\sigma_M = 2$ points
2. No. The standard error is computed by dividing the standard deviation by the square root of n. The standard error is always less than or equal to the standard deviation.
3. **a.** A sample of $n = 4$ or larger.
   b. A sample of $n = 9$ or larger.

7.5 LOOKING AHEAD TO INFERENTIAL STATISTICS

Inferential statistics are methods that use sample data as the basis for drawing general conclusions about populations. However, we have noted that a sample is not expected to give a perfectly accurate reflection of its population. In particular, there will be some error or discrepancy between a sample statistic and the corresponding population parameter. In this chapter, we have observed that a sample mean is not exactly equal to the population mean. The standard error of M specifies how much difference is expected on average between the mean for a sample and the mean for the population.

The natural differences that exist between samples and populations introduce a degree of uncertainty and error into all inferential processes. Specifically, there is always a margin of error that must be considered whenever a researcher uses a sample mean as the basis for drawing a conclusion about a population mean. Remember that the sample mean is not perfect. In the next seven chapters we introduce a variety of statistical methods that all use sample means to draw inferences about population means.

In each case, the distribution of sample means and the standard error are critical elements in the inferential process. Before we begin this series of chapters, we pause briefly to demonstrate how the distribution of sample means, along with z-scores and probability, can help us use sample means to draw inferences about population means.

EXAMPLE 7.5

Suppose that a psychologist is planning a research study to evaluate the effect of a new growth hormone. It is known that regular adult rats (with no hormone) weigh an average of $\mu = 400$ grams. Of course, not all rats are the same size, and the distribution of their weights is normal with $\sigma = 20$. The psychologist plans to select a sample of $n = 25$ newborn rats, inject them with the hormone, and then measure their weights when they become adults. The structure of this research study is shown in Figure 7.11.

The psychologist makes a decision about the effect of the hormone by comparing the sample of treated rats with the regular untreated rats in the original population. If the treated rats in the sample are noticeably different from untreated rats, then the researcher has evidence that the hormone has an effect. The problem is to determine exactly how much difference is necessary before we can say that the sample is *noticeably different.*

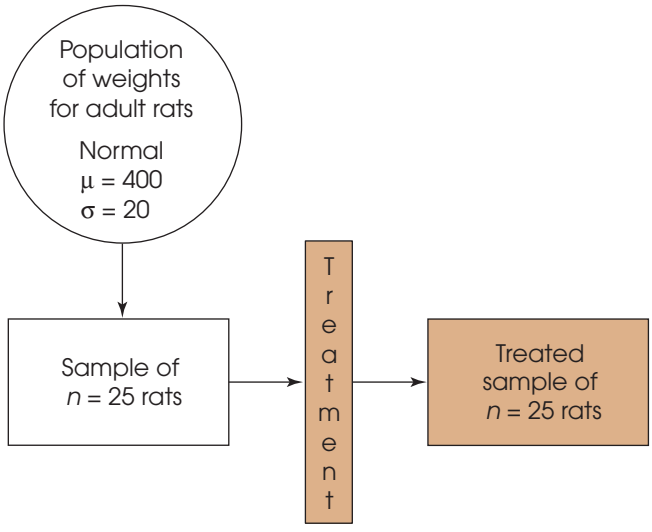
The distribution of sample means and the standard error can help researchers make this decision. In particular, the distribution of sample means can be used to show exactly what would be expected for a sample of rats who do not receive any hormone injections. This allows researchers to make a simple comparison between

a. The sample of treated rats (from the research study)

b. Samples of untreated rats (from the distribution of sample means)

FIGURE 7.11

The structure of the research study described in Example 7.5. The purpose of the study is to determine whether the treatment (a growth hormone) has an effect on weight for rats.



If our treated sample is noticeably different from the untreated samples, then we have evidence that the treatment has an effect. On the other hand, if our treated sample still looks like one of the untreated samples, then we must conclude that the treatment does not appear to have any effect.

We begin with the original population of untreated rats and consider the distribution of sample means for all of the possible samples of $n = 25$ rats. The distribution of sample means has the following characteristics:

1. It is a normal distribution, because the population of rat weights is normal.
2. It has an expected value of 400, because the population mean for untreated rats is $\mu = 400$.
3. It has a standard error of $\sigma_M = \dfrac{20}{\sqrt{25}} = \dfrac{20}{5} = 4$, because the population standard deviation is $\sigma = 20$ and the sample size is $n = 25$.

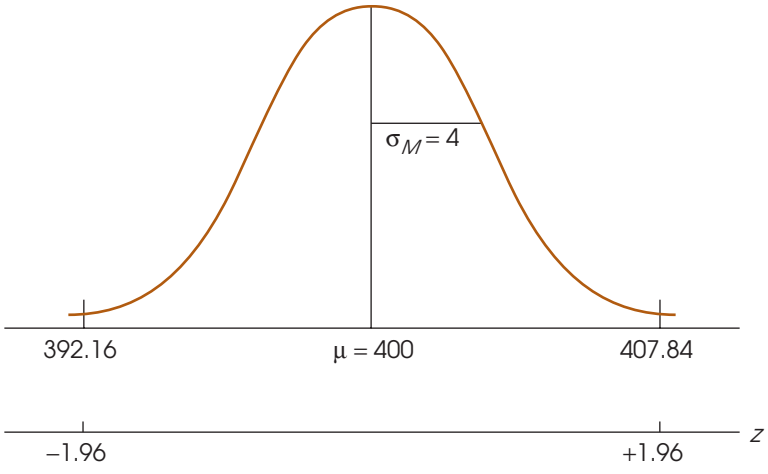
The distribution of sample means is shown in Figure 7.12. Notice that a sample of $n = 25$ untreated rats (without the hormone) should have a mean weight around 400 grams. To be more precise, we can use z-scores to determine the middle 95% of all the possible sample means. As demonstrated in Chapter 6 (p. 190), the middle 95% of a normal distribution is located between z-score boundaries of $z = +1.96$ and $z = -1.96$ (check the unit normal table). These z-score boundaries are shown in Figure 7.12. With a standard error of $\sigma_M = 4$ points, a z-score of $z = 1.96$ corresponds to a distance of $1.96(4) = 7.84$ points from the mean. Thus, the z-score boundaries of ± 1.96 correspond to sample means of 392.16 and 407.84.

We have demonstrated that a sample of untreated rats is almost guaranteed (95% probability) to have a sample mean between 392.16 and 407.84. If our sample has a mean within this range, then we must conclude that our sample of treated rats is not noticeably different from samples of untreated rats. In this case, we conclude that the treatment does not appear to have any effect.

On the other hand, if the mean for the treated sample is outside the 95% range, then we can conclude that our sample of treated rats is noticeably different from the

FIGURE 7.12

The distribution of sample means for samples of $n = 25$ untreated rats (from Example 7.5).



samples that would be obtained without any treatment. In this case, the research results provide evidence that the treatment has an effect.

In Example 7.5 we used the distribution of sample means, together with *z*-scores and probability, to provide a description of what is reasonable to expect for an untreated sample. Then, we evaluated the effect of a treatment by determining whether the treated sample was noticeably different from an untreated sample. This procedure forms the foundation for the inferential technique known as *hypothesis testing,* which is introduced in Chapter 8 and repeated throughout the remainder of this book.

STANDARD ERROR AS A MEASURE OF RELIABILITY

The research situation shown in Figure 7.11 introduces one final issue concerning sample means and standard error. In Figure 7.11, as in most research studies, the researcher must rely on a *single* sample to provide an accurate representation of the population being investigated. As we have noted, however, if you take two different samples from the same population, the samples will have different individuals with different scores and different sample means. Thus, every researcher must face the nagging question, "If I had taken a different sample, would I have obtained different results?"

The importance of this question is directly related to the degree of similarity among all the different samples. For example, if there is a high level of consistency from one sample to another, then a researcher can be reasonably confident that the specific sample being studied provides a good measurement of the population. That is, when all of the samples are similar, then it does not matter which one you have selected. On the other hand, if there are big differences from one sample to another, then the researcher is left with some doubts about the accuracy of his or her specific sample. In this case, a different sample could have produced vastly different results.

The relationship between the number of scores in the sample and the size of the standard error is shown in Figure 7.3 on page 208.

In this context, the standard error can be viewed as a measure of the *reliability* of a sample mean. The term *reliability* refers to the consistency of different measurements of the same thing. More specifically, a measurement procedure is said to be reliable if you make two different measurements of the same thing and obtain identical (or nearly identical) values. If you view a sample as a "measurement" of a population, then a sample mean is a "measurement" of the population mean.

If the standard error is small, then all of the possible sample means are clustered close together and a researcher can be confident that any individual sample mean provides a reliable measure of the population. On the other hand, a large standard error indicates that there are relatively large differences from one sample mean to another, and a researcher must be concerned that a different sample could produce a different conclusion. Fortunately, the size of the standard error can be controlled. In particular, if a researcher is concerned about a large standard error and the potential for big differences from one sample to another, then the researcher has the option of reducing the standard error by selecting a larger sample. Thus, the ability to compute the value of the standard error provides researchers with the ability to control the reliability of their samples.

The reliability of a sample mean is directly related to the degree of confidence that a specific sample mean is a stable and accurate representative of the population. If a researcher suspects that adding one or two new scores to a sample might produce a substantial change in the sample mean, then the sample is not reliable and the researcher has no confidence that it is stable and accurate. There are two factors that influence whether a few new scores might substantially change a sample mean.

1. The number of scores in the sample. If there are only 2 or 3 scores in a sample, then a few new scores can have a huge influence on the sample mean. On the other hand, if a sample already has 100 scores, then one or two new ones cannot have much effect.
2. The size of the population standard deviation. When the standard deviation is large, it means that the scores are spread over a wide range of values. In this situation it is possible to select one or two extreme scores that are very different from the others. As we noted in Chapter 3 (p. 90), adding one or two extreme scores to a sample can have a large influence on the sample mean. With a small standard deviation, however, all of the scores are close together and a few new scores should be similar to the ones already in the sample.

Notice that these two factors are the same values that are used to calculate the standard error. A large sample means that the standard error is small and the sample mean is reliable. Also, a small population standard deviation means that the standard error is small and the sample mean is reliable. In either case, a researcher can be confident that adding a few new scores to an existing sample will not have a significant influence on the sample mean.

LEARNING CHECK

1. A population forms a normal distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$.
   - **a.** If single score is selected from this population, how much distance would you expect, on average, between the score and the population mean?
   - **b.** If a sample of $n = 100$ scores is selected from this population, how much distance would you expect, on average, between the sample mean and the population mean?
2. A population forms a normal shaped distribution with $\mu = 40$ and $\sigma = 8$.
   - **a.** A sample of $n = 16$ scores from this population has a mean of $M = 36$. Would you describe this as a relatively typical sample, or is the sample mean an extreme value? Explain your answer.
   - **b.** If the sample from part a had $n = 4$ scores, would it be considered typical or extreme?

3. The SAT scores for the entering freshman class at a local college form a normal distribution with a mean of $\mu = 530$ and a standard deviation of $\sigma = 80$.

a. For a random sample of $n = 16$ students, what range of values for the sample mean would be expected 95% of the time?

b. What range of values would be expected 95% of the time if the sample size were $n = 100$?

4. An automobile manufacturer claims that a new model will average $\mu = 45$ miles per gallon with $\sigma = 4$. A sample of $n = 16$ cars is tested and averages only $M = 43$ miles per gallon. Is this sample mean likely to occur if the manufacturer's claim is true? Specifically, is the sample mean within the range of values that would be expected 95% of the time? (Assume that the distribution of mileage scores is normal.)

ANSWERS

1. a. For a single score, the standard distance from the mean is the standard deviation, $\sigma = 20$.

b. For a sample of $n = 100$ scores, the average distance between the sample mean and the population mean is the standard error, $\sigma_M = 20/\sqrt{100} = 2$.

2. a. With $n = 16$ the standard error is 2, and the sample mean corresponds to $z = -2.00$. This is an extreme value.

b. With $n = 4$ the standard error is 4, and the sample mean corresponds to $z = -1.00$. This is a relatively typical value.

3. a. With $n = 16$ the standard error is $\sigma_M = 20$ points. Using $z = \pm 1.96$, the 95% range extends from 490.8 to 569.2.

b. With $n = 100$ the standard error is only 8 points and the range extends from 514.32 to 545.68.

4. With $n = 16$, the standard error is $\sigma_M = 1$. If the real mean is $\mu = 45$, then 95% of all sample means should be within $1.96(1) = 1.96$ points of $\mu = 45$. This is a range of values from 43.04 to 46.96. Our sample mean is outside this range, so it is not the kind of sample that ought to be obtained if the manufacturer's claim is true.

SUMMARY

1. The distribution of sample means is defined as the set of *M*s for all the possible random samples for a specific sample size (n) that can be obtained from a given population. According to the *central limit theorem,* the parameters of the distribution of sample means are as follows:

a. *Shape*. The distribution of sample means is normal if either one of the following two conditions is satisfied:

(1) The population from which the samples are selected is normal.

(2) The size of the samples is relatively large ($n = 30$ or more).

b. *Central Tendency*. The mean of the distribution of sample means is identical to the mean of the population from which the samples are selected. The mean of the distribution of sample means is called the expected value of *M*.

c. *Variability*. The standard deviation of the distribution of sample means is called the standard error of *M* and is defined by the formula

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \sigma_M = \sqrt{\frac{\sigma^2}{n}}$$

Standard error measures the standard distance between a sample mean (M) and the population mean (μ).

2. One of the most important concepts in this chapter is standard error. The standard error is the standard

deviation of the distribution of sample means. It measures the standard distance between a sample mean (M) and the population mean (μ). The standard error tells how much error to expect if you are using a sample mean to represent a population mean.

3. The location of each M in the distribution of sample means can be specified by a z-score:

$$z = \frac{M - \mu}{\sigma_M}$$

Because the distribution of sample means tends to be normal, we can use these z-scores and the unit normal table to find probabilities for specific sample means. In particular, we can identify which sample means are likely and which are very unlikely to be obtained from any given population. This ability to find probabilities for samples is the basis for the inferential statistics in the chapters ahead.

4. In general terms, the standard error measures how much discrepancy you should expect between a sample statistic and a population parameter. Statistical inference involves using sample statistics to make a general conclusion about a population parameter. Thus, standard error plays a crucial role in inferential statistics.

KEY TERMS

sampling error (201)
distribution of sample means (201)
sampling distribution (202)
central limit theorem (205)
expected value of M (206)
standard error of M (207)
law of large numbers (207)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 7 on the book companion website. The website also provides access to two workshops entitled *Standard Error* and *Central Limit Theorem* that review the material covered in Chapter 7.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGE brain com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

The statistical computer package SPSS is not structured to compute the standard error or a *z*-score for a sample mean. In later chapters, however, we introduce new inferential statistics that are included in SPSS. When these new statistics are computed, SPSS typically includes a report of standard error that describes how accurately, on average, the sample represents its population.

FOCUS ON PROBLEM SOLVING

1. Whenever you are working probability questions about sample means, you must use the distribution of sample means. Remember that every probability question can be restated as a proportion question. Probabilities for sample means are equivalent to proportions of the distribution of sample means.
2. When computing probabilities for sample means, the most common error is to use standard deviation (σ) instead of standard error (σ_M) in the *z*-score formula. Standard deviation measures the typical deviation (or error) for a single score. Standard error measures the typical deviation (or error) for a sample. Remember: The larger the sample is, the more accurately the sample represents the population. Thus, sample size (n) is a critical part of the standard error.

$$\text{Standard error} = \sigma_M = \frac{\sigma}{\sqrt{n}}$$

Although the distribution of sample means is often normal, it is not always a normal distribution. Check the criteria to be certain that the distribution is normal before you use the unit normal table to find probabilities (see item 1a of the Summary). Remember that all probability problems with a normal distribution are easier to solve if you sketch the distribution and shade in the area of interest.

DEMONSTRATION 7.1

PROBABILITY AND THE DISTRIBUTION OF SAMPLE MEANS

A population forms a normal distribution with a mean of $\mu = 60$ and a standard deviation of $\sigma = 12$. For a sample of $n = 36$ scores from this population, what is the probability of obtaining a sample mean greater than 64?

$$p(M > 64) = ?$$

STEP 1 **Rephrase the probability question as a proportion question.** Out of all of the possible sample means for $n = 36$, what proportion has values greater than 64? *All of the possible sample means* is simply the distribution of sample means, which is normal, with a mean of $\mu = 60$ and a standard error of

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = \frac{12}{6} = 2$$

The distribution is shown in Figure 7.13(a). Because the problem is asking for the proportion greater than $M = 64$, this portion of the distribution is shaded in Figure 7.13(b).

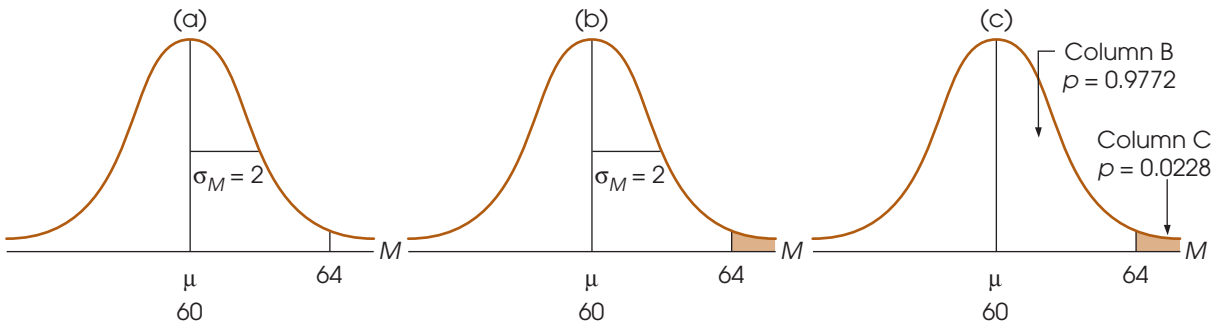


FIGURE 7.13
Sketches of the distributions for Demonstration 7.1.

STEP 2 **Compute the *z*-score for the sample mean.** A sample mean of $M = 64$ corresponds to a *z*-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{64 - 60}{2} = \frac{4}{2} = 2.00$$

Therefore, $p(M > 64) = p(z > 2.00)$

STEP 3 **Look up the proportion in the unit normal table.** Find $z = 2.00$ in column A and read across the row to find $p = 0.0228$ in column C. This is the answer as shown in Figure 7.13(c).

$$p(M > 64) = p(z > 2.00) = 0.0228 \text{ (or } 2.28\%)$$

PROBLEMS

1. Briefly define each of the following:
   a. Distribution of sample means
   b. Expected value of M
   c. Standard error of M

2. Describe the distribution of sample means (shape, expected value, and standard error) for samples of $n = 36$ selected from a population with a mean of $\mu = 100$ and a standard deviation of $\sigma = 12$.

3. A sample is selected from a population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 8$.
   a. If the sample has $n = 4$ scores, what is the expected value of M and the standard error of M?
   b. If the sample has $n = 16$ scores, what is the expected value of M and the standard error of M?

4. The distribution of sample means is not always a normal distribution. Under what circumstances is the distribution of sample means *not* normal?

5. A population has a standard deviation of $\sigma = 30$.
   a. On average, how much difference should exist between the population mean and the sample mean for $n = 4$ scores randomly selected from the population?
   b. On average, how much difference should exist for a sample of $n = 25$ scores?
   c. On average, how much difference should exist for a sample of $n = 100$ scores?

6. For a population with a mean of $\mu = 70$ and a standard deviation of $\sigma = 20$, how much error, on average, would you expect between the sample mean (M) and the population mean for each of the following sample sizes?
   a. $n = 4$ scores
   b. $n = 16$ scores
   c. $n = 25$ scores

7. For a population with a standard deviation of $\sigma = 20$, how large a sample is necessary to have a standard error that is:
- **a.** less than or equal to 5 points?
- **b.** less than or equal to 2 points?
- **c.** less than or equal to 1 point?

8. If the population standard deviation is $\sigma = 8$, how large a sample is necessary to have a standard error that is:
- **a.** less than 4 points?
- **b.** less than 2 points?
- **c.** less than 1 point?

9. For a sample of $n = 25$ scores, what is the value of the population standard deviation (σ) necessary to produce each of the following a standard error values?
- **a.** $\sigma_M = 10$ points?
- **b.** $\sigma_M = 5$ points?
- **c.** $\sigma_M = 2$ points?

10. For a population with a mean of $\mu = 80$ and a standard deviation of $\sigma = 12$, find the z-score corresponding to each of the following samples.
- **a.** $M = 83$ for a sample of $n = 4$ scores
- **b.** $M = 83$ for a sample of $n = 16$ scores
- **c.** $M = 83$ for a sample of $n = 36$ scores

11. A sample of $n = 4$ scores has a mean of $M = 75$. Find the z-score for this sample:
- **a.** If it was obtained from a population with $\mu = 80$ and $\sigma = 10$.
- **b.** If it was obtained from a population with $\mu = 80$ and $\sigma = 20$.
- **c.** If it was obtained from a population with $\mu = 80$ and $\sigma = 40$.

12. A population forms a normal distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 15$. For each of the following samples, compute the z-score for the sample mean and determine whether the sample mean is a typical, representative value or an extreme value for a sample of this size.
- **a.** $M = 84$ for $n = 9$ scores
- **b.** $M = 84$ for $n = 100$ scores

13. A random sample is obtained from a normal population with a mean of $\mu = 30$ and a standard deviation of $\sigma = 8$. The sample mean is $M = 33$.
- **a.** Is this a fairly typical sample mean or an extreme value for a sample of $n = 4$ scores?
- **b.** Is this a fairly typical sample mean or an extreme value for a sample of $n = 64$ scores?

14. The population of IQ scores forms a normal distribution with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. What is the probability of obtaining a sample mean greater than $M = 97$,
- **a.** for a random sample of $n = 9$ people?
- **b.** for a random sample of $n = 25$ people?

15. The scores on a standardized mathematics test for 8th-grade children in New York State form a normal distribution with a mean of $\mu = 70$ and a standard deviation of $\sigma = 10$.
- **a.** What proportion of the students in the state have scores less than $X = 75$?
- **b.** If samples of $n = 4$ are selected from the population, what proportion of the samples will have means less than $M = 75$?
- **c.** If samples of $n = 25$ are selected from the population, what proportion of the samples will have means less than $M = 75$?

16. A population of scores forms a normal distribution with a mean of $\mu = 40$ and a standard deviation of $\sigma = 12$.
- **a.** What is the probability of randomly selecting a score less than $X = 34$?
- **b.** What is the probability of selecting a sample of $n = 9$ scores with a mean less than $M = 34$?
- **c.** What is the probability of selecting a sample of $n = 36$ scores with a mean less than $M = 34$?

17. A population of scores forms a normal distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 10$.
- **a.** What proportion of the scores have values between 75 and 85?
- **b.** For samples of $n = 4$, what proportion of the samples will have means between 75 and 85?
- **c.** For samples of $n = 16$, what proportion of the samples will have means between 75 and 85?

18. At the end of the spring semester, the Dean of Students sent a survey to the entire freshman class. One question asked the students how much weight they had gained or lost since the beginning of the school year. The average was a gain of $\mu = 9$ pounds with a standard deviation of $\sigma = 6$. The distribution of scores was approximately normal. A sample of $n = 4$ students is selected and the average weight change is computed for the sample.
- **a.** What is the probability that the sample mean will be greater than $M = 10$ pounds? In symbols, what is $p(M > 10)$?
- **b.** Of all of the possible samples, what proportion will show an average weight loss? In symbols, what is $p(M < 0)$?
- **c.** What is the probability that the sample mean will be a gain of between $M = 9$ and $M = 12$ pounds? In symbols, what is $p(9 < M < 12)$?

19. The machinery at a food-packing plant is able to put exactly 12 ounces of juice in every bottle. However, some items such as apples come in variable sizes so it is almost impossible to get exactly 3 pounds of apples in a bag labeled "3 lbs." Therefore, the machinery is set to put an average of $\mu = 50$ ounces (3 pounds and 2 ounces) in each bag. The distribution of bag weights is approximately normal with a standard deviation of $\sigma = 4$ ounces.

a. What is the probability of randomly picking a bag of apples that weighs less than 48 ounces (3 pounds)?

b. What is the probability of randomly picking $n = 4$ bags of apples that have an average weight less than $M = 48$ ounces?

20. The average age for licensed drivers in the county is $\mu = 40.3$ years with a standard deviation of $\sigma = 13.2$ years.

a. A researcher obtained a random sample of $n = 16$ parking tickets and computed an average age of $M = 38.9$ years for the drivers. Compute the z-score for the sample mean and find the probability of obtaining an average age this young or younger for a random sample of licensed drivers. Is it reasonable to conclude that this set of $n = 16$ people is a representative sample of licensed drivers?

b. The same researcher obtained a random sample of $n = 36$ speeding tickets and computed an average age of $M = 36.2$ years for the drivers. Compute the z-score for the sample mean and find the probability of obtaining an average age this young or younger for a random sample of licensed drivers. Is it reasonable to conclude that this set of $n = 36$ people is a representative sample of licensed drivers?

21. People are selected to serve on juries by randomly picking names from the list of registered voters. The average age for registered voters in the county is $\mu = 44.3$ years with a standard deviation of $\sigma = 12.4$. A statistician computes the average age for a group of $n = 12$ people currently serving on a jury and obtains a mean of $M = 48.9$ years.

a. How likely is it to obtain a random sample of $n = 12$ jurors with an average age equal to or greater than 48.9?

b. Is it reasonable to conclude that this set of $n = 12$ people is not a representative random sample of registered voters?

22. Welsh, Davis, Burke, and Williams (2002) conducted a study to evaluate the effectiveness of a carbohydrate-electrolyte drink on sports performance and endurance. Experienced athletes were given either a carbohydrate-electrolyte drink or a placebo while they were tested on a series of high-intensity exercises. One measure was how much time it took for the athletes to run to fatigue. Data similar to the results obtained in the study are shown in the following table.

Time to Run to Fatigue (in minutes)

	Mean	*SE*
Placebo	21.7	2.2
Carbohydrate-electrolyte	28.6	2.7

a. Construct a bar graph that incorporates all of the information in the table.

b. Looking at your graph, do you think that the carbohydrate-electrolyte drink helps performance?

23. In the Preview section for this chapter, we discussed a research study demonstrating that 8-month-old infants appear to recognize which samples are likely to be obtained from a population and which are not. In the study, the infants watched as a sample of $n = 5$ ping pong balls was selected from a large box. In one condition, the sample consisted of 1 red ball and 4 white balls. After the sample was selected, the front panel of the box was removed to reveal the contents. In the *expected* condition, the box contained primarily white balls like the sample and the infants looked at it for an average of $M = 7.5$ seconds. In the *unexpected* condition, the box had primarily red balls, unlike the sample, and the infants looked at it for $M = 9.9$ seconds. Assuming that the standard error for both means is ($\sigma_M = 1$ second, draw a bar graph showing the two sample means using brackets to show the size of the standard error for each mean.

CHAPTER

8

Introduction to Hypothesis Testing

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- z-Scores (Chapter 5)
- Distribution of sample means (Chapter 7)
  - Expected value
  - Standard error
  - Probability and sample means

Preview

Most of us spend more time looking down at our mobile devices than we do looking up at the clouds. But if you do watch the clouds and have a little imagination, you occasionally see them form into familiar shapes. Figure 8.1 is a photograph of a cloud formation seen over Kansas City around Christmas in 2008. Do you recognize a familiar image?

The cloud pattern shown in Figure 8.1 was formed simply by chance. Specifically, it was the random forces of wind and air currents that produced a portrait of Santa Claus. The clouds did not conspire to form the image, and it was not deliberately created by a team of professional skywriters. The point we would like to make is that what appear to be meaningful patterns can be produced by random chance.

The Problem Researchers often find meaningful patterns in the sample data obtained in research studies. The problem is deciding whether the patterns found in a sample reflect real patterns that exist in the population or are simply random, chance occurrences.

The Solution To differentiate between real, systematic patterns and random, chance occurrences, researchers rely on a statistical technique known as *hypothesis testing,* which is introduced in this chapter. As you will see, a hypothesis test first determines the probability that the pattern could have been produced by chance alone. If this probability is large enough, then we conclude that the pattern can reasonably be explained by chance. However, if the probability is extremely small, then we can rule out chance as a plausible explanation and conclude that some meaningful, systematic force has created the pattern. For example, it is reasonable, once in a lifetime, to see a cloud formation that resembles Santa Claus. However, it would be extremely unlikely if the clouds also included the words "Merry Christmas" spelled out beneath Santa's face. If this happened, we would conclude that the pattern was not produced by the random forces of chance, but rather was created by a deliberate, systematic act.

FIGURE 8.1

A cloud formation seen over Kansas City.

Mark Gravetter

8.1 THE LOGIC OF HYPOTHESIS TESTING

It usually is impossible or impractical for a researcher to observe every individual in a population. Therefore, researchers usually collect data from a sample and then use the sample data to help answer questions about the population. Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest.

Hypothesis testing is one of the most commonly used inferential procedures. In fact, most of the remainder of this book examines hypothesis testing in a variety of different situations and applications. Although the details of a hypothesis test change from one situation to another, the general process remains constant. In this chapter, we introduce the general procedure for a hypothesis test. You should notice that we use the statistical techniques that have been developed in the preceding three chapters—that is, we combine the concepts of *z*-scores, probability, and the distribution of sample means to create a new statistical procedure known as a *hypothesis test*.

DEFINITION

A hypothesis test is a statistical method that uses sample data to evaluate a hypothesis about a population.

In very simple terms, the logic underlying the hypothesis-testing procedure is as follows:

1. First, we state a hypothesis about a population. Usually the hypothesis concerns the value of a population parameter. For example, we might hypothesize that American adults gain an average of $\mu = 7$ pounds between Thanksgiving and New Year's Day each year.
2. Before we select a sample, we use the hypothesis to predict the characteristics that the sample should have. For example, if we predict that the average weight gain for the population is $\mu = 7$ pounds, then we would predict that our sample should have a mean *around* 7 pounds. Remember: The sample should be similar to the population, but you always expect a certain amount of error.
3. Next, we obtain a random sample from the population. For example, we might select a sample of $n = 200$ American adults and measure the average weight change for the sample between Thanksgiving and New Year's Day.
4. Finally, we compare the obtained sample data with the prediction that was made from the hypothesis. If the sample mean is consistent with the prediction, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the prediction, then we decide that the hypothesis is wrong.

A hypothesis test is typically used in the context of a research study. That is, a researcher completes a research study and then uses a hypothesis test to evaluate the results. Depending on the type of research and the type of data, the details of the hypothesis test change from one research situation to another. In later chapters, we examine different versions of hypothesis testing that are used for different kinds of research. For now, however, we focus on the basic elements that are common to all hypothesis tests. To accomplish this general goal, we examine a hypothesis test as it applies to the simplest possible situation—using a sample mean to test a hypothesis about a population mean.

In the six chapters that follow, we consider hypothesis testing in more complex research situations involving sample means and mean differences. In Chapters 15 and 16,

we look at correlational research and examine how the relationships obtained for sample data are used to evaluate hypotheses about relationships in the population. In Chapters 17 and 18, we examine how the proportions that exist in a sample are used to test hypotheses about the corresponding proportions in the population. Chapter 19 reviews the complete set of hypothesis tests and presents a guide to help you find the appropriate test for a specific set of data.

Once again, we introduce hypothesis testing with a situation in which a researcher is using one sample mean to evaluate a hypothesis about one unknown population mean.

The unknown population Figure 8.2 shows the general research situation that we use to introduce the process of hypothesis testing. Notice that the researcher begins with a known population. This is the set of individuals as they exist *before treatment*. For this example, we are assuming that the original set of scores forms a normal distribution with $\mu = 80$ and $\sigma = 20$. The purpose of the research is to determine the effect of a treatment on the individuals in the population. That is, the goal is to determine what happens to the population *after the treatment is administered*.

To simplify the hypothesis-testing situation, one basic assumption is made about the effect of the treatment: If the treatment has any effect, it is simply to add a constant amount to (or subtract a constant amount from) each individual's score. You should recall from Chapters 3 and 4 that adding (or subtracting) a constant changes the mean but does not change the shape of the population, nor does it change the standard deviation. Thus, we assume that the population after treatment has the same shape as the original population and the same standard deviation as the original population. This assumption is incorporated into the situation shown in Figure 8.2.

Note that the unknown population, after treatment, is the focus of the research question. Specifically, the purpose of the research is to determine what would happen if the treatment were administered to every individual in the population.

The sample in the research study The goal of the hypothesis test is to determine whether the treatment has any effect on the individuals in the population (see Figure 8.2). Usually, however, we cannot administer the treatment to the entire population, so the actual research study is conducted using a sample. Figure 8.3 shows the structure of the research study from the point of view of the hypothesis test. The original population, before treatment, is shown on the left-hand side. The unknown population, after treatment, is shown on the right-hand side. Note that the unknown population is actually *hypothetical* (the treatment is never administered to the entire population). Instead, we are asking *what would happen if* the treatment were administered to the entire population. The research

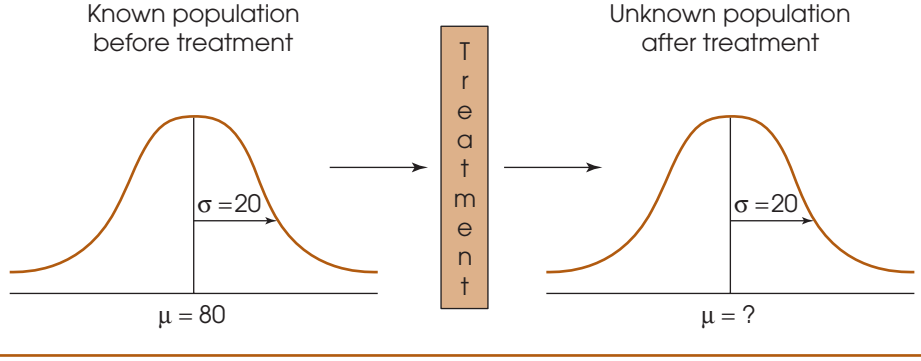
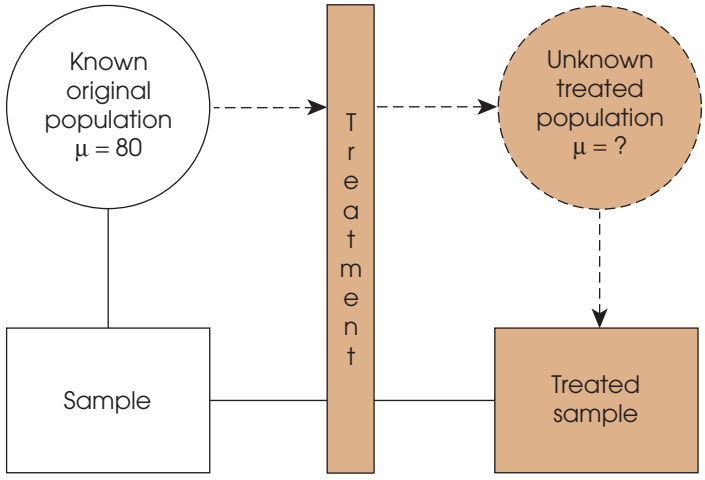


FIGURE 8.2
The basic experimental situation for hypothesis testing. It is assumed that the parameter μ is known for the population before treatment. The purpose of the experiment is to determine whether the treatment has an effect on the population mean.

FIGURE 8.3

From the point of view of the hypothesis test, the entire population receives the treatment and then a sample is selected from the treated population. In the actual research study, a sample is selected from the original population and the treatment is administered to the sample. From either perspective, the result is a treated sample that represents the treated population.



study involves selecting a sample from the original population, administering the treatment to the sample, and then recording scores for the individuals in the treated sample. Notice that the research study produces a treated sample. Although this sample was obtained indirectly, it is equivalent to a sample that is obtained directly from the unknown treated population. The hypothesis test uses the treated sample on the right-hand side of Figure 8.3 to evaluate a hypothesis about the unknown treated population on the right side of the figure.

A hypothesis test is a formalized procedure that follows a standard series of operations. In this way, researchers have a standardized method for evaluating the results of their research studies. Other researchers recognize and understand exactly how the data were evaluated and how conclusions were reached. To emphasize the formal structure of a hypothesis test, we present hypothesis testing as a four-step process that is used throughout the rest of the book. The following example provides a concrete foundation for introducing the hypothesis-testing procedure.

EXAMPLE 8.1

Researchers have noted a decline in cognitive functioning as people age (Bartus, 1990). However, the results from other research suggest that the antioxidants in foods such as blueberries can reduce and even reverse these age-related declines, at least in laboratory rats (Joseph et al., 1999). Based on these results, one might theorize that the same antioxidants might also benefit elderly humans. Suppose a researcher is interested in testing this theory.

Standardized neuropsychological tests such as the Wisconsin Card Sorting Test can be used to measure conceptual thinking ability and mental flexibility (Heaton, Chelune, Talley, Kay, & Curtiss, 1993). Performance on this type of test declines gradually with age. Suppose that our researcher selects a test for which adults older than 65 have an average score of $\mu = 80$ with a standard deviation of $\sigma = 20$. The distribution of test scores is approximately normal. The researcher's plan is to obtain a sample of $n = 25$ adults who are older than 65, and give each participant a daily dose of a blueberry supplement that is very high in antioxidants. After taking the supplement for 6 months, the participants are given the neuropsychological test to measure their level of cognitive function. If the mean score for the sample is

noticeably different from the mean for the general population of elderly adults, then the researcher can conclude that the supplement does appear to have an effect on cognitive function. On the other hand, if the sample mean is around 80 (the same as the general population mean), the researcher must conclude that the supplement does not appear to have any effect.

THE FOUR STEPS OF A HYPOTHESIS TEST

Figure 8.3 depicts the research situation that was described in the preceding example. Notice that the population after treatment is unknown. Specifically, we do not know what will happen to the mean score if the entire population of elderly adults is given the blueberry supplement. However, we do have a sample of $n = 25$ participants who have received the supplement and we can use this sample to help draw inferences about the unknown population. The following four steps outline the hypothesis-testing procedure that allows us to use sample data to answer questions about an unknown population.

STEP 1: STATE THE HYPOTHESIS

As the name implies, the process of hypothesis testing begins by stating a hypothesis about the unknown population. Actually, we state two opposing hypotheses. Notice that both hypotheses are stated in terms of population parameters.

The first, and most important, of the two hypotheses is called the *null hypothesis*. The null hypothesis states that the treatment has no effect. In general, the null hypothesis states that there is no change, no effect, no difference—nothing happened, hence the name *null*. The null hypothesis is identified by the symbol H_0. (The H stands for *hypothesis*, and the zero subscript indicates that this is the *zero-effect* hypothesis.) For the study in Example 8.1, the null hypothesis states that the blueberry supplement has no effect on cognitive functioning for the population of adults who are more than 65 years old. In symbols, this hypothesis is

The goal of inferential statistics is to make general statements about the population by using sample data. Therefore, when testing hypotheses, we make our predictions about the population parameters.

$$H_0: \mu_{\text{with supplement}} = 80$$ (Even with the supplement, the mean test score is still 80.)

DEFINITION

The **null hypothesis** (H_0) states that in the general population there is no change, no difference, or no relationship. In the context of an experiment, H_0 predicts that the independent variable (treatment) *has no effect* on the dependent variable (scores) for the population.

The second hypothesis is simply the opposite of the null hypothesis, and it is called the *scientific,* or *alternative, hypothesis* (H_1). This hypothesis states that the treatment has an effect on the dependent variable.

DEFINITION

The **alternative hypothesis** (H_1) states that there is a change, a difference, or a relationship for the general population. In the context of an experiment, H_1 predicts that the independent variable (treatment) *does have an effect* on the dependent variable.

The null hypothesis and the alternative hypothesis are mutually exclusive and exhaustive. They cannot both be true, and one of them must be true. The data determine which one should be rejected.

For this example, the alternative hypothesis states that the supplement does have an effect on cognitive functioning for the population and will cause a change in the mean score. In symbols, the alternative hypothesis is represented as

$$H_1: \mu_{\text{with supplement}} \neq 80$$ (With the supplement, the mean test score is different from 80.)

Notice that the alternative hypothesis simply states that there will be some type of change. It does not specify whether the effect will be increased or decreased test scores. In some circumstances, it is appropriate for the alternative hypothesis to specify the direction of the effect. For example, the researcher might hypothesize that the supplement will increase neuropsychological test scores ($\mu > 80$). This type of hypothesis results in a directional hypothesis test, which is examined in detail later in this chapter. For now we concentrate on nondirectional tests, for which the hypotheses simply state that the treatment has no effect (H_0) or has some effect (H_1).

STEP 2: SET THE CRITERIA FOR A DECISION

Eventually the researcher uses the data from the sample to evaluate the credibility of the null hypothesis. The data either provide support for the null hypothesis or tend to refute the null hypothesis. In particular, if there is a big discrepancy between the data and the null hypothesis, then we conclude that the null hypothesis is wrong.

To formalize the decision process, we use the null hypothesis to predict the kind of sample mean that ought to be obtained. Specifically, we determine exactly which sample means are consistent with the null hypothesis and which sample means are at odds with the null hypothesis.

For our example, the null hypothesis states that the supplement has no effect and the population mean is still $\mu = 80$. If this is true, then the sample mean should have a value around 80. Therefore, a sample mean near 80 is consistent with the null hypothesis. On the other hand, a sample mean that is very different from 80 is not consistent with the null hypothesis. To determine exactly which values are "near" 80 and which values are "very different from" 80, we examine all of the possible sample means that could be obtained if the null hypothesis is true. For our example, this is the distribution of sample means for $n = 25$. According to the null hypothesis, this distribution is centered at $\mu = 80$. The distribution of sample means is then divided into two sections:

1. Sample means that are likely to be obtained if H_0 is true; that is, sample means that are close to the null hypothesis
2. Sample means that are very unlikely to be obtained if H_0 is true; that is, sample means that are very different from the null hypothesis

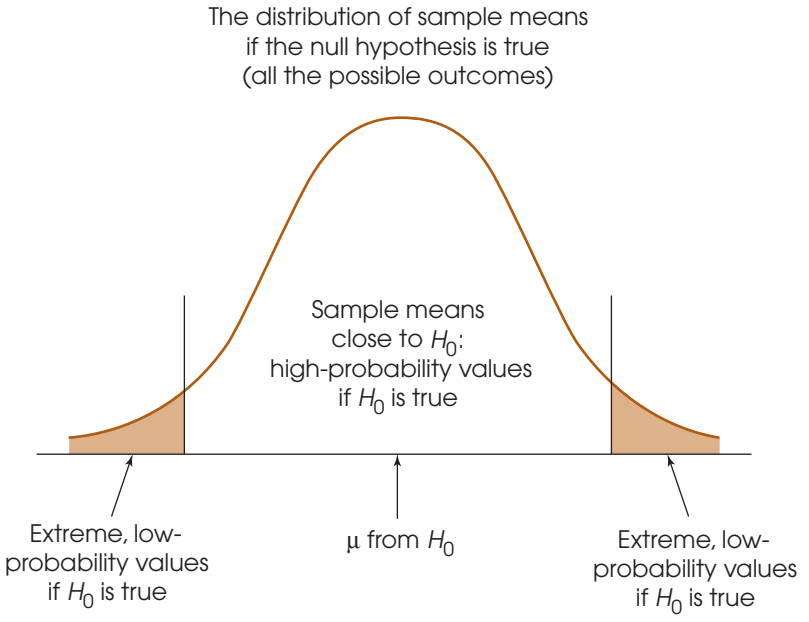
Figure 8.4 shows the distribution of sample means divided into these two sections. Notice that the high-probability samples are located in the center of the distribution and have sample means close to the value specified in the null hypothesis. On the other hand, the low-probability samples are located in the extreme tails of the distribution. After the distribution has been divided in this way, we can compare our sample data with the values in the distribution. Specifically, we can determine whether our sample mean is consistent with the null hypothesis (like the values in the center of the distribution) or whether our sample mean is very different from the null hypothesis (like the values in the extreme tails).

The alpha level To find the boundaries that separate the high-probability samples from the low-probability samples, we must define exactly what is meant by "low" probability and "high" probability. This is accomplished by selecting a specific probability value, which is known as the *level of significance,* or the *alpha level,* for the hypothesis test. The alpha (α) value is a small probability that is used to identify the low-probability samples. By convention, commonly used alpha levels are $\alpha = .05$ (5%), $\alpha = .01$ (1%), and $\alpha = .001$ (0.1%). For example, with $\alpha = .05$, we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values).

With rare exceptions, an alpha level is never larger than .05.

FIGURE 8.4

The set of potential samples is divided into those that are likely to be obtained and those that are very unlikely to be obtained if the null hypothesis is true.



The extremely unlikely values, as defined by the alpha level, make up what is called the *critical region*. These extreme values in the tails of the distribution define outcomes that are not consistent with the null hypothesis; that is, they are very unlikely to occur if the null hypothesis is true. Whenever the data from a research study produce a sample mean that is located in the critical region, we conclude that the data are not consistent with the null hypothesis, and we reject the null hypothesis.

DEFINITIONS

The **alpha level,** or the **level of significance,** is a probability value that is used to define the concept of "very unlikely" in a hypothesis test.

The **critical region** is composed of the extreme sample values that are very unlikely (as defined by the alpha level) to be obtained if the null hypothesis is true. The boundaries for the critical region are determined by the alpha level. If sample data fall in the critical region, the null hypothesis is rejected.

Technically, the critical region is defined by sample outcomes that are *very unlikely* to occur if the treatment has no effect (that is, if the null hypothesis is true). Reversing the point of view, we can also define the critical region as sample values that provide *convincing evidence* that the treatment really does have an effect. For our example, the regular population of elderly adults has a mean test score of $\mu = 80$. We selected a sample from this population and administered a treatment (the blueberry supplement) to the individuals in the sample. What kind of sample mean would convince you that the treatment has an effect? It should be obvious that the most convincing evidence would be a sample mean that is really different from $\mu = 80$. In a hypothesis test, the critical region is determined by sample values that are "really different" from the original population.

The boundaries for the critical region To determine the exact location for the boundaries that define the critical region, we use the alpha-level probability and the unit normal table. In most cases, the distribution of sample means is normal, and the unit normal table provides the precise z-score location for the critical region boundaries. With $\alpha = .05$, for example, the boundaries separate the extreme 5% from the middle 95%. Because the extreme 5% is split between two tails of the distribution, there is exactly 2.5% (or 0.0250) in each tail. In the unit normal table, you can look up a proportion of 0.0250 in column C (the tail) and find that the z-score boundary is $z = 1.96$. Thus, for any normal distribution, the extreme 5% is in the tails of the distribution beyond $z = +1.96$ and $z = -1.96$. These values define the boundaries of the critical region for a hypothesis test using $\alpha = .05$ (Figure 8.5).

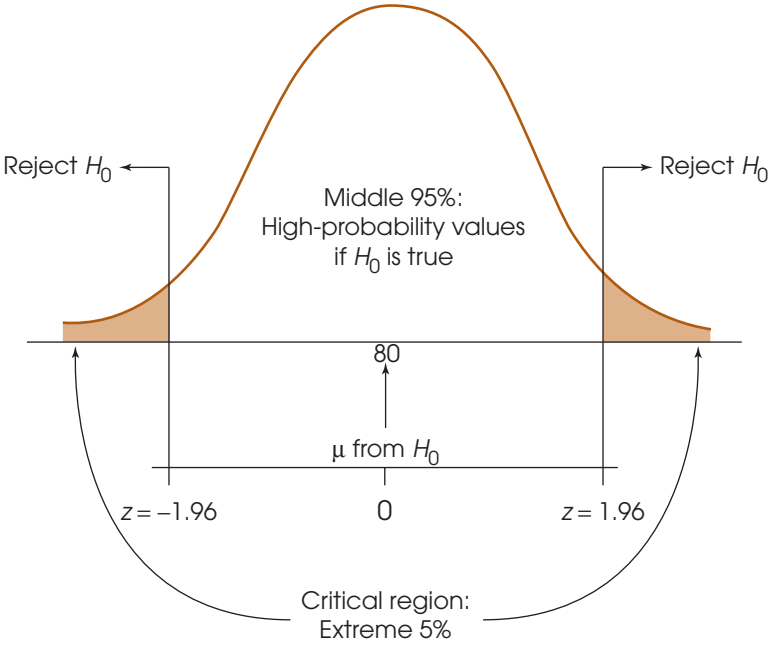
Similarly, an alpha level of $\alpha = .01$ means that 1%, or .0100, is split between the two tails. In this case, the proportion in each tail is .0050, and the corresponding z-score boundaries are $z = \pm 2.58$ (± 2.57 is equally good). For $\alpha = .001$, the boundaries are located at $z = \pm 3.30$. You should verify these values in the unit normal table and be sure that you understand exactly how they are obtained.

LEARNING CHECK

1. The city school district is considering increasing class size in the elementary schools. However, some members of the school board are concerned that larger classes may have a negative effect on student learning. In words, what would the null hypothesis say about the effect of class size on student learning?
2. If the alpha level is increased from $\alpha = .01$ to $\alpha = .05$, then the boundaries for the critical region move farther away from the center of the distribution. (True or false?)
3. If a researcher conducted a hypothesis test with an alpha level of $\alpha = .02$, what z-score values would form the boundaries for the critical region?

FIGURE 8.5

The critical region (very unlikely outcomes) for $\alpha = .05$.



ANSWERS

1. The null hypothesis would say that class size has no effect on student learning.
2. False. A larger alpha means that the boundaries for the critical region move closer to the center of the distribution.
3. The .02 would be split between the two tails, with .01 in each tail. The z-score boundaries would be $z = +2.33$ and $z = -2.33$.

STEP 3: COLLECT DATA AND COMPUTE SAMPLE STATISTICS

At this time, we select a sample of adults who are more than 65 years old and give each one a daily dose of the blueberry supplement. After 6 months, the neuropsychological test is used to measure cognitive function for the sample of participants. Notice that the data are collected *after* the researcher has stated the hypotheses and established the criteria for a decision. This sequence of events helps to ensure that a researcher makes an honest, objective evaluation of the data and does not tamper with the decision criteria after the experimental outcome is known.

Next, the raw data from the sample are summarized with the appropriate statistics: For this example, the researcher would compute the sample mean. Now it is possible for the researcher to compare the sample mean (the data) with the null hypothesis. This is the heart of the hypothesis test: comparing the data with the hypothesis.

The comparison is accomplished by computing a z-score that describes exactly where the sample mean is located relative to the hypothesized population mean from H_0. In step 2, we constructed the distribution of sample means that would be expected if the null hypothesis were true—that is, the entire set of sample means that could be obtained if the treatment has no effect (see Figure 8.5). Now we calculate a z-score that identifies where our sample mean is located in this hypothesized distribution. The z-score formula for a sample mean is

$$z = \frac{M - \mu}{\sigma_M}$$

In the formula, the value of the sample mean (M) is obtained from the sample data, and the value of μ is obtained from the null hypothesis. Thus, the z-score formula can be expressed in words as follows:

$$z = \frac{\text{sample mean} - \text{hypothesized population mean}}{\text{standard error between } M \text{ and } \mu}$$

Notice that the top of the z-score formula measures how much difference there is between the data and the hypothesis. The bottom of the formula measures the standard distance that ought to exist between a sample mean and the population mean.

STEP 4: MAKE A DECISION

In the final step, the researcher uses the z-score value obtained in step 3 to make a decision about the null hypothesis according to the criteria established in step 2. There are two possible outcomes:

1. The sample data are located in the critical region. By definition, a sample value in the critical region is very unlikely to occur if the null hypothesis is true. Therefore, we conclude that the sample is not consistent with H_0 and our decision is to *reject the null hypothesis*. Remember, the null hypothesis states that there is no treatment effect, so rejecting H_0 means that we are concluding that the treatment did have an effect.

For the example we have been considering, suppose that the sample produced a mean of $M = 92$ after taking the supplement for 6 months. The null hypothesis states that the population mean is $\mu = 80$ and, with $n = 25$ and $\sigma = 20$, the standard error for the sample mean is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

Thus, a sample mean of $M = 92$ produces a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{92 - 80}{4} = \frac{12}{4} = 3.00$$

With an alpha level of $\alpha = .05$, this z-score is far beyond the boundary of 1.96. Because the sample z-score is in the critical region, we reject the null hypothesis and conclude that the blueberry supplement did have an effect on cognitive functioning.

2. The second possibility is that the sample data are not in the critical region. In this case, the sample mean is reasonably close to the population mean specified in the null hypothesis (in the center of the distribution). Because the data do not provide strong evidence that the null hypothesis is wrong, our conclusion is to *fail to reject the null hypothesis*. This conclusion means that the treatment does not appear to have an effect.

   For the research study examining the blueberry supplement, suppose our sample produced a mean test score of $M = 84$. As before, the standard error for a sample of $n = 25$ is $\sigma_M = 4$, and the null hypothesis states that $\mu = 80$. These values produce a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{84 - 80}{4} = \frac{4}{4} = 1.00$$

   The z-score of 1.00 is not in the critical region. Therefore, we would fail to reject the null hypothesis and conclude that the blueberry supplement does not appear to have an effect on cognitive functioning.

In general, the final decision is made by comparing our treated sample with the distribution of sample means that would be obtained for untreated samples. If our treated sample looks much the same as samples that do not receive the blueberry treatment, we conclude that the treatment does not appear to have any effect. On the other hand, if the treated sample is noticeably different from the majority of untreated samples, we conclude that the treatment does have an effect.

An Analogy for Hypothesis Testing It may seem awkward to phrase both of the two possible decisions in terms of rejecting the null hypothesis; either we reject H_0 or we fail to reject H_0. These two decisions may be easier to understand if you think of a research study as an attempt to gather evidence to prove that a treatment works. From this perspective, the process of conducting a hypothesis test is similar to the process that takes place during a jury trial. For example,

1. The test begins with a null hypothesis stating that there is no treatment effect. The trial begins with a null hypothesis that the defendant did not commit a crime (innocent until proven guilty).

2. The research study gathers evidence to show that the treatment actually does have an effect, and the police gather evidence to show that the defendant really did commit a crime. Note that both are trying to refute the null hypothesis.
3. If there is enough evidence, the researcher rejects the null hypothesis and concludes that there really is a treatment effect. If there is enough evidence, the jury rejects the hypothesis and concludes that the defendant is guilty of a crime.
4. If there is not enough evidence, the researcher fails to reject the null hypothesis. Note that the researcher does not conclude that there is no treatment effect, simply that there is not enough evidence to conclude that there is an effect. Similarly, if there is not enough evidence, the jury fails to find the defendant guilty. Note that the jury does not conclude that the defendant is innocent, simply that there is not enough evidence for a guilty verdict.

A CLOSER LOOK AT THE *z*-SCORE STATISTIC

The *z*-score statistic that is used in the hypothesis test is the first specific example of what is called a *test statistic*. The term *test statistic* simply indicates that the sample data are converted into a single, specific statistic that is used to test the hypotheses. In the chapters that follow, we introduce several other test statistics that are used in a variety of different research situations. However, most of the new test statistics have the same basic structure and serve the same purpose as the *z*-score. We have already described the *z*-score equation as a formal method for comparing the sample data and the population hypothesis. In this section, we discuss the *z*-score from two other perspectives that may give you a better understanding of hypothesis testing and the role that *z*-scores play in this inferential technique. In each case, keep in mind that the *z*-score serves as a general model for other test statistics that come in future chapters.

The *z*-score formula as a recipe The *z*-score formula, like any formula, can be viewed as a recipe. If you follow instructions and use all of the right ingredients, the formula produces a *z*-score. In the hypothesis-testing situation, however, you do not have all of the necessary ingredients. Specifically, you do not know the value for the population mean (μ), which is one component, or ingredient, in the formula.

This situation is similar to trying to follow a cake recipe in which one of the ingredients is not clearly listed. For example, the recipe may call for flour but there is a grease stain that makes it impossible to read how much flour. Faced with this situation, you might try the following steps:

1. Make a hypothesis about the amount of flour. For example, hypothesize that the correct amount is 2 cups.
2. To test your hypothesis, add the rest of the ingredients along with the hypothesized amount of flour and bake the cake.
3. If the cake turns out to be good, you can reasonably conclude that your hypothesis was correct. But if the cake is terrible, you conclude that your hypothesis was wrong.

In a hypothesis test with *z*-scores, we do essentially the same thing. We have a formula (recipe) for *z*-scores but one ingredient is missing. Specifically, we do not know the value for the population mean, μ. Therefore, we try the following steps:

1. Make a hypothesis about the value of μ. This is the null hypothesis.
2. Plug the hypothesized value in the formula along with the other values (ingredients).

3. If the formula produces a z-score near zero (which is where z-scores are supposed to be), we conclude that the hypothesis was correct. On the other hand, if the formula produces an extreme value (a very unlikely result), we conclude that the hypothesis was wrong.

The z-score formula as a ratio In the context of a hypothesis test, the z-score formula has the following structure:

$$z = \frac{M - \mu}{\sigma_M} = \frac{\text{sample mean} - \text{hypothesized population mean}}{\text{standard error between } M \text{ and } \mu}$$

Notice that the numerator of the formula involves a direct comparison between the sample data and the null hypothesis. In particular, the numerator measures the obtained difference between the sample mean and the hypothesized population mean. The standard error in the denominator of the formula measures the standard amount of distance that exists naturally between a sample mean and the population mean without any treatment effect causing the sample to be different. Thus, the z-score formula (and most other test statistics) forms a ratio

$$z = \frac{\text{actual difference between the sample } (M) \text{ and the hypothesis } (\mu)}{\text{standard difference between } M \text{ and } \mu \text{ with no treatment effect}}$$

Thus, for example, a z-score of $z = 3.00$ means that the obtained difference between the sample and the hypothesis is 3 times bigger than would be expected if the treatment had no effect.

In general, a large value for a test statistic like the z-score indicates a large discrepancy between the sample data and the null hypothesis. Specifically, a large value indicates that the sample data are very unlikely to have occurred by chance alone. Therefore, when we obtain a large value (in the critical region), we conclude that it must have been caused by a treatment effect.

LEARNING CHECK

1. A researcher selects a sample of $n = 16$ individuals from a normal population with a mean of $\mu = 40$ and $\sigma = 8$. A treatment is administered to the sample and, after treatment, the sample mean is $M = 43$. If the researcher uses a hypothesis test to evaluate the treatment effect, what z-score would be obtained for this sample?
2. A small value (near zero) for the z-score statistic is evidence that the sample data are consistent with the null hypothesis. (True or false?)
3. A z-score value in the critical region means that you should reject the null hypothesis. (True or false?)

ANSWERS

1. The standard error is 2 points and $z = 3/2 = 1.50$.
2. True. A z-score near zero indicates that the data support the null hypothesis.
3. True. A z-score value in the critical region means that the sample is not consistent with the null hypothesis.

8.2 UNCERTAINTY AND ERRORS IN HYPOTHESIS TESTING

Hypothesis testing is an *inferential process,* which means that it uses limited information as the basis for reaching a general conclusion. Specifically, a sample provides only limited or incomplete information about the whole population, and yet a hypothesis test uses a sample to draw a conclusion about the population. In this situation, there is always the possibility that an incorrect conclusion will be made. Although sample data are usually representative of the population, there is always a chance that the sample is misleading and will cause a researcher to make the wrong decision about the research results. In a hypothesis test, there are two different kinds of errors that can be made.

TYPE I ERRORS

It is possible that the data will lead you to reject the null hypothesis when in fact the treatment has no effect. Remember: Samples are not expected to be identical to their populations, and some extreme samples can be very different from the populations that they are supposed to represent. If a researcher selects one of these extreme samples by chance, then the data from the sample may give the appearance of a strong treatment effect, even though there is no real effect. In the previous section, for example, we discussed a research study examining how a food supplement that is high in antioxidants affects the cognitive functioning of elderly adults. Suppose that the researcher selects a sample of $n = 25$ people who already have cognitive functioning that is well above average. Even if the blueberry supplement (the treatment) has no effect at all, these people will still score higher than average on the neuropsychological test when they are tested after 6 months of taking the supplement. In this case, the researcher is likely to conclude that the treatment does have an effect, when in fact it really does not. This is an example of what is called a *Type I error.*

DEFINITION

A **Type I error** occurs when a researcher rejects a null hypothesis that is actually true. In a typical research situation, a Type I error means that the researcher concludes that a treatment does have an effect when, in fact, it has no effect.

You should realize that a Type I error is not a stupid mistake in the sense that a researcher is overlooking something that should be perfectly obvious. On the contrary, the researcher is looking at sample data that appear to show a clear treatment effect. The researcher then makes a careful decision based on the available information. The problem is that the information from the sample is misleading.

In most research situations, the consequences of a Type I error can be very serious. Because the researcher has rejected the null hypothesis and believes that the treatment has a real effect, it is likely that the researcher will report or even publish the research results. A Type I error, however, means that this is a false report. Thus, Type I errors lead to false reports in the scientific literature. Other researchers may try to build theories or develop other experiments based on the false results. A lot of precious time and resources may be wasted.

The Probability of a Type I Error A Type I error occurs when a researcher unknowingly obtains an extreme, nonrepresentative sample. Fortunately, the hypothesis test is structured to minimize the risk that this will occur. Figure 8.5 shows the distribution of sample means and the critical region for the research study we have been discussing. This distribution contains all of the possible sample means for samples of $n = 25$ if the null hypothesis is true. Notice that most of the sample means are near the

hypothesized population mean, $\mu = 80$, and that means in the critical region are very unlikely to occur.

With an alpha level of $\alpha = .05$, only 5% of the samples have means in the critical region. Therefore, there is only a 5% probability ($p = .05$) that one of these samples will be obtained. Thus, the alpha level determines the probability of obtaining a sample mean in the critical region when the null hypothesis is true. In other words, the alpha level determines the probability of a Type I error.

DEFINITION

The **alpha level** for a hypothesis test is the probability that the test will lead to a Type I error. That is, the alpha level determines the probability of obtaining sample data in the critical region even though the null hypothesis is true.

In summary, whenever the sample data are in the critical region, the appropriate decision for a hypothesis test is to reject the null hypothesis. Normally this is the correct decision because the treatment has caused the sample to be different from the original population; that is, the treatment effect has pushed the sample mean into the critical region. In this case, the hypothesis test has correctly identified a real treatment effect. Occasionally, however, sample data are in the critical region just by chance, without any treatment effect. When this occurs, the researcher makes a Type I error; that is, the researcher concludes that a treatment effect exists when in fact it does not. Fortunately, the risk of a Type I error is small and is under the control of the researcher. Specifically, the probability of a Type I error is equal to the alpha level.

TYPE II ERRORS

Whenever a researcher rejects the null hypothesis, there is a risk of a Type I error. Similarly, whenever a researcher fails to reject the null hypothesis, there is a risk of a *Type II error*. By definition, a Type II error is the failure to reject a false null hypothesis. In more straightforward English, a Type II error means that a treatment effect really exists, but the hypothesis test fails to detect it.

DEFINITION

A **Type II error** occurs when a researcher fails to reject a null hypothesis that is really false. In a typical research situation, a Type II error means that the hypothesis test has failed to detect a real treatment effect.

A Type II error occurs when the sample mean is not in the critical region even though the treatment has had an effect on the sample. Often this happens when the effect of the treatment is relatively small. In this case, the treatment does influence the sample, but the magnitude of the effect is not big enough to move the sample mean into the critical region. Because the sample is not substantially different from the original population (it is not in the critical region), the statistical decision is to fail to reject the null hypothesis and to conclude that there is not enough evidence to say that there is a treatment effect.

The consequences of a Type II error are usually not as serious as those of a Type I error. In general terms, a Type II error means that the research data do not show the results that the researcher had hoped to obtain. The researcher can accept this outcome and conclude that the treatment either has no effect or has only a small effect that is not worth pursuing, or the researcher can repeat the experiment (usually with some improvement, such as a larger sample) and try to demonstrate that the treatment really does work.

Unlike a Type I error, it is impossible to determine a single, exact probability for a Type II error. Instead, the probability of a Type II error depends on a variety of factors and therefore is a function, rather than a specific number. Nonetheless, the probability of a Type II error is represented by the symbol β, the Greek letter *beta.*

In summary, a hypothesis test always leads to one of two decisions:

1. The sample data provide sufficient evidence to reject the null hypothesis and conclude that the treatment has an effect.
2. The sample data do not provide enough evidence to reject the null hypothesis. In this case, you fail to reject H_0 and conclude that the treatment does not appear to have an effect.

In either case, there is a chance that the data are misleading and the decision is wrong. The complete set of decisions and outcomes is shown in Table 8.1. The risk of an error is especially important in the case of a Type I error, which can lead to a false report. Fortunately, the probability of a Type I error is determined by the alpha level, which is completely under the control of the researcher. At the beginning of a hypothesis test, the researcher states the hypotheses and selects the alpha level, which immediately determines the risk that a Type I error will be made.

SELECTING AN ALPHA LEVEL

As you have seen, the alpha level for a hypothesis test serves two very important functions. First, the alpha level helps to determine the boundaries for the critical region by defining the concept of "very unlikely" outcomes. At the same time, the alpha level determines the probability of a Type I error. When you select a value for alpha at the beginning of a hypothesis test, your decision influences both of these functions.

The primary concern when selecting an alpha level is to minimize the risk of a Type I error. Thus, alpha levels tend to be very small probability values. By convention, the largest permissible value is $\alpha = .05$. When there is no treatment effect, an alpha level of .05 means that there is still a 5% risk, or a 1-in-20 probability, of rejecting the null hypothesis and committing a Type I error. Because the consequences of a Type I error can be relatively serious, many individual researchers and many scientific publications prefer to use a more conservative alpha level such as .01 or .001 to reduce the risk that a false report is published and becomes part of the scientific literature. (For more information on the origins of the .05 level of significance, see the excellent short article by Cowles and Davis, 1982.)

TABLE 8.1
Possible outcomes of a statistical decision

		Actual Situation	
		No Effect, H_0 True	Effect Exists, H_0 False
Experimenter's Decision	Reject H_0	Type I error	Decision correct
	Retain H_0	Decision correct	Type II error

At this point, it may appear that the best strategy for selecting an alpha level is to choose the smallest possible value to minimize the risk of a Type I error. However, there is a different kind of risk that develops as the alpha level is lowered. Specifically, a lower alpha level means less risk of a Type I error, but it also means that the hypothesis test demands more evidence from the research results.

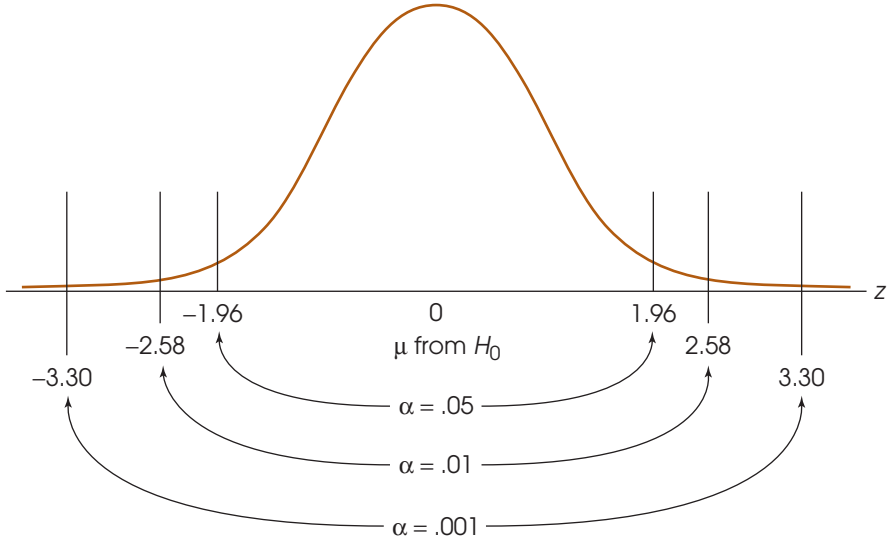
The trade-off between the risk of a Type I error and the demands of the test is controlled by the boundaries of the critical region. For the hypothesis test to conclude that the treatment does have an effect, the sample data must be in the critical region. If the treatment really has an effect, it should cause the sample to be different from the original population; essentially, the treatment should push the sample into the critical region. However, as the alpha level is lowered, the boundaries for the critical region move farther out and become more difficult to reach. Figure 8.6 shows how the boundaries for the critical region move farther into the tails as the alpha level decreases. Notice that $z = 0$, in the center of the distribution, corresponds to the value of μ specified in the null hypothesis. The boundaries for the critical region determine how much distance between the sample mean and μ is needed to reject the null hypothesis. As the alpha level gets smaller, this distance gets larger.

Thus, an extremely small alpha level, such as .000001 (one in a million), would mean almost no risk of a Type I error but would push the critical region so far out that it would become essentially impossible to ever reject the null hypothesis; that is, it would require an enormous treatment effect before the sample data would reach the critical boundaries.

In general, researchers try to maintain a balance between the risk of a Type I error and the demands of the hypothesis test. Alpha levels of .05, .01, and .001 are considered reasonably good values because they provide a low risk of error without placing excessive demands on the research results.

FIGURE 8.6

The locations of the critical region boundaries for three different levels of significance: $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.



LEARNING CHECK

1. Define a Type I error.
2. Define a Type II error.
3. Under what circumstances is a Type II error likely to occur?
4. If a sample mean is in the critical region with $\alpha = .05$, it would still (always) be in the critical region if alpha were changed to $\alpha = .01$. (True or false?)
5. If a sample mean is in the critical region with $\alpha = .01$, it would still (always) be in the critical region if alpha were changed to $\alpha = .05$. (True or false?)

ANSWERS

1. A Type I error is rejecting a true null hypothesis—that is, saying that the treatment has an effect when, in fact, it does not.
2. A Type II error is the failure to reject a false null hypothesis. In terms of a research study, a Type II error occurs when a study fails to detect a treatment effect that really exists.
3. A Type II error is likely to occur when the treatment effect is very small. In this case, a research study is more likely to fail to detect the effect.
4. False. With $\alpha = .01$, the boundaries for the critical region move farther out into the tails of the distribution. It is possible that a sample mean could be beyond the .05 boundary but not beyond the .01 boundary.
5. True. With $\alpha = .01$, the boundaries for the critical region are farther out into the tails of the distribution than for $\alpha = .05$. If a sample mean is beyond the .01 boundary it is definitely beyond the .05 boundary.

8.3 AN EXAMPLE OF A HYPOTHESIS TEST

At this time, we have introduced all the elements of a hypothesis test. In this section, we present a complete example of the hypothesis-testing process and discuss how the results from a hypothesis test are presented in a research report. For purposes of demonstration, the following scenario is used to provide a concrete background for the hypothesis-testing process.

EXAMPLE 8.2

Alcohol appears to be involved in a variety of birth defects, including low birth weight and retarded growth. A researcher would like to investigate the effect of prenatal alcohol exposure on birth weight. A random sample of $n = 16$ pregnant rats is obtained. The mother rats are given daily doses of alcohol. At birth, one pup is selected from each litter to produce a sample of $n = 16$ newborn rats. The average weight for the sample is $M = 15$ grams. The researcher would like to compare the sample with the general population of rats. It is known that regular newborn rats (not exposed to alcohol) have an average weight of $\mu = 18$ grams. The distribution of weights is normal with $\sigma = 4$. Figure 8.7 shows the overall research situation. Notice that the researcher's question concerns the unknown population that is exposed to alcohol. Also notice that we have a sample representing the unknown population, and we have a hypothesis about the unknown population mean. Specifically, the null hypothesis says that the alcohol has no effect and the unknown mean is still $\mu = 18$. The goal of the hypothesis test is to determine whether the sample data are compatible with the hypothesis.

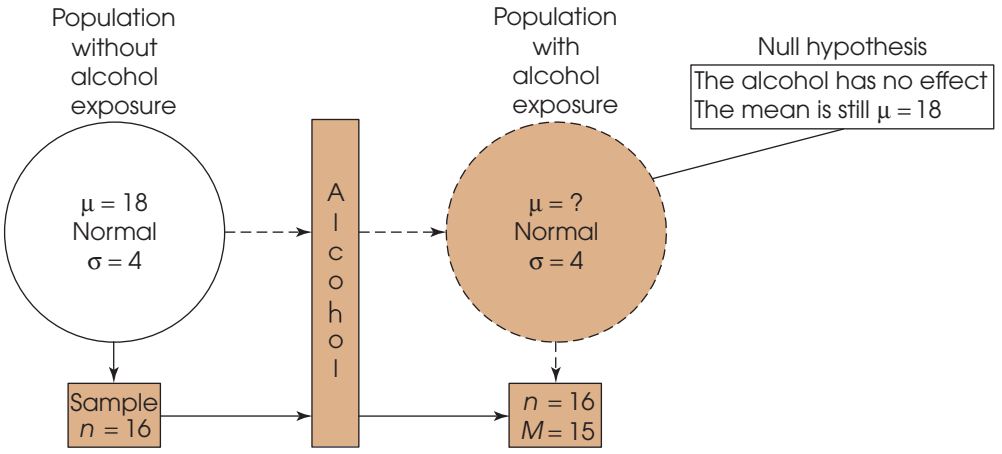


FIGURE 8.7

The structure of a research study to determine whether prenatal alcohol affects birth weight. A sample is selected from the original population and is exposed to alcohol. The question is what would happen if the entire population were exposed to alcohol. The treated sample provides information about the unknown treated population.

The following steps outline the hypothesis test that evaluates the effect of alcohol exposure on birth weight.

STEP 1 *State the hypotheses, and select the alpha level.* Both hypotheses concern the unknown population that is exposed to alcohol (the population on the right-hand side of Figure 8.7). The null hypothesis states that exposure to alcohol has no effect on birth weight. Thus, the population of rats with alcohol exposure should have the same mean birth weight as the regular, unexposed rats. In symbols,

H_0: $\mu_{\text{alcohol exposure}} = 18$ (Even with alcohol exposure, the rats still average 18 grams at birth.)

The alternative hypothesis states that alcohol exposure does affect birth weight, so the exposed population should be different from the regular rats. In symbols,

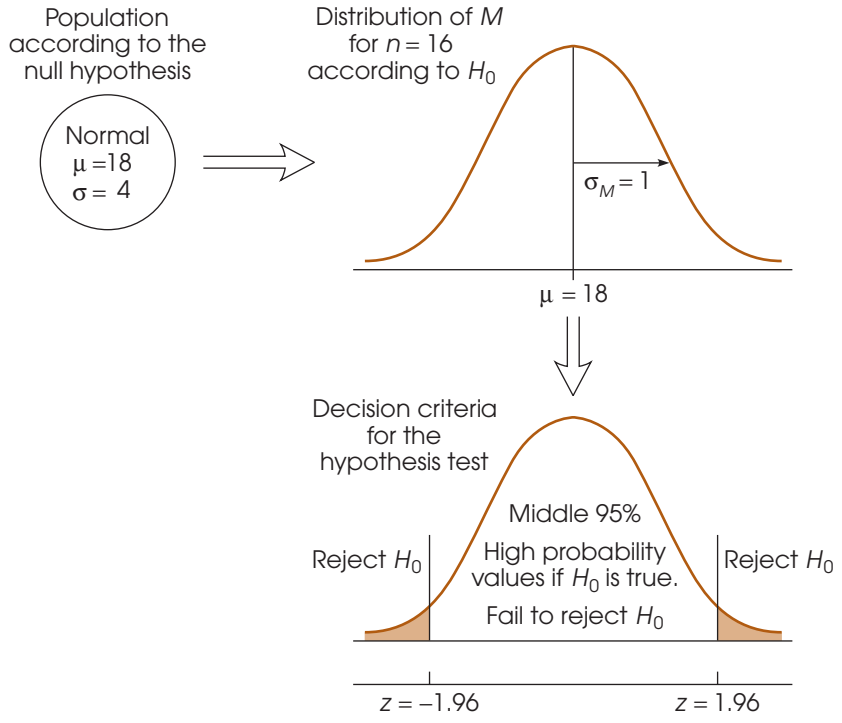
H_1: $\mu_{\text{alcohol exposure}} \neq 18$ (Alcohol exposure will change birth weight.)

Notice that both hypotheses concern the unknown population. For this test, we will use an alpha level of $\alpha = .05$. That is, we are taking a 5% risk of committing a Type I error.

STEP 2 *Set the decision criteria by locating the critical region.* By definition, the critical region consists of outcomes that are very unlikely if the null hypothesis is true. To locate the critical region we go through a three-stage process that is portrayed in Figure 8.8. We begin with the null hypothesis, which states that the alcohol has no effect on newborn rats. If H_0 is true, the population treated with alcohol is the same as the original population: that is, a normal distribution with $\mu = 18$ and $\sigma = 4$. Next, we consider all

FIGURE 8.8

Locating the critical region as a three-step process. You begin with the population of scores that is predicted by the null hypothesis. Then, you construct the distribution of sample means for the sample size that is being used. The distribution of sample means corresponds to all the possible outcomes that could be obtained if H_0 is true. Finally, you use z-scores to separate the extreme outcomes (as defined by the alpha level) from the high-probability outcomes. The extreme values determine the critical region.



the possible outcomes for a sample of $n = 16$ newborn rats. This is the distribution of sample means for $n = 16$. For this example, the distribution of sample means is normal, is centered at $\mu = 18$ (according to H_0), and has a standard error of $\sigma_M = \frac{4}{\sqrt{16}} = 1$.

Finally, we use the distribution of sample means to identify the critical region, which consists of those outcomes that are very unlikely if the null hypothesis is true. With $\alpha = .05$, the critical region consists of the extreme 5% of the distribution. As we saw earlier, for any normal distribution, z-scores of $z = \pm 1.96$ separate the middle 95% from the extreme 5% (a proportion of 0.0250 in each tail). Thus, we have identified the sample means that, according to the null hypothesis, are very unlikely to occur. It is the unlikely sample means, those with z-score values beyond ± 1.96, that form the critical region for the test. If we obtain a sample mean that is in the critical region, we conclude that the sample is not compatible with the null hypothesis and we reject H_0.

STEP 3 *Collect the data, and compute the test statistic.* At this point, we would select one newborn pup from each of the $n = 16$ mothers that received alcohol during pregnancy. The birth weight is recorded for each pup and the sample mean is computed. For this example, we obtained a sample mean of $M = 15$ grams. The sample mean is then converted to a z-score, which is our test statistic.

$$z = \frac{M - \mu}{\sigma_M} = \frac{15 - 18}{1} = \frac{-3}{1} = -3.00$$

STEP 4 *Make a decision*. The z-score computed in step 3 has a value of -3.00, which is beyond the boundary of -1.96. Therefore, the sample mean is located in the critical region. This is a very unlikely outcome if the null hypothesis is true, so our decision is to reject the null hypothesis. In addition to this statistical decision concerning the null hypothesis, it is customary to state a conclusion about the results of the research study. For this example, we conclude that prenatal exposure to alcohol does have a significant effect on birth weight.

IN THE LITERATURE

REPORTING THE RESULTS OF THE STATISTICAL TEST

A special jargon and notational system are used in published reports of hypothesis tests. When you are reading a scientific journal, for example, you typically are not told explicitly that the researcher evaluated the data using a z-score as a test statistic with an alpha level of .05. Nor are you told that "the null hypothesis is rejected." Instead, you see a statement such as:

> The treatment with alcohol had a significant effect on the birth weight of newborn rats, $z = 3.00$, $p < .05$.

Let us examine this statement, piece by piece. First, what is meant by the word *significant*? In statistical tests, a *significant* result means that the null hypothesis has been rejected, which means that the result is very unlikely to have occurred merely by chance. For this example, the null hypothesis stated that the alcohol has no effect, however the data clearly indicate that the alcohol did have an effect. Specifically, it is very unlikely that the data would have been obtained if the alcohol did not have an effect.

DEFINITION

A result is said to be **significant,** or **statistically significant,** if it is very unlikely to occur when the null hypothesis is true. That is, the result is sufficient to reject the null hypothesis. Thus, a treatment has a significant effect if the decision from the hypothesis test is to reject H_0.

Next, what is the meaning of $z = 3.00$? The z indicates that a z-score was used as the test statistic to evaluate the sample data and that its value is 3.00. Finally, what is meant by $p < .05$? This part of the statement is a conventional way of specifying the alpha level that was used for the hypothesis test. It also acknowledges the possibility (and the probability) of a Type I error. Specifically, the researcher is reporting that the treatment had an effect but admits that this could be a false report. That is, it is possible that the sample mean was in the critical region even though the alcohol had no effect. However, the probability (p) of obtaining a sample mean in the critical region is extremely small (less than .05) if there is no treatment effect.

In circumstances in which the statistical decision is to *fail to reject* H_0, the report might state that.

> There was no evidence that the alcohol had an effect on birth weight, $z = 1.30$, $p > .05$.

The APA style does not use a leading zero in a probability value that refers to a level of significance.

In that case, we would be saying that the obtained result, $z = 1.30$, is not unusual (not in the critical region) and that it has a relatively high probability of occurring (greater than .05) even if the null hypothesis is true and there is no treatment effect.

Sometimes students become confused trying to differentiate between $p < .05$ and $p > .05$. Remember that you reject the null hypothesis with extreme, low-probability values, located in the critical region in the tails of the distribution. Thus, a significant result that rejects the null hypothesis corresponds to $p < .05$ (Figure 8.9).

When a hypothesis test is conducted using a computer program, the printout often includes not only a z-score value but also an exact value for p, the probability that the result occurred without any treatment effect. In this case, researchers are encouraged to report the exact p value instead of using the less-than or greater-than notation. For example, a research report might state that the treatment effect was significant, with $z = 2.45$, $p = .0142$. When using exact values for p, however, you must still satisfy the traditional criterion for significance; specifically, the p value must be smaller than .05 to be considered statistically significant. Remember: The p value is the probability that the result would occur if H_0 were true (without any treatment effect), which is also the probability of a Type I error. It is essential that this probability be very small.

FACTORS THAT INFLUENCE A HYPOTHESIS TEST

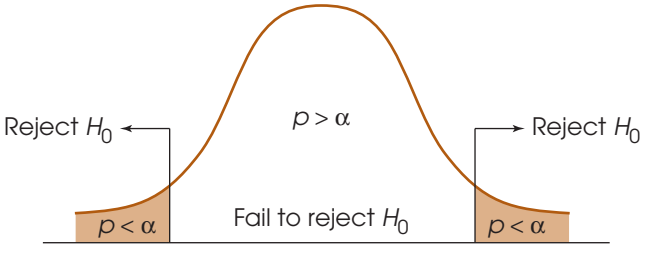
The final decision in a hypothesis test is determined by the value obtained for the z-score statistic. If the z-score is large enough to be in the critical region, then we reject the null hypothesis and conclude that there is a significant treatment effect. Otherwise, we fail to reject H_0 and conclude that the treatment does not have a significant effect. The most obvious factor influencing the size of the z-score is the difference between the sample mean and the hypothesized population mean from H_0. A big mean difference indicates that the treated sample is noticeably different from the untreated population and usually supports a conclusion that the treatment effect is significant. In addition to the mean difference, however, there are other factors that help determine whether the z-score is large enough to reject H_0. In this section we examine two factors that can influence the outcome of a hypothesis test.

1. The variability of the scores, which is measured by either the standard deviation or the variance. The variability influences the size of the standard error in the denominator of the z-score.
2. The number of scores in the sample. This value also influences the size of the standard error in the denominator.

We use the research study from Example 8.2, shown in Figure 8.7, to examine each of these factors. The study used a sample of $n = 16$ newborn rats and concluded that alcohol has a significant effect on birth weight, $z = -3.00$, $p < .05$.

FIGURE 8.9

Sample means that fall in the critical region (shaded areas) have a probability *less than* alpha ($p < \alpha$). In this case, H_0 should be rejected. Sample means that do not fall in the critical region have a probability *greater than* alpha ($p > \alpha$).



The variability of the scores In Chapter 4 (p. 124) we noted that high variability can make it very difficult to see any clear patterns in the results from a research study. In a hypothesis test, higher variability can reduce the chances of finding a significant treatment effect. For the study in Figure 8.7, the standard deviation is $\sigma = 4$. With a sample of $n = 16$, this produced a standard error of $\sigma_M = 1$ point and a significant z-score of $z = -300$. Now consider what happens if the standard deviation is increased to $\sigma = 12$. With the increased variability, the standard error becomes $\sigma_M = 12/\sqrt{16} = 3$ points. Using the same 3-points mean difference from the original example the new z-score becomes

$$z = \frac{M - \mu}{\sigma_M} = \frac{15 - 18}{3} = \frac{-3}{3} = -1.00$$

The z-score is no longer beyond the critical boundary of 1.96, so the statistical decision is to fail to reject the null hypothesis. The increased variability means that the sample data are no longer sufficient to conclude that the treatment has a significant effect. In general, increasing the variability of the scores produces a larger standard error and a smaller value (closer to zero) for the z-score. If other factors are held constant, then the larger the variability, the lower the likelihood of finding a significant treatment effect.

The number of scores in the sample The second factor that influences the outcome of a hypothesis test is the number of scores in the sample. The study in Figure 8.7 used a sample of $n = 16$ rats obtained a standard error of $\sigma_M = 4/\sqrt{16} = 1$ point and a significant z-score of $z = -3.00$. Now consider what happens if we increase the sample size to $n = 64$ rats. With $n = 64$, the standard error becomes $\sigma_M = 4/\sqrt{64} = 0.5$ points, and the z-score becomes

$$z = \frac{M - \mu}{\sigma_M} = \frac{15 - 18}{0.5} = \frac{-3}{0.5} = -6.00$$

Increasing the sample size from $n = 16$ to $n = 64$ has doubled the size of the z-score. In general, increasing the number of scores in the sample produces a smaller standard error and a larger value for the z-score. If all other factors are held constant, the larger the sample size, the greater the likelihood of finding a significant treatment effect. In simple terms, finding a 3-point treatment effect with large sample is more convincing than finding a 3-point effect with a small sample.

ASSUMPTIONS FOR HYPOTHESIS TESTS WITH *z*-SCORES

The mathematics used for a hypothesis test are based on a set of assumptions. When these assumptions are satisfied, you can be confident that the test produces a justified conclusion. However, if the assumptions are not satisfied, then the hypothesis test may be compromised. In practice, researchers are not overly concerned with the assumptions underlying a hypothesis test because the tests usually work well even when the assumptions are violated. However, you should be aware of the fundamental conditions that are associated with each type of statistical test to ensure that the test is being used appropriately. The assumptions for hypothesis tests with z-scores are summarized as follows.

Random sampling It is assumed that the participants used in the study were selected randomly. Remember, we wish to generalize our findings from the sample to the population. Therefore, the sample must be representative of the population from which it has been drawn. Random sampling helps to ensure that it is representative.

Independent observations The values in the sample must consist of *independent* observations. In everyday terms, two observations are independent if there is no consistent, predictable relationship between the first observation and the second. More precisely, two events (or observations) are independent if the occurrence of the first event has no effect on the probability of the second event. Specific examples of independence and non-independence are examined in Box 8.1. Usually, this assumption is satisfied by using a *random* sample, which also helps to ensure that the sample is representative of the population and that the results can be generalized to the population.

The value of σ is unchanged by the treatment A critical part of the *z*-score formula in a hypothesis test is the standard error, σ_M. To compute the value for the standard error, we must know the sample size (n) and the population standard deviation (σ). In a hypothesis test, however, the sample comes from an *unknown* population (see Figures 8.3 and 8.7). If the population is really unknown, it would suggest that we do not know the standard deviation and, therefore, we cannot calculate the standard error. To solve this dilemma, we have made an assumption. Specifically, we assume that the standard deviation for the unknown population (after treatment) is the same as it was for the population before treatment.

BOX 8.1 INDEPENDENT OBSERVATIONS

Independent observations are a basic requirement for nearly all hypothesis tests. The critical concern is that each observation or measurement is not influenced by any other observation or measurement. An example of independent observations is the set of outcomes obtained in a series of coin tosses. Assuming that the coin is balanced, each toss has a 50–50 chance of coming up either heads or tails. More important, each toss is *independent* of the tosses that came before. On the fifth toss, for example, there is a 50% chance of heads no matter what happened on the previous four tosses; the coin does not remember what happened earlier and is not influenced by the past. (*Note:* Many people fail to believe in the independence of events. For example, after a series of four tails in a row, it is tempting to think that the probability of heads must increase because the coin is overdue to come up heads. This is a mistake, called the "gambler's fallacy." Remember that the coin does not know what happened on the preceding tosses and cannot be influenced by previous outcomes.)

In most research situations, the requirement for independent observations is satisfied by using a random sample of separate, unrelated individuals. Thus, the measurement obtained for each individual is not influenced by other participants in the study. The following two situations demonstrate circumstances in which the observations are *not* independent.

1. A researcher is interested in examining television preferences for children. To obtain a sample of $n = 20$ children, the researcher selects 4 children from family A, 3 children from family B, 5 children from family C, 2 children from family D, and 6 children from family E.

   It should be obvious that the researcher does *not* have 20 independent observations. Within each family, the children probably share television preference (at least, they watch the same shows). Thus, the response, for each child is likely to be related to the responses of his or her siblings.
2. The principle of independent observations is violated if the sample is obtained using *sampling without replacement.* For example, if you are selecting from a group of 20 potential participants, each individual has a 1 in 20 chance of being selected first. After the first person is selected, however, there are only 19 people remaining and the probability of being selected changes to 1 in 19. Because the probability of the second selection depends on the first, the two selections are not independent.

Actually, this assumption is the consequence of a more general assumption that is part of many statistical procedures. This general assumption states that the effect of the treatment is to add a constant amount to (or subtract a constant amount from) every score in the population. You should recall that adding (or subtracting) a constant changes the mean but has no effect on the standard deviation. You also should note that this assumption is a theoretical ideal. In actual experiments, a treatment generally does not show a perfect and consistent additive effect.

Normal sampling distribution To evaluate hypotheses with z-scores, we have used the unit normal table to identify the critical region. This table can be used only if the distribution of sample means is normal.

LEARNING CHECK

1. After years of teaching driver's education, an instructor knows that students hit an average of $\mu = 10.5$ orange cones while driving the obstacle course in their final exam. The distribution of run-over cones is approximately normal with a standard deviation of $\sigma = 4.8$. To test a theory about text messaging and driving, the instructor recruits a sample of $n = 16$ student drivers to attempt the obstacle course while sending a text message. The individuals in this sample hit an average of $M = 15.9$ cones.
   a. Do the data indicate that texting has a significant effect on driving? Test with $\alpha = .01$.
   b. Write a sentence describing the outcome of the hypothesis test as it would appear in a research report.
2. In a research report, the term *significant* is used when the null hypothesis is rejected. (True or false?)
3. In a research report, the results of a hypothesis test include the phrase "$z = 3.15$, $p < .01$." This means that the test failed to reject the null hypothesis. (True or false?)
4. If other factors are held constant, increasing the size of the sample increases the likelihood of rejecting the null hypothesis. (True or false?)
5. If other factors are held constant, are you more likely to reject the null hypothesis with a standard deviation of $\sigma = 2$ or with $\sigma = 10$?

ANSWERS

1. a. With $\alpha = .01$, the critical region consists of z-scores in the tails beyond $z = \pm 2.58$. For these data, the standard error is 1.2 and $z = 4.50$. Reject the null hypothesis and conclude that texting has a significant effect on driving.
   b. Texting while driving had a significant effect on the number of cones hit by the participants, $z = 4.50$, $p < .01$.
2. True.
3. False. The probability is *less than* .01, which means it is very unlikely that the result occurred without any treatment effect. In this case, the data are in the critical region, and H_0 is rejected.
4. True. A larger sample produces a smaller standard error, which leads to a larger z-score.
5. $\sigma = 2$. A smaller standard deviation produces a smaller standard error, which leads to larger z-score.

8.4 DIRECTIONAL (ONE-TAILED) HYPOTHESIS TESTS

The hypothesis-testing procedure presented in Section 8.3 is the standard, or *two-tailed,* test format. The term *two-tailed* comes from the fact that the critical region is divided between the two tails of the distribution. This format is by far the most widely accepted procedure for hypothesis testing. Nonetheless, there is an alternative that is discussed in this section.

Usually a researcher begins an experiment with a specific prediction about the direction of the treatment effect. For example, a special training program is expected to *increase* student performance, or alcohol consumption is expected to *slow* reaction times. In these situations, it is possible to state the statistical hypotheses in a manner that incorporates the directional prediction into the statement of H_0 and H_1. The result is a directional test, or what commonly is called a *one-tailed test*.

DEFINITION

In a **directional hypothesis test,** or a **one-tailed test,** the statistical hypotheses (H_0 and H_1) specify either an increase or a decrease in the population mean. That is, they make a statement about the direction of the effect.

The following example demonstrates the elements of a one-tailed hypothesis test.

EXAMPLE 8.3

Earlier, in Example 8.1, we discussed a research study that examined the effect of antioxidants (such as those found in blueberries) on the cognitive skills of elderly adults. In the study, each participant in a sample of $n = 25$ received a blueberry supplement every day for 6 months and then was given a standardized test to measure cognitive skill. For the general population of elderly adults (without any supplement), the test scores form a normal distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$. For this example, the expected effect is that the blueberry supplement will improve cognitive performance. If the researcher obtains a sample mean of $M = 87$ for the $n = 25$ participants, is the result sufficient to conclude that the supplement really works?

THE HYPOTHESIS FOR A DIRECTIONAL TEST

Because a specific direction is expected for the treatment effect, it is possible for the researcher to perform a directional test. The first step (and the most critical step) is to state the statistical hypotheses. Remember that the null hypothesis states that there is no treatment effect and the alternative hypothesis says that there is an effect. For this example, the predicted effect is that the blueberry supplement will increase test scores. Thus, the two hypotheses would state:

H_0: Test scores are not increased. (The treatment does not work.)

H_1: Test scores are increased. (The treatment works as predicted.)

To express directional hypotheses in symbols, it usually is easier to begin with the alternative hypothesis (H_1). Again, we know that the general population has an average test score of $\mu = 80$, and H_1 states that test scores will be increased by the blueberry supplement. Therefore, expressed in symbols, H_1 states,

H_1: $\mu > 80$ (With the supplement, the average score is greater than 80.)

The null hypothesis states that the supplement does not increase scores. In symbols,

H_0: $\mu \leq 80$ (With the supplement, the average score is not greater than 80.)

Note again that the two hypotheses are mutually exclusive and cover all of the possibilities.

THE CRITICAL REGION FOR DIRECTIONAL TESTS

If the prediction is that the treatment will produce a *decrease* in scores, then the critical region is located entirely in the left-hand tail of the distribution.

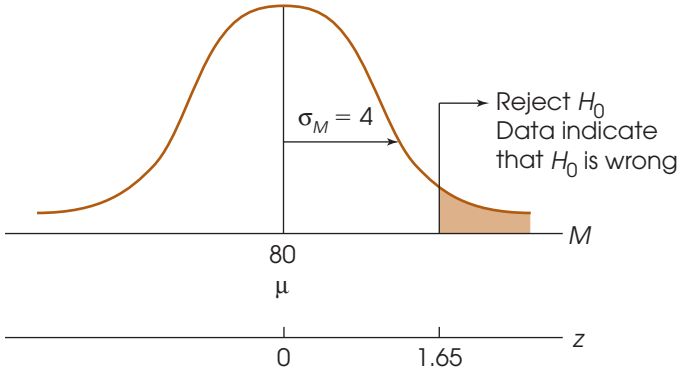
The critical region is defined by sample outcomes that are very unlikely to occur if the null hypothesis is true (that is, if the treatment has no effect). Earlier (p. 238), we noted that the critical region can also be defined in terms of sample values that provide *convincing evidence* that the treatment really does have an effect. For a directional test, the concept of "convincing evidence" is the simplest way to determine the location of the critical region. We begin with all of the possible sample means that could be obtained if the null hypothesis is true. This is the distribution of sample means and it is normal (because the population of test scores is normal), has an expected value of $\mu = 80$ (from H_0), and, for a sample of $n = 25$, has a standard error of $\sigma_M = 20/\sqrt{25} = 4$. The distribution is shown in Figure 8.10.

For this example, the treatment is expected to increase test scores. If untreated adults average $\mu = 80$ on the test, then a sample mean that is substantially more than 80 would provide convincing evidence that the treatment worked. Thus, the critical region is located entirely in the right-hand tail of the distribution corresponding to sample means much greater than $\mu = 80$ (see Figure 8.10). Because the critical region is contained in one tail of the distribution, a directional test is commonly called a *one-tailed* test. Also note that the proportion specified by the alpha level is not divided between two tails, but rather is contained entirely in one tail. Using $\alpha = .05$ for example, the whole 5% is located in one tail. In this case, the z-score boundary for the critical region is $z = 1.65$, which is obtained by looking up a proportion of .05 in column C (the tail) of the unit normal table.

Notice that a directional (one-tailed) test requires two changes in the step-by-step hypothesis-testing procedure.

1. In the first step of the hypothesis test, the directional prediction is incorporated into the statement of the hypotheses.
2. In the second step of the process, the critical region is located entirely in one tail of the distribution.

FIGURE 8.10
Critical region for Example 8.3.



After these two changes, the remainder of a one-tailed test proceeds exactly the same as a regular two-tailed test. Specifically, you calculate the z-score statistic and then make a decision about H_0 depending on whether the z-score is in the critical region.

For this example, the researcher obtained a mean of $M = 87$ for the 25 participants who received the blueberry supplement. This sample mean corresponds to a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{87 - 80}{4} = \frac{7}{4} = 1.75$$

A z-score of $z = 1.75$ is in the critical region for a one-tailed test (see Figure 8.10). This is a very unlikely outcome if H_0 is true. Therefore, we reject the null hypothesis and conclude that the blueberry supplement produces a significant increase in cognitive performance scores. In the literature, this result would be reported as follows:

> The supplement produced a significant increase in scores, $z = 1.75$, $p < .05$, one tailed.

Note that the report clearly acknowledges that a one-tailed test was used.

COMPARISON OF ONE-TAILED VERSUS TWO-TAILED TESTS

The general goal of hypothesis testing is to determine whether a particular treatment has any effect on a population. The test is performed by selecting a sample, administering the treatment to the sample, and then comparing the result with the original population. If the treated sample is noticeably different from the original population, then we conclude that the treatment has an effect, and we reject H_0. On the other hand, if the treated sample is still similar to the original population, then we conclude that there is no convincing evidence for a treatment effect, and we fail to reject H_0. The critical factor in this decision is the *size of the difference* between the treated sample and the original population. A large difference is evidence that the treatment worked; a small difference is not sufficient to say that the treatment had any effect.

The major distinction between one-tailed and two-tailed tests is the criteria that they use for rejecting H_0. A one-tailed test allows you to reject the null hypothesis when the difference between the sample and the population is relatively small, provided that the difference is in the specified direction. A two-tailed test, on the other hand, requires a relatively large difference independent of direction. This point is illustrated in the following example.

EXAMPLE 8.4

Consider again the one-tailed test evaluating the effect of an antioxidant supplement. If we had used a standard two-tailed test, the hypotheses would be

H_0: $\mu = 80$ (The supplement has no effect on test scores.)

H_1: $\mu \neq 80$ (The supplement does have an effect on test scores.)

For a two-tailed test with $\alpha = .05$, the critical region consists of z-scores beyond ± 1.96. The data from Example 8.3 produced a sample mean of $M = 87$ and $z = 1.75$. For the two-tailed test, this z-score is not in the critical region, and we conclude that the supplement does not have a significant effect.

With the two-tailed test in Example 8.4, the 7-point difference between the sample mean and the hypothesized population mean ($M = 87$ and $\mu = 80$) is not big enough to reject the null hypothesis. However, with the one-tailed test introduced in

Example 8.3, the same 7-point difference is large enough to reject H_0 and conclude that the treatment had a significant effect.

All researchers agree that one-tailed tests are different from two-tailed tests. However, there are several ways to interpret the difference. One group of researchers contends that a two-tailed test is more rigorous and, therefore, more convincing than a one-tailed test. Remember that the two-tailed test demands more evidence to reject H_0 and thus provides a stronger demonstration that a treatment effect has occurred.

Other researchers feel that one-tailed tests are preferable because they are more sensitive. That is, a relatively small treatment effect may be significant with a one-tailed test but fail to reach significance with a two-tailed test. Also, there is the argument that one-tailed tests are more precise because they test hypotheses about a specific directional effect instead of an indefinite hypothesis about a general effect.

In general, two-tailed tests should be used in research situations when there is no strong directional expectation or when there are two competing predictions. For example, a two-tailed test would be appropriate for a study in which one theory predicts an increase in scores but another theory predicts a decrease. One-tailed tests should be used only in situations in which the directional prediction is made before the research is conducted and there is a strong justification for making the directional prediction. In particular, if a two-tailed test fails to reach significance, you should never follow up with a one-tailed test as a second attempt to salvage a significant result for the same data.

LEARNING CHECK

1. If a researcher predicts that a treatment will increase scores, then the critical region for a one-tailed test would be located in the right-hand tail of the distribution. (True or false?)
2. If the sample data are sufficient to reject the null hypothesis for a one-tailed test, then the same data would also reject H_0 for a two-tailed test. (True or false?)
3. A researcher obtains $z = 2.43$ for a hypothesis test. Using $\alpha = .01$, the researcher should reject the null hypothesis for a one-tailed test but fail to reject for a two-tailed test. (True or false?)

ANSWERS

1. True. A large sample mean, in the right-hand tail, would indicate that the treatment worked as predicted.
2. False. Because a two-tailed test requires a larger mean difference, it is possible for a sample to be significant for a one-tailed test but not for a two-tailed test.
3. True. The one-tailed critical value is $z = 2.33$ and the two-tailed value is $z = 2.58$.

8.5 CONCERNS ABOUT HYPOTHESIS TESTING: MEASURING EFFECT SIZE

Although hypothesis testing is the most commonly used technique for evaluating and interpreting research data, a number of scientists have expressed a variety of concerns about the hypothesis testing procedure (for example, see Loftus, 1996; Hunter, 1997; and Killeen, 2005).

There are two serious limitations with using a hypothesis test to establish the significance of a treatment effect. The first concern is that the focus of a hypothesis test is on the data rather than the hypothesis. Specifically, when the null hypothesis is rejected,

we are actually making a strong probability statement about the sample data, not about the null hypothesis. A significant result permits the following conclusion: "This specific sample mean is very unlikely ($p < .05$) if the null hypothesis is true." Note that the conclusion does not make any definite statement about the probability of the null hypothesis being true or false. The fact that the data are very unlikely *suggests* that the null hypothesis is also very unlikely, but we do not have any solid grounds for making a probability statement about the null hypothesis. Specifically, you cannot conclude that the probability of the null hypothesis being true is less than 5% simply because you rejected the null hypothesis with $\alpha = .05$ (see Box 8.2).

A second concern is that demonstrating a *significant* treatment effect does not necessarily indicate a *substantial* treatment effect. In particular, statistical significance

BOX 8.2 A FLAW IN THE LOGIC OF HYPOTHESIS TESTING

Suppose that you do a hypothesis test and reject the null hypothesis with $\alpha = .05$. Can you conclude that there is a 5% probability that you are making a Type I error? Can you also conclude that there is a 95% probability that your decision is correct and the treatment does have an effect? For both questions, the answer is no.

The problem is that the probabilities for a hypothesis test are well defined only when the null hypothesis is true. Specifically, a hypothesis test using $\alpha = .05$ is structured so that the error rate is $p < .05$ and the accuracy rate is $p \geq .95$ *if the null hypothesis is true*. If H_0 is false, however, these probabilities start to fall apart. When there is a treatment effect (H_0 is false), the probability that a hypothesis test will detect it and reject H_0 depends on a variety of factors. For example, if the treatment effect is very small, then a hypothesis test is unlikely to detect it. With a large treatment effect, the hypothesis test is more likely to detect it and the probability of rejecting H_0 increases. Thus, whenever there is a treatment effect (H_0 is false), it becomes impossible to define precisely the probability of rejecting the null hypothesis.

Most researchers begin research studies believing that there is a good likelihood that the null hypothesis is false and there really is a treatment effect. They are hoping that the study will provide evidence of the effect so they can convince their colleagues. Thus, most research begins with some probability that the null hypothesis is false. For the sake of argument, let's assume that there is an 80% probability that the null hypothesis is true.

p(there is no treatment effect—H_0 is true) = 0.80 and
p(there is a treatment effect—H_0 is false) = 0.20

In this situation, suppose that 125 researchers are all doing hypothesis tests with $\alpha = .05$. Of these researchers, 80% ($n = 100$) are testing a true H_0. For these researchers, the probability of rejecting the null hypothesis (and making a Type I error) is $\alpha = .05$. Therefore, the 100 hypothesis tests for this group should produce, at most, 5 tests that reject H_0.

Meanwhile, the other 20% of the researchers ($n = 25$) are testing a false null hypothesis. For this group, the probability of rejecting the null hypothesis is unknown. For the sake of argument, however, let's assume that the probability of detecting the treatment effect and correctly rejecting H_0 is 60%. This means that the 25 hypothesis tests should result in 15 tests (60%) that reject H_0 and 10 that fail to reject H_0.

Notice that there could be as many as 20 hypothesis tests that reject the null hypothesis (5 from the first group and 15 from the second group). Thus, a total of 20 researchers will find a statistically significant effect. Of these 20 "significant" results, however, the 5 from the first group are making a Type I error. In this case, the probability of a Type I error is 5 out of 20, or $p = 5/20 = .25$, which is five times greater than the alpha level of .05.

Based on this kind of argument, many scientists suspect that a large number of the results and conclusions published in research journals are simply wrong. Specifically, the Type I error rate in published research is almost certainly higher than the alpha levels used in the hypothesis tests that support the results (Siegfried, 2010).

does not provide any real information about the absolute size of a treatment effect. Instead, the hypothesis test has simply established that the results obtained in the research study are very unlikely to have occurred if there is no treatment effect. The hypothesis test reaches this conclusion by (1) calculating the standard error, which measures how much difference is reasonable to expect between M and μ, and (2) demonstrating that the obtained mean difference is substantially bigger than the standard error.

Notice that the test is making a *relative* comparison: the size of the treatment effect is being evaluated relative to the standard error. If the standard error is very small, then the treatment effect can also be very small and still be large enough to be significant. Thus, a significant effect does not necessarily mean a big effect.

The idea that a hypothesis test evaluates the relative size of a treatment effect, rather than the absolute size, is illustrated in the following example.

EXAMPLE 8.5

We begin with a population of scores that forms a normal distribution with $\mu = 50$ and $\sigma = 10$. A sample is selected from the population and a treatment is administered to the sample. After treatment, the sample mean is found to be $M = 51$. Does this sample provide evidence of a statistically significant treatment effect?

Although there is only a 1-point difference between the sample mean and the original population mean, the difference may be enough to be significant. In particular, the outcome of the hypothesis test depends on the sample size.

For example, with a sample of $n = 25$ the standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2.00$$

and the z-score for $M = 51$ is

$$z = \frac{M - \mu}{\sigma_M} = \frac{51 - 50}{2} = \frac{1}{2} = 0.50$$

This z-score fails to reach the critical boundary of $z = 1.96$, so we fail to reject the null hypothesis. In this case, the 1-point difference between M and μ is not significant because it is being evaluated relative to a standard error of 2 points.

Now consider the outcome with a sample of $n = 400$. With a larger sample, the standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{400}} = \frac{10}{20} = 0.50$$

and the z-score for $M = 51$ is

$$z = \frac{M - \mu}{\sigma_M} = \frac{51 - 50}{0.5} = \frac{1}{0.5} = 2.00$$

Now the z-score is beyond the 1.96 boundary, so we reject the null hypothesis and conclude that there is a significant effect. In this case, the 1-point difference between M and μ is considered statistically significant because it is being evaluated relative to a standard error of only 0.5 points.

The point of Example 8.5 is that a small treatment effect can still be statistically significant. If the sample size is large enough, any treatment effect, no matter how small, can be enough for us to reject the null hypothesis.

MEASURING EFFECT SIZE

As noted in the previous section, one concern with hypothesis testing is that a hypothesis test does not really evaluate the absolute size of a treatment effect. To correct this problem, it is recommended that whenever researchers report a statistically significant effect, they also provide a report of the effect size (see the guidelines presented by L. Wilkinson and the APA Task Force on Statistical Inference, 1999). Therefore, as we present different hypothesis tests we also present different options for measuring and reporting *effect size.*

DEFINITION

A measure of **effect size** is intended to provide a measurement of the absolute magnitude of a treatment effect, independent of the size of the sample(s) being used.

One of the simplest and most direct methods for measuring effect size is *Cohen's d.* Cohen (1988) recommended that effect size can be standardized by measuring the mean difference in terms of the standard deviation. The resulting measure of effect size is computed as

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{\mu_{\text{treatment}} - \mu_{\text{no treatment}}}{\sigma} \tag{8.1}$$

For the *z*-score hypothesis test, the mean difference is determined by the difference between the population mean before treatment and the population mean after treatment. However, the population mean after treatment is unknown. Therefore, we must use the mean for the treated sample in its place. Remember, the sample mean is expected to be representative of the population mean and provides the best measure of the treatment effect. Thus, the actual calculations are really estimating the value of Cohen's *d* as follows:

Cohen's *d* measures the distance between two means and is typically reported as a positive number even when the formula produces a negative value.

$$\text{estimated Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{M_{\text{treatment}} - \mu_{\text{no treatment}}}{\sigma} \tag{8.2}$$

The standard deviation is included in the calculation to standardize the size of the mean difference in much the same way that *z*-scores standardize locations in a distribution. For example, a 15-point mean difference can be a relatively large treatment effect or a relatively small effect depending on the size of the standard deviation. This phenomenon is demonstrated in Figure 8.11. The top portion of the figure (part a) shows the results of a treatment that produces a 15-point mean difference in SAT scores; before treatment, the average SAT score is $\mu = 500$, and after treatment the average is 515. Notice that the standard deviation for SAT scores is $\sigma = 100$, so the 15-point difference appears to be small. For this example, Cohen's *d* is

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{15}{100} = 0.15$$

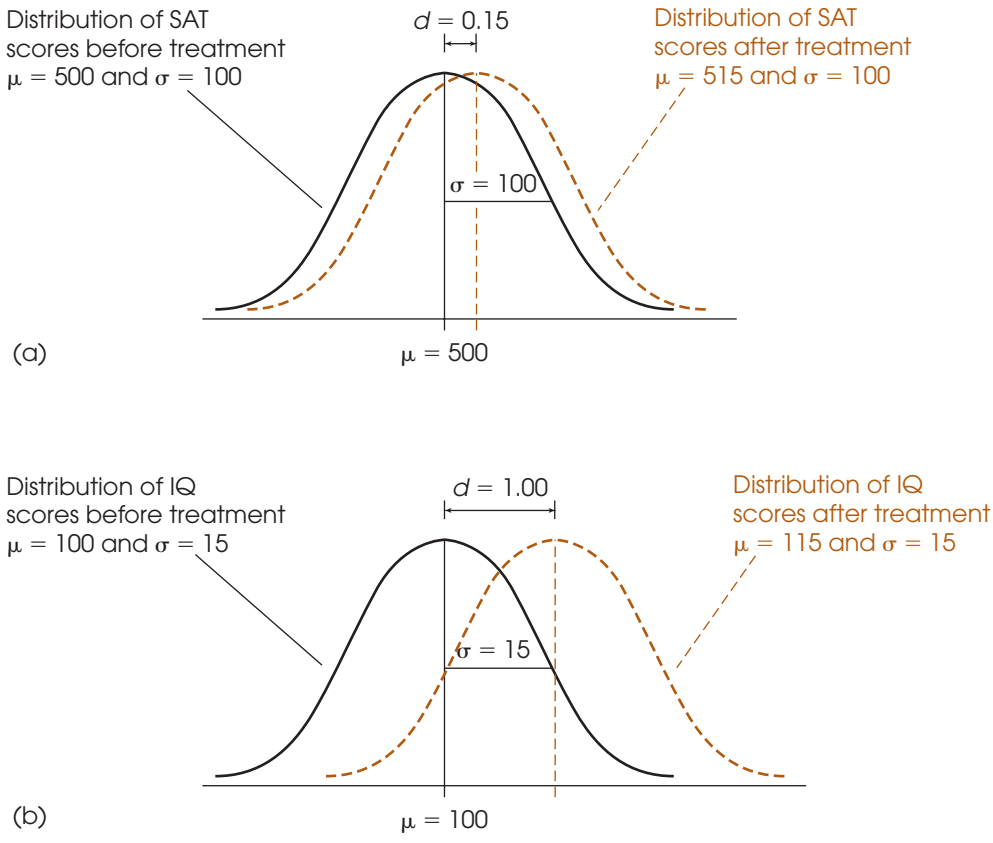


FIGURE 8.11

The appearance of a 15-point treatment effect in two different situations. In part (a), the standard deviation is $\sigma = 100$ and the 15-point effect is relatively small. In part (b), the standard deviation is $\sigma = 15$ and the 15-point effect is relatively large. Cohen's *d* uses the standard deviation to help measure effect size.

Now consider the treatment effect shown in Figure 8.11(b). This time, the treatment produces a 15-point mean difference in IQ scores; before treatment the average IQ is 100, and after treatment the average is 115. Because IQ scores have a standard deviation of $\sigma = 15$, the 15-point mean difference now appears to be large. For this example, Cohen's *d* is

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{15}{15} = 1.00$$

Notice that Cohen's *d* measures the size of the treatment effect in terms of the standard deviation. For example, a value of $d = 0.50$ indicates that the treatment changed the mean by half of a standard deviation; similarly, a value of $d = 1.00$ indicates that the size of the treatment effect is equal to one whole standard deviation. (See Box 8.3.)

Cohen (1988) also suggested criteria for evaluating the size of a treatment effect as shown in Table 8.2.

BOX 8.3 OVERLAPPING DISTRIBUTIONS

Figure 8.11(b) shows the results of a treatment with a Cohen's d of 1.00; that is, the effect of the treatment is to increase the mean by one full standard deviation. According to the guidelines in Table 8.2, a value of $d = 1.00$ is considered a large treatment effect. However, looking at the figure, you may get the impression that there really isn't that much difference between the distribution before treatment and the distribution after treatment. In particular, there is substantial overlap between the two distributions, so that many of the individuals who receive the treatment are not any different from the individuals who do not receive the treatment.

The overlap between distributions is a basic fact of life in most research situations; it is extremely rare for the scores after treatment to be *completely different* (no overlap) from the scores before treatment. Consider, for example, children's heights at different ages. Everyone knows that 8-year-old children are taller than 6-year-old children; on average, the difference is 3 or 4 inches. However, this does not mean that all 8-year-old children are taller than all 6-year-old children. In fact, there is considerable overlap between the two distributions, so that the tallest among the 6-year-old children are actually taller than most 8-year-old children. In fact, the height distributions for the two age groups would look a lot like the two distributions in Figure 8.10(b). Although there is a clear *mean difference* between the two distributions, there still can be substantial overlap.

Cohen's d measures the degree of separation between two distributions, and a separation of one standard deviation ($d = 1.00$) represents a large difference. Eight-year-old children really are bigger than 6-year-old children.

TABLE 8.2
Evaluating effect size with Cohen's *d*.

Magnitude of *d*	Evaluation of Effect Size
$d = 0.2$	Small effect (mean difference around 0.2 standard deviation)
$d = 0.5$	Medium effect (mean difference around 0.5 standard deviation)
$d = 0.8$	Large effect (mean difference around 0.8 standard deviation)

As one final demonstration of Cohen's d, consider the two hypothesis tests in Example 8.5. For each test, the original population had a mean of $\mu = 50$ with a standard deviation of $\sigma = 10$. For each test, the mean for the treated sample was $M = 51$. Although one test used a sample of $n = 25$ and the other test used a sample of $n = 400$, the sample size is not considered when computing Cohen's d. Therefore, both of the hypothesis tests would produce the same value:

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{15}{15} = 1.00$$

Notice that Cohen's d simply describes the size of the treatment effect and is not influenced by the number of scores in the sample. For both hypothesis tests, the original population mean was $\mu = 50$ and, after treatment, the sample mean was $M = 51$. Thus, treatment appears to have increased the scores by 1 point, which is equal to one-tenth of a standard deviation (Cohen's $d = 0.1$).

LEARNING CHECK

1. **a.** How does increasing sample size influence the outcome of a hypothesis test?
   b. How does increasing sample size influence the value of Cohen's *d*?
2. A researcher selects a sample from a population with $\mu = 45$ and $\sigma = 8$. A treatment is administered to the sample and, after treatment, the sample mean is found to be $M = 47$. Compute Cohen's *d* to measure the size of the treatment effect.

ANSWERS

1. **a.** Increasing sample size increases the likelihood of rejecting the null hypothesis.
   b. Cohen's *d* is not influenced at all by the sample size.
2. $d = 2/8 = 0.25$

8.6 STATISTICAL POWER

Instead of measuring effect size directly, an alternative approach to determining the size or strength of a treatment effect is to measure the power of the statistical test. The *power* of a test is defined as the probability that the test will reject the null hypothesis if the treatment really has an effect.

DEFINITION

The **power** of a statistical test is the probability that the test will correctly reject a false null hypothesis. That is, power is the probability that the test will identify a treatment effect if one really exists.

Whenever a treatment has an effect, there are only two possible outcomes for a hypothesis test: either fail to reject H_0 or reject H_0. Because there are only two possible outcomes, the probability for the first and the probability for the second must add up to 1.00. The first outcome, failing to reject H_0 when there is a real effect, was defined earlier (p. 245) as a Type II error with a probability identified as $p = \beta$. Therefore, the second outcome must have a probability of $1 - \beta$. However, the second outcome, rejecting H_0 when there is a real effect, is the power of the test. Thus, the power of a hypothesis test is equal to $1 - \beta$. In the examples that follow, we demonstrate the calculation of power for a hypothesis test; that is, the probability that the test will correctly reject the null hypothesis. At the same time, however, we are computing the probability that the test will result in a Type II error. For example, if the power of the test is 70% ($1 - \beta$) then the probability of a Type II error must be 30% (β).

Researchers typically calculate power as a means of determining whether a research study is likely to be successful. Thus, researchers usually calculate the power of a hypothesis test *before* they actually conduct the research study. In this way, they can determine the probability that the results will be significant (reject H_0) before investing time and effort in the actual research. To calculate power, however, it is first necessary to make assumptions about a variety of factors that influence the outcome of a hypothesis test. Factors such as the sample size, the size of the treatment effect, and the value chosen for the alpha level can all influence a hypothesis test. The following example demonstrates the calculation of power for a specific research situation.

EXAMPLE 8.6

We start with a normal-shaped population with a mean of $\mu = 80$ and a standard deviation of $\sigma = 10$. A researcher plans to select a sample of $n = 25$ individuals from

this population and administer a treatment to each individual. It is expected that the treatment will have an 8-point effect; that is, the treatment will add 8 points to each individual's score.

Figure 8.12 shows the original population distribution and two possible outcomes:

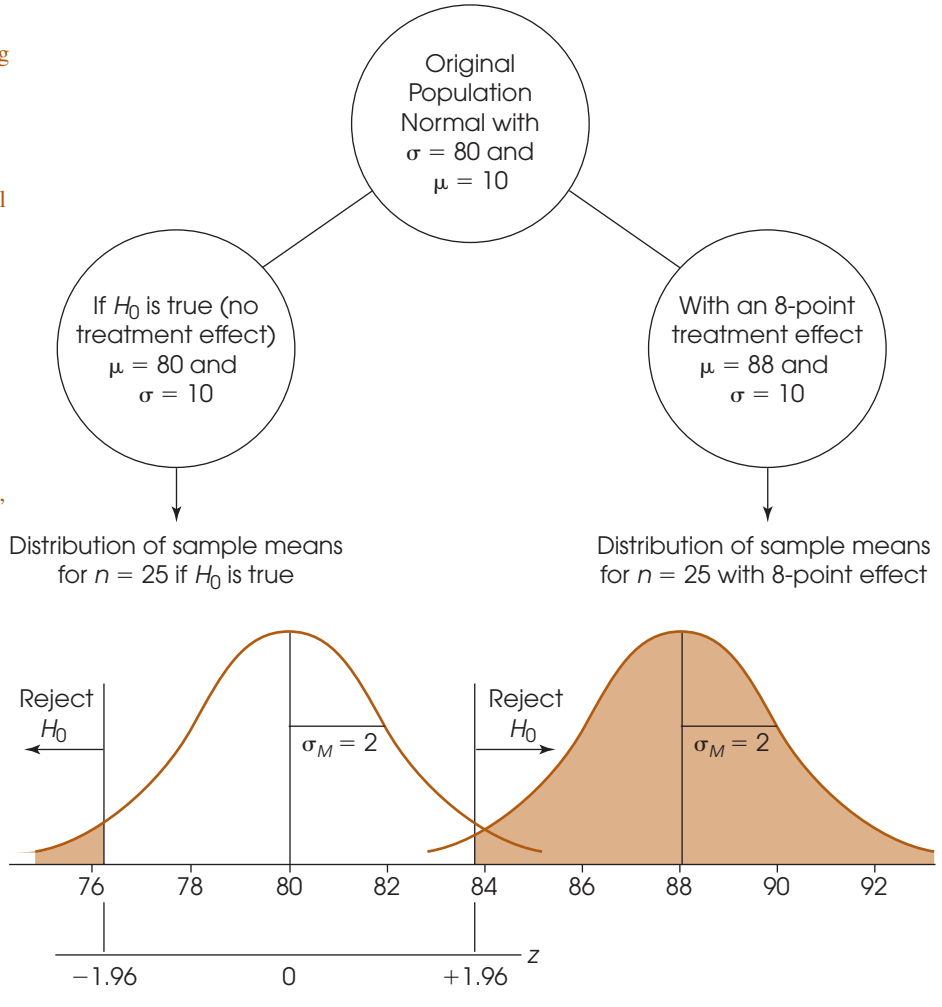
1. If the null hypothesis is true and there is no treatment effect.
2. If the researcher's expectation is correct and there is an 8-point effect.

The left-hand side of the figure shows what should happen according to the null hypothesis. In this case, the treatment has no effect and the population mean is still $\mu = 80$. On the right-hand side of the figure we show what would happen if the treatment has an 8-point effect. If the treatment adds 8 points to each person's score, then the population mean after treatment increases to $\mu = 88$.

Beneath each of the two populations, Figure 8.12 shows the distribution of sample means for $n = 25$. According to the null hypothesis, the sample means are centered

FIGURE 8.12

A demonstration of measuring power for a hypothesis test. The left-hand side shows the distribution of sample means that would occur if the null hypothesis is true. The critical region is defined for this distribution. The right-hand side shows the distribution of sample means that would be obtained if there were an 8-point treatment effect. Notice that if there is an 8-point effect, essentially all of the sample means would be in the critical region. Thus, the probability of rejecting H_0 (the power of the test) would be nearly 100% for an 8-point treatment effect.



around $\mu = 80$. With an 8-point treatment effect, the sample means are centered around $\mu = 88$. Both distributions have a standard error of

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

Notice that the distribution on the left shows all of the possible sample means if the null hypothesis is true. This is the distribution we use to locate the critical region for the hypothesis test. Using $\alpha = .05$, the critical region consists of extreme values in this distribution, specifically sample means beyond $z = 1.96$ or $z = -1.96$. These values are shown in Figure 8.12, and we have shaded all of the sample means located in the critical region.

Now turn your attention to the distribution on the right, which shows all of the possible sample means if there is an 8-point treatment effect. Notice that most of these sample means are located beyond the $z = 1.96$ boundary. This means that, if there is an 8-point treatment effect, you are almost guaranteed to obtain a sample mean in the critical region and reject the null hypothesis. Thus, the power of the test (the probability of rejecting H_0) is close to 100% if there is an 8-point treatment effect.

To calculate the exact value for the power of the test we must determine what portion of the distribution on the right-hand side is shaded. Thus, we must locate the exact boundary for the critical region, then find the probability value in the unit normal table. For the distribution on the left-hand side, the critical boundary of $z = +1.96$ corresponds to a location that is above $\mu = 80$ by a distance equal to

$$1.96\mu_M = 1.96(2) = 3.92 \text{ points}$$

Thus, the critical boundary of $z = +1.96$ corresponds to a sample mean of $M = 80 + 3.92 = 83.92$. Any sample mean greater than $M = 83.92$ is in the critical region and would lead to rejecting the null hypothesis. Next, we determine what proportion of the treated samples are greater than $M = 83.92$. For the treated distribution (right-hand side), the population mean is $\mu = 88$ and a sample mean of $M = 83.92$ corresponds to a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{83.92 - 88}{2} = \frac{-4.08}{2} = -2.04$$

Finally, look up $z = -2.04$ in the unit normal table and determine that the shaded area ($z > -2.04$) corresponds to $p = 0.9793$ (or 97.93%). Thus, if the treatment has an 8-point effect, 97.93% of all the possible sample means will be in the critical region and we will reject the null hypothesis. In other words, the power of the test is 97.93%. In practical terms, this means that the research study is almost guaranteed to be successful. If the researcher selects a sample of $n = 25$ individuals, and if the treatment really does have an 8-point effect, then 97.93% of the time the hypothesis test will conclude that there is a significant effect.

POWER AND EFFECT SIZE

Logically, it should be clear that power and effect size are related. Figure 8.12 shows the calculation of power for an 8-point treatment effect. Now consider what would happen if the treatment effect were only 4 points. With a 4-point treatment effect, the distribution on the right-hand side would shift to the left so that it is centered at $\mu = 84$.

In this new position, only about 50% of the treated sample means would be beyond the $z = 1.96$ boundary. Thus, with a 4-point treatment effect, there is only a 50% probability of selecting a sample that leads to rejecting the null hypothesis. In other words, the power of the test is only about 50% for a 4-point effect compared to nearly 98% with an 8-point effect (Example 8.6). Again, it is possible to find the z-score corresponding to the exact location of the critical boundary and to look up the probability value for power in the unit normal table. In this case, you should obtain $z = -0.04$ and the exact power of the test is $p = 0.5160$, or 51.60%.

In general, as the effect size increases, the distribution of sample means on the right-hand side moves even farther to the right so that more and more of the samples are beyond the $z = 1.96$ boundary. Thus, as the effect size increases, the probability of rejecting H_0 also increases, which means that the power of the test increases. Thus, measures of effect size such as Cohen's d and measures of power both provide an indication of the strength or magnitude of a treatment effect.

OTHER FACTORS THAT AFFECT POWER

Although the power of a hypothesis test is directly influenced by the size of the treatment effect, power is not meant to be a pure measure of effect size. Instead, power is influenced by several factors, other than effect size, that are related to the hypothesis test. Some of these factors are considered in the following section.

Sample size One factor that has a huge influence on power is the size of the sample. In Example 8.6 we demonstrated power for an 8-point treatment effect using a sample of $n = 25$. If the researcher decided to conduct the study using a sample of $n = 4$, then the power would be dramatically different. With $n = 4$, the standard error for the sample means would be

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{4}} = \frac{10}{2} = 5$$

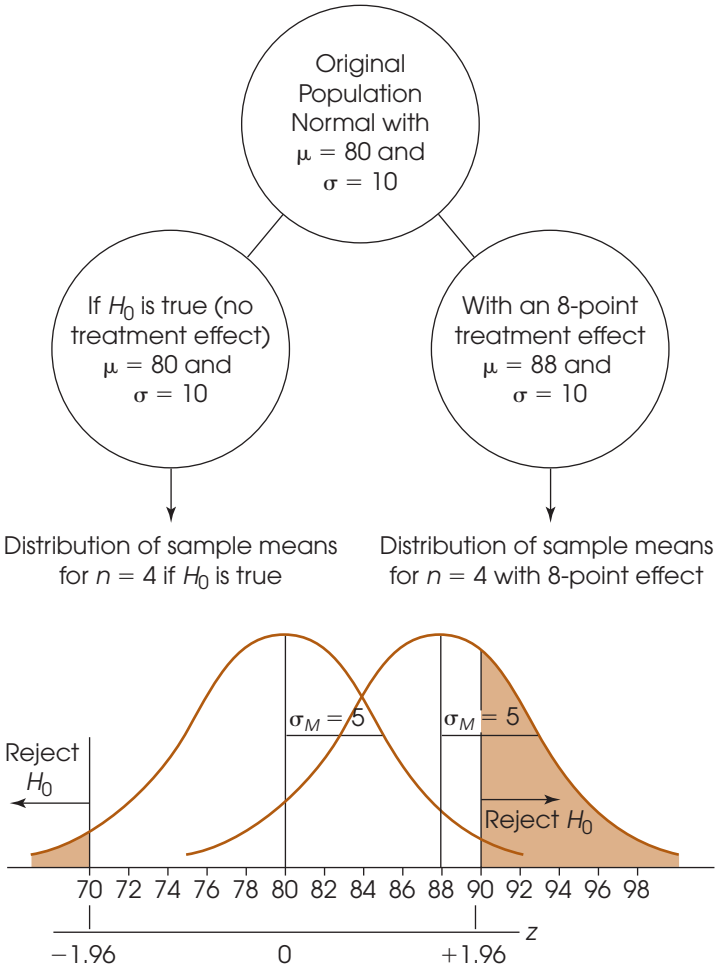
Figure 8.13 shows the two distributions of sample means with $n = 4$ and a standard error of $\sigma_M = 5$ points. Again, the distribution on the left is centered at $\mu = 80$ and shows all of the possible sample means if H_0 is true. As always, this distribution is used to locate the critical boundaries for the hypothesis test, $z = -1.96$ and $z = +1.96$. The distribution on the right is centered at $\mu = 88$ and shows all of the possible sample means if there is an 8-point treatment effect. Note that less than half of the treated sample means in the right-hand distribution are now located beyond the 1.96 boundary. Thus, with a sample of $n = 4$, there is less than a 50% probability that the hypothesis test would reject H_0, even though the treatment has an 8-point effect. Earlier, in Example 8.6, we found power equal to 97.93% for a sample of $n = 25$. However, when the sample size is reduced to $n = 4$, power decreases to less than 50%. In general, a larger sample produces greater power for a hypothesis test.

Because power is directly related to sample size, one of the primary reasons for computing power is to determine what sample size is necessary to achieve a reasonable probability for a successful research study. Before a study is conducted, researchers can compute power to determine the probability that their research will successfully reject the null hypothesis. If the probability (power) is too small, they always have the option of increasing sample size to increase power.

Alpha level Reducing the alpha level for a hypothesis test also reduces the power of the test. For example, lowering α from .05 to .01 lowers the power of the hypothesis

FIGURE 8.13

A demonstration of how sample size affects the power of a hypothesis test. As in Figure 8.12, the left-hand side shows the distribution of sample means if the null hypothesis were true. The critical region is defined for this distribution. The right-hand side shows the distribution of sample means that would be obtained if there were an 8-point treatment effect. Notice that reducing the sample size to $n = 4$ has reduced the power of the test to less than 50% compared to a power of nearly 100% with a sample of $n = 25$ in Figure 8.12.



test. The effect of reducing the alpha level can be seen by referring again to Figure 8.13. In this figure, the boundaries for the critical region are drawn using $\alpha = .05$. Specifically, the critical region on the right-hand side begins at $z = 1.96$. If α were changed to .01, the boundary would be moved farther to the right, out to $z = 2.58$. It should be clear that moving the critical boundary to the right means that a smaller portion of the treatment distribution (the distribution on the right-hand side) will be in the critical region. Thus, there would be a lower probability of rejecting the null hypothesis and a lower value for the power of the test.

One-tailed versus two-tailed tests Changing from a regular two-tailed test to a one-tailed test increases the power of the hypothesis test. Again, this effect can be seen by referring to Figure 8.13. The figure shows the boundaries for the critical region using a two-tailed test with $\alpha = .05$ so that the critical region on the right-hand side begins at $z = 1.96$. Changing to a one-tailed test would move the critical boundary to the left to a value of $z = 1.65$. Moving the boundary to the left would cause a larger proportion of the treatment distribution to be in the critical region and, therefore, would increase the power of the test.

LEARNING CHECK

1. For a particular hypothesis test, the power is .50 (50%) for a 5-point treatment effect. Will the power be greater or less than .50 for a 10-point treatment effect?
2. As the power of a test increases, what happens to the probability of a Type II error?
3. How does increasing sample size influence the power of a hypothesis test?
4. Find the exact value of the power for the hypothesis test shown in Figure 8.13.

ANSWERS

1. The hypothesis test is more likely to detect a 10-point effect, so power will be greater.
2. As power increases, the probability of a Type II error decreases.
3. Increasing sample size increases the power of a test.
4. With $n = 4$, the critical boundary of $z = 1.96$ corresponds to a sample mean of $M = 89.8$, and the exact value for power is $p = 0.3594$ or 35.945%.

SUMMARY

1. Hypothesis testing is an inferential procedure that uses the data from a sample to draw a general conclusion about a population. The procedure begins with a hypothesis about an unknown population. Then a sample is selected, and the sample data provide evidence that either supports or refutes the hypothesis.
2. In this chapter, we introduced hypothesis testing using the simple situation in which a sample mean is used to test a hypothesis about an unknown population mean; usually the mean for a population that has received a treatment. The question is to determine whether the treatment has an effect on the population mean (see Figure 8.2).
3. Hypothesis testing is structured as a four-step process that is used throughout the remainder of the book.
   a. State the null hypothesis (H_0), and select an alpha level. The null hypothesis states that there is no effect or no change. In this case, H_0 states that the mean for the treated population is the same as the mean before treatment. The alpha level, usually $\alpha = .05$ or $\alpha = .01$, provides a definition of the term *very unlikely* and determines the risk of a Type I error. Also state an alternative hypothesis (H_1), which is the exact opposite of the null hypothesis.
   b. Locate the critical region. The critical region is defined as extreme sample outcomes that would be very unlikely to occur if the null hypothesis is true. The alpha level defines "very unlikely."
   c. Collect the data, and compute the test statistic. The sample mean is transformed into a z-score by the formula

$$z = \frac{M - \mu}{\sigma_M}$$

   The value of μ is obtained from the null hypothesis. The z-score test statistic identifies the location of the sample mean in the distribution of sample means.
   d. Make a decision. If the obtained z-score is in the critical region, reject H_0 because it is very unlikely that these data would be obtained if H_0 were true. In this case, conclude that the treatment has changed the population mean. If the z-score is not in the critical region, fail to reject H_0 because the data are not significantly different from the null hypothesis. In this case, the data do not provide sufficient evidence to indicate that the treatment has had an effect.
4. Whatever decision is reached in a hypothesis test, there is always a risk of making the incorrect decision. There are two types of errors that can be committed.

   A Type I error is defined as rejecting a true H_0. This is a serious error because it results in falsely reporting a treatment effect. The risk of a Type I error is determined by the alpha level and, therefore, is under the experimenter's control.

   A Type II error is defined as the failure to reject a false H_0. In this case, the experiment fails to detect an effect

that actually occurred. The probability of a Type II error cannot be specified as a single value and depends in part on the size of the treatment effect. It is identified by the symbol β (beta).

5. When a researcher expects that a treatment will change scores in a particular direction (increase or decrease), it is possible to do a directional, or one-tailed, test. The first step in this procedure is to incorporate the directional prediction into the hypotheses. For example, if the prediction is that a treatment will increase scores, the null hypothesis says that there is no increase and the alternative hypothesis states that there is an increase. To locate the critical region, you must determine what kind of data would refute the null hypothesis by demonstrating that the treatment worked as predicted. These outcomes are located entirely in one tail of the distribution, so the entire critical region (5%, 1%, or 0.1% depending on α) will be in one tail.

6. A one-tailed test is used when there is prior justification for making a directional prediction. These *a priori* reasons may be previous reports and findings or theoretical considerations. In the absence of the *a priori* basis, a two-tailed test is appropriate. In this situation, you might be unsure of what to expect in the study, or you might be testing competing theories.

7. In addition to using a hypothesis test to evaluate the *significance* of a treatment effect, it is recommended that you also measure and report the *effect size*. One measure of effect size is Cohen's *d*, which is a standardized measure of the mean difference. Cohen's *d* is computed as

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}}$$

8. The power of a hypothesis test is defined as the probability that the test will correctly reject the null hypothesis.

9. To determine the power for a hypothesis test, you must first identify the treatment and null distributions. Also, you must specify the magnitude of the treatment effect. Next, you locate the critical region in the null distribution. The power of the hypothesis test is the portion of the treatment distribution that is located beyond the boundary (critical value) of the critical region.

10. As the size of the treatment effect increases, statistical power increases. Also, power is influenced by several factors that can be controlled by the experimenter:

    a. Increasing the alpha level increases power.
    b. A one-tailed test has greater power than a two-tailed test.
    c. A large sample results in more power than a small sample.

KEY TERMS

hypothesis test (233)
null hypothesis (236)
alternative hypothesis (236)
level of significance (237)
alpha level (237)
critical region (238)
test statistic (242)
Type I error (244)
Type II error (245)
beta (246)
significant (251)
directional test (256)
one-tailed test (256)
effect size (262)
Cohen's *d* (262)
power (265)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 8 on the book companion website. The website also provides access to a workshop titled *Hypothesis Testing,* which reviews the concept and logic of hypothesis testing.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

The statistical computer package SPSS is not structured to conduct hypothesis tests using *z*-scores. In truth, the *z*-score test presented in this chapter is rarely used in actual research situations. The problem with the *z*-score test is that it requires that you know the value of the population standard deviation, and this information is usually not available. Researchers rarely have detailed information about the populations that they wish to study. Instead, they must obtain information entirely from samples. In the following chapters we introduce new hypothesis-testing techniques that are based entirely on sample data. These new techniques are included in SPSS.

FOCUS ON PROBLEM SOLVING

1. Hypothesis testing involves a set of logical procedures and rules that enable us to make general statements about a population when all we have are sample data. This logic is reflected in the four steps that have been used throughout this chapter. Hypothesis-testing problems are easier to tackle when you learn to follow the steps.

STEP 1 State the hypotheses and set the alpha level.

STEP 2 Locate the critical region.

STEP 3 Compute the test statistic (in this case, the *z*-score) for the sample.

STEP 4 Make a decision about H_0 based on the result of step 3.

2. Students often ask, "What alpha level should I use?" Or a student may ask, "Why is an alpha of .05 used?" as opposed to something else. There is no single correct answer to either of these questions. Keep in mind that the aim of setting an alpha level in the first place: *to reduce the risk of committing a Type I error*. Therefore,

the maximum acceptable value is $\alpha = .05$. However, some researchers prefer to take even less risk and use alpha levels of .01 or smaller.

Most statistical tests are now done with computer programs that provide an exact probability (*p* value) for a Type I error. Because an exact value is available, most researchers simply report the *p* value from the computer printout rather than setting an alpha level at the beginning of the test. However, the same criterion still applies: A result is not significant unless the *p* value is less than .05.

3. Take time to consider the implications of your decision about the null hypothesis. The null hypothesis states that there is no effect. Therefore, if your decision is to reject H_0, you should conclude that the sample data provide evidence for a treatment effect. However, it is an entirely different matter if your decision is to fail to reject H_0. Remember that when you fail to reject the null hypothesis, the results are inconclusive. It is impossible to *prove* that H_0 is correct; therefore, you cannot state with certainty that "there is no effect" when H_0 is not rejected. At best, all you can state is that "there is insufficient evidence for an effect."

4. It is very important that you understand the structure of the *z*-score formula (p. 242). It will help you understand many of the other hypothesis tests that are covered later.

5. When you are doing a directional hypothesis test, read the problem carefully, and watch for key words (such as increase or decrease, raise or lower, and more or less) that tell you which direction the researcher is predicting. The predicted direction determines the alternative hypothesis (H_1) and the critical region. For example, if a treatment is expected to *increase* scores, H_1 would contain a *greater than* symbol, and the critical region would be in the tail associated with high scores.

DEMONSTRATION 8.1

HYPOTHESIS TEST WITH *Z*

A researcher begins with a known population—in this case, scores on a standardized test that are normally distributed with $\mu = 65$ and $\sigma = 15$. The researcher suspects that special training in reading skills will produce a change in the scores for the individuals in the population. Because it is not feasible to administer the treatment (the special training) to everyone in the population, a sample of $n = 25$ individuals is selected, and the treatment is given to this sample. Following treatment, the average score for this sample is $M = 70$. Is there evidence that the training has an effect on test scores?

STEP 1 **State the hypothesis and select an alpha level.** The null hypothesis states that the special training has no effect. In symbols,

H_0: $\mu = 65$ (After special training, the mean is still 65.)

The alternative hypothesis states that the treatment does have an effect.

H_1: $\mu \neq 65$ (After training, the mean is different from 65.)

At this time you also select the alpha level. For this demonstration, we will use $\alpha = .05$. Thus, there is a 5% risk of committing a Type I error if we reject H_0.

STEP 2 **Locate the critical region.** With $\alpha = .05$, the critical region consists of sample means that correspond to z-scores beyond the critical boundaries of $z = \pm 1.96$.

STEP 3 **Obtain the sample data, and compute the test statistic.** For this example, the distribution of sample means, according to the null hypothesis, is normal with an expected value of $\mu = 65$ and a standard error of

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

In this distribution, our sample mean of $M = 70$ corresponds to a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{70 - 65}{3} = \frac{5}{3} = +1.67$$

STEP 4 **Make a decision about H_0, and state the conclusion.** The z-score we obtained is not in the critical region. This indicates that our sample mean of $M = 70$ is not an extreme or unusual value to be obtained from a population with $\mu = 65$. Therefore, our statistical decision is to *fail to reject* H_0. Our conclusion for the study is that the data do not provide sufficient evidence that the special training changes test scores.

DEMONSTRATION 8.2

EFFECT SIZE USING COHEN'S *D*

We will compute Cohen's d using the research situation and the data from Demonstration 8.1. Again, the original population mean was $\mu = 65$ and, after treatment (special training), the sample mean was $M = 70$. Thus, there is a 5-point mean difference. Using the population standard deviation, $\sigma = 15$, we obtain an effect size of

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{5}{15} = 0.33$$

According to Cohen's evaluation standards (see Table 8.2), this is a medium treatment effect.

PROBLEMS

1. In the z-score formula as it is used in a hypothesis test,
   a. Explain what is measured by $M - \mu$ in the numerator.
   b. Explain what is measured by the standard error in the denominator.
2. The value of the z-score in a hypothesis test is influenced by a variety of factors. Assuming that all other variables are held constant, explain how the value of z is influenced by each of the following:
   a. Increasing the difference between the sample mean and the original population mean.
   b. Increasing the population standard deviation.
   c. Increasing the number of scores in the sample.
3. In words, define the alpha level and the critical region for a hypothesis test.
4. If the alpha level is changed from $\alpha = .05$ to $\alpha = .01$,
   a. What happens to the boundaries for the critical region?
   b. What happens to the probability of a Type I error?
5. Although there is a popular belief that herbal remedies such as ginkgo biloba and ginseng may improve

learning and memory in healthy adults, these effects are usually not supported by well-controlled research (Persson, Bringlov, Nilsson, & Nyberg, 2004). In a typical study, a researcher obtains a sample of $n = 36$ participants and has each person take the herbal supplements every day for 90 days. At the end of the 90 days, each person takes a standardized memory test. For the general population, scores from the test are normally distributed with a mean of $\mu = 80$ and a standard deviation of $\sigma = 18$. The sample of research participants had an average of $M = 84$.

a. Assuming a two-tailed test, state the null hypothesis in a sentence that includes the two variables being examined.

b. Using symbols, state the hypotheses (H_0 and H_1) for the two-tailed test.

c. Sketch the appropriate distribution, and locate the critical region for $\alpha = .05$.

d. Calculate the test statistic (z-score) for the sample.

e. What decision should be made about the null hypothesis, and what decision should be made about the effect of the herbal supplements?

6. Childhood participation in sports, cultural groups, and youth groups appears to be related to improved self-esteem for adolescents (McGee, Williams, Howden-Chapman, Martin, & Kawachi, 2006). In a representative study, a sample of $n = 100$ adolescents with a history of group participation is given a standardized self-esteem questionnaire. For the general population of adolescents, scores on this questionnaire form a normal distribution with a mean of $\mu = 40$ and a standard deviation of $\sigma = 12$. The sample of group-participation adolescents had an average of $M = 43.84$.

a. Does this sample provide enough evidence to conclude that self-esteem scores for these adolescents are significantly different from those of the general population? Use a two-tailed test with $\alpha = .01$.

b. Compute Cohen's d to measure the size of the difference.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

7. A local college requires an English composition course for all freshmen. This year they are evaluating a new online version of the course. A random sample of $n = 16$ freshmen is selected and the students are placed in the online course. At the end of the semester, all freshmen take the same English composition exam. The average score for the sample is $M = 76$. For the general population of freshmen who took the traditional lecture class, the exam scores form a normal distribution with a mean of $\mu = 80$.

a. If the final exam scores for the population have a standard deviation of $\sigma = 12$, does the sample provide enough evidence to conclude that the new online course is significantly different from the traditional class? Assume a two-tailed test with $\alpha = .05$.

b. If the population standard deviation is $\sigma = 6$, is the sample sufficient to demonstrate a significant difference? Again, assume a two-tailed test with $\alpha = .05$.

c. Comparing your answers for parts a and b, explain how the magnitude of the standard deviation influences the outcome of a hypothesis test.

8. A random sample is selected from a normal population with a mean of $\mu = 50$ and a standard deviation of $\sigma = 12$. After a treatment is administered to the individuals in the sample, the sample mean is found to be $M = 55$.

a. If the sample consists of $n = 16$ scores, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.

b. If the sample consists of $n = 36$ scores, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.

c. Comparing your answers for parts a and b, explain how the size of the sample influences the outcome of a hypothesis test.

9. A random sample of $n = 36$ scores is selected from a normal population with a mean of $\mu = 60$. After a treatment is administered to the individuals in the sample, the sample mean is found to be $M = 52$.

a. If the population standard deviation is $\sigma = 18$, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.

b. If the population standard deviation is $\sigma = 30$, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.

c. Comparing your answers for parts a and b, explain how the magnitude of the standard deviation influences the outcome of a hypothesis test.

10. Miller (2008) examined the energy drink consumption of college undergraduates and found that males use energy drinks significantly more often than females. To further investigate this phenomenon, suppose that a researcher selects a random sample of $n = 36$ male undergraduates and a sample of $n = 25$ females. On average, the males reported consuming $M = 2.45$ drinks per month and females had an average of $M = 1.28$. Assume that the overall level of consumption for college undergraduates averages $\mu = 1.85$ energy drinks per month, and that the distribution of monthly consumption scores is approximately normal with a standard deviation of $\sigma = 1.2$.

a. Did this sample of males consume significantly more energy drinks than the overall population average? Use a one-tailed test with $\alpha = .01$.

b. Did this sample of females consume significantly fewer energy drinks than the overall population average? Use a one-tailed test with $\alpha = .01$

11. A random sample is selected from a normal population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 10$. After a treatment is administered to the individuals in the sample, the sample mean is found to be $M = 42$.

a. How large a sample is necessary for this sample mean to be statistically significant? Assume a two-tailed test with $\alpha = .05$.

b. If the sample mean were $M = 41$, what sample size is needed to be significant for a two-tailed test with $\alpha = .05$?

12. There is some evidence that REM sleep, associated with dreaming, may also play a role in learning and memory processing. For example, Smith and Lapp (1991) found increased REM activity for college students during exam periods. Suppose that REM activity for a sample of $n = 16$ students during the final exam period produced an average score of $M = 143$. Regular REM activity for the college population averages $\mu = 110$ with a standard deviation of $\sigma = 50$. The population distribution is approximately normal.

a. Do the data from this sample provide evidence for a significant increase in REM activity during exams? Use a one-tailed test with $\alpha = .01$.

b. Compute Cohen's *d* to estimate the size of the effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

13. There is some evidence indicating that people with visible tattoos are viewed more negatively than people without visible tattoos (Resenhoeft, Villa, & Wiseman, 2008). In a similar study, a researcher first obtained overall ratings of attractiveness for a woman with no tattoos shown in a color photograph. On a 7-point scale, the woman received an average rating of $\mu = 4.9$, and the distribution of ratings was normal with a standard deviation of $\sigma = 0.84$. The researcher then modified the photo by adding a tattoo of a butterfly on the woman's left arm. The modified photo was then shown to a sample of $n = 16$ students at a local community college and the students used the same 7-point scale to rate the attractiveness of the woman. The average score for the photo with the tattoo was $M = 4.2$.

a. Do the data indicate a significant difference in rated attractiveness when the woman appeared to have a tattoo? Use a two-tailed test with $\alpha = .05$.

b. Compute Cohen's *d* to measure the size of the effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

14. A psychologist is investigating the hypothesis that children who grow up as the only child in the household develop different personality characteristics than those who grow up in larger families. A sample of $n = 30$ only children is obtained and each child is given a standardized personality test. For the general population, scores on the test from a normal distribution with a mean of $\mu = 50$ and a standard deviation of $\sigma = 15$. If the mean for the sample is $M = 58$, can the researcher conclude that there is a significant difference in personality between only children and the rest of the population? Use a two-tailed test with $\alpha = .05$.

15. A researcher is testing the hypothesis that consuming a sports drink during exercise improves endurance. A sample of $n = 50$ male college students is obtained and each student is given a series of three endurance tasks and asked to consume 4 ounces of the drink during each break between tasks. The overall endurance score for this sample is $M = 53$. For the general population of male college students, without any sports drink, the scores for this task average $\mu = 50$ with a standard deviation of $\sigma = 12$.

a. Can the researcher conclude that endurance scores with the sports drink are significantly higher than scores without the drink? Use a one-tailed test with $\alpha = .05$.

b. Can the researcher conclude that endurance scores with the sports drink are significantly different than scores without the drink? Use a two-tailed test with $\alpha = .05$.

c. You should find that the two tests lead to different conclusions. Explain why.

16. Montarello and Martins (2005) found that fifth-grade students completed more mathematics problems correctly when simple problems were mixed in with their regular math assignments. To further explore this phenomenon, suppose that a researcher selects a standardized mathematics achievement test that produces a normal distribution of scores with a mean of $\mu = 100$ and a standard deviation of $\sigma = 18$. The researcher modifies the test by inserting a set of very easy problems among the standardized questions, and gives the modified test to a sample of $n = 36$ students. If the average test score for the sample is $M = 104$, is this result sufficient to conclude that inserting the easy questions improves student performance? Use a one-tailed test with $\alpha = .01$.

17. Researchers have often noted increases in violent crimes when it is very hot. In fact, Reifman, Larrick, and Fein (1991) noted that this relationship even extends to baseball. That is, there is a much greater chance of a batter being hit by a pitch when the temperature increases. Consider the following hypothetical data. Suppose that over the past 30 years, during any given week of the major-league season, an average of $\mu = 12$ players are hit by wild pitches. Assume that the distribution is nearly normal with $\sigma = 3$. For a sample of $n = 4$ weeks in which the daily temperature was extremely hot, the weekly average of hit-by-pitch players was $M = 15.5$. Are players more likely to get hit by pitches during hot weeks? Set alpha to .05 for a one-tailed test.

18. A researcher plans to conduct an experiment testing the effect of caffeine on reaction time during a driving simulation task. A sample of $n = 9$ participants is selected and each person receives a standard dose of caffeine before being tested on the simulator. The caffeine is expected to lower reaction time by an average of 30 msec. Scores on the simulator task for the regular population (without caffeine) form a normal distribution with $\mu = 240$ msec. and $\sigma = 30$.

a. If the researcher uses a two-tailed test with $\alpha = .05$, what is the power of the hypothesis test?

b. Again assuming a two-tailed test with $\alpha = .05$, what is the power of the hypothesis test if the sample size is increased to $n = 25$?

19. A sample of $n = 40$ is selected from a normal population with $\mu = 75$ msec. and $\sigma = 12$, and a treatment is administered to the sample. The treatment is expected to increase scores by an average of 4 points.

a. If the treatment effect is evaluated with a two-tailed hypothesis test using $\alpha = .05$, what is the power of the test?

b. What is the power of the test if the researcher uses a one-tailed test with $\alpha = .05$?

20. Briefly explain how increasing sample size influences each of the following. Assume that all other factors are held constant.

a. The size of the z-score in a hypothesis test.

b. The size of Cohen's d.

c. The power of a hypothesis test.

21. Explain how the power of a hypothesis test is influenced by each of the following. Assume that all other factors are held constant.

a. Increasing the alpha level from .01 to .05.

b. Changing from a one-tailed test to a two-tailed test.

22. A researcher is investigating the effectiveness of a new medication for lowering blood pressure for individuals with systolic pressure greater than 140. For this population, systolic scores average $\mu = 160$ with a standard deviation of $\sigma = 20$, and the scores form a normal-shaped distribution. The researcher plans to select a sample of $n = 25$ individuals, and measure their systolic blood pressure after they take the medication for 60 days. If the researcher uses a two-tailed test with $\alpha = .05$,

a. What is the power of the test if the medication has a 5-point effect?

b. What is the power of the test if the medication has a 10-point effect?

23. A researcher is evaluating the influence of a treatment using a sample selected from a normally distributed population with a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$. The researcher expects a 12-point treatment effect and plans to use a two-tailed hypothesis test with $\alpha = .05$.

a. Compute the power of the test if the researcher uses a sample of $n = 16$ individuals. (See Example 8.6.)

b. Compute the power of the test if the researcher uses a sample of $n = 25$ individuals.

This page intentionally left blank

REVIEW

After completing this part, you should understand the basic procedures that form the foundation of inferential statistics. These include:

1. The ability to transform scores into *z*-scores to describe locations within a distribution and to standardize entire distributions.
2. The ability to determine probabilities associated with individual scores selected from a distribution, especially for scores from normal distributions.
3. The ability to transform sample means into *z*-scores and to determine the probabilities associated with sample means.
4. The ability to use a sample mean to evaluate a hypothesis about an unknown population mean.

The general goal of inferential statistics is to use the limited information from a sample to answer general questions about an unknown population. In Chapter 8, we introduced hypothesis testing, one of the most commonly used inferential procedures. The hypothesis test presented in Chapter 8 integrates *z*-scores from Chapter 5, probability from Chapter 6, and the distribution of sample means from Chapter 7 into a single procedure that allows researchers to use a sample from an unknown population to evaluate a hypothesis about the population mean. The researcher first obtains a sample from the unknown population and computes the sample mean. The sample mean and a hypothesized value for the population mean are then used to compute a *z*-score. If the resulting *z*-score is a high-probability value, near the center of the distribution of sample means, then the researcher concludes that the sample data fit the hypothesis and the decision is to fail to reject the hypothesis. On the other hand, if the resulting *z*-score is a low-probability value, out in the tails of the distribution of sample means, then the researcher concludes that the sample data do not fit the hypothesis and the decision is to reject the hypothesis.

REVIEW EXERCISES

1. Find each of the requested values for a population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 8$.

- **a.** What is the *z*-score corresponding to $X = 52$?
- **b.** What is the X value corresponding to $z = -0.50$?
- **c.** If all of the scores in the population are transformed into *z*-scores, what will be the values for the mean and standard deviation for the complete set of *z*-scores?
- **d.** What is the *z*-score corresponding to a sample mean of $M = 42$ for a sample of $n = 4$ scores?
- **e.** What is the *z*-score corresponding to a sample mean of $M = 42$ for a sample of $n = 16$ scores?

2. A survey of female high school seniors shows that the average amount of time spent on clothes, hair, and makeup each morning before school is $\mu = 35$ minutes. Assume that the distribution of preparation times is approximately normal with a standard deviation of $\sigma = 14$ minutes, and find each of the requested values.

- **a.** What proportion of female high school seniors spend more than 40 minutes preparing themselves for going to school each morning?
- **b.** What is the probability of randomly selecting a female high school senior who spends less than 10 minutes on her clothes, hair, and makeup each morning?
- **c.** What is the probability of obtaining a mean preparation time less than $M = 30$ minutes for a sample of $n = 49$ female high school students?

3. Brunt, Rhee, and Zhong (2008) surveyed 557 undergraduate college students to examine their weight status, health behaviors, and diet. Using body mass index (BMI), they classified the students into four categories: underweight, healthy weight, overweight, and obese. They also measured dietary variety by counting the number of different foods each student ate from several food groups. Note that the researchers are not measuring the amount of food eaten, but rather the number of different foods eaten (variety, not quantity). Nonetheless, it was somewhat surprising that the results showed no differences among the four weight categories that were related to eating fatty and/or sugary snacks.

Suppose a researcher conducting a follow up study obtains a sample of $n = 25$ students classified as healthy weight and a sample of $n = 36$ students classified as overweight. Each student completes the food variety questionnaire, and the healthy-weight group produces a mean of $M = 4.01$ for the fatty, sugary snack category compared to a mean of $M = 4.48$ for the overweight group. The results from the Brunt, Rhee, and Zhong study showed an overall mean variety score of $\mu = 4.22$ for the discretionary sweets or fats food group. Assume that the distribution of scores is approximately normal with a standard deviation of $\sigma = 0.60$.

- **a.** Does the sample of $n = 36$ indicate that number of fatty, sugary snacks eaten by overweight students is significantly different from the overall population mean? Use a two-tailed test with $\alpha = .05$.
- **b.** Based on the sample of $n = 25$ healthy-weight students, can you conclude that healthy-weight students eat significantly fewer fatty, sugary snacks than the overall population? Use a one-tailed test with $\alpha = .05$.

This page intentionally left blank

PART

III

Using *t* Statistics for Inferences About Population Means and Mean Differences

In Part II we presented the foundation for inferential statistics. In this part, we begin to introduce some of the inferential procedures that are actually used in behavioral science research. Specifically, we look at a family of *t* statistics that use sample means and mean differences to draw inferences about the corresponding population means and mean differences. The *t* statistics are all modeled after the *z*-score for sample means that was introduced in Chapter 7 and used for hypothesis testing in Chapter 8. However, the *t* statistics do not require any prior knowledge about the population being evaluated. The three *t* statistics introduced in this part apply to three distinct research situations:

1. Using a single sample to draw an inference about the unknown mean for a single population.
2. Using two separate samples to draw an inference about the mean difference between two unknown populations.
3. Using one sample, with each individual tested in two different treatment conditions, to draw an inference about the population mean difference between the two conditions.

In addition to the hypothesis testing procedure introduced in Chapter 8, this part introduces a new inferential technique known as *confidence intervals*. Confidence intervals allow researchers to use sample data to estimate population means or mean differences by computing a range of values that is highly likely to contain the unknown parameter.

This page intentionally left blank

CHAPTER

9

Introduction to the *t* Statistic

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Sample standard deviation (Chapter 4)
- Degrees of freedom (Chapter 4)
- Standard error (Chapter 7)
- Hypothesis testing (Chapter 8)

Preview

Numerous accounts suggest that for many animals, including humans, direct stare from another animal is aversive (e.g., Cook, 1977). Try it out for yourself. Make direct eye contact with a stranger in a cafeteria. Chances are the person will display avoidance by averting his or her gaze or turning away from you. Some insects, such as moths, have even developed eye-spot patterns on the wings or body to ward off predators (mostly birds) who may have a natural fear of eyes (Blest, 1957). Suppose that a comparative psychologist is interested in determining whether the birds that feed on these insects show an avoidance of eye-spot patterns.

Using methods similar to those of Scaife (1976), the researcher performed the following experiment. A sample of $n = 16$ moth-eating birds is selected. The animals are tested in an apparatus that consists of a two-chambered box. The birds are free to roam from one side of the box to the other through a doorway in the partition that separates the two chambers. In one chamber, there are two eye-spot patterns painted on the wall. The other side of the box has plain walls. One at a time, the researcher tests each bird by placing it in the doorway between the chambers. Each subject is left in the apparatus for 60 minutes, and the amount of time spent in the plain chamber is recorded.

The null hypothesis for this study would state that eye-spot patterns have no effect on the behavior of moth-eating birds. If this is true, then birds placed in the apparatus should wander randomly from side to side during the 60-minute period, averaging half of the time on each side. Therefore, for the general population of moth-eating birds, the null hypothesis states

$$H_0: \mu_{\text{plain side}} = 30 \text{ minutes}$$

The Problem: The researcher has most of the information needed to conduct a hypothesis test. In particular, the researcher has a hypothesis about the population ($\mu = 30$ minutes) and a sample of $n = 16$ scores that produces a sample mean (M). However, the researcher does not know the population standard deviation (σ). This value is needed to compute the standard error for the sample mean (σ_M) that appears in the denominator of the z-score equation. Recall that the standard error measures how much difference is reasonable to expect between a sample mean (M) and the population mean (μ). The value of the standard error is critical to deciding whether the sample data are consistent with the null hypothesis or refute the null hypothesis. Without the standard error, it is impossible to conduct a z-score hypothesis test.

The Solution: Because it is impossible to compute the standard error, a z-score cannot be used for the hypothesis test. However, it is possible to estimate the standard error using the sample data. The estimated standard error can then be used to compute a new statistic that is similar in structure to the z-score. The new statistic is called a *t statistic* and it can be used to conduct a new kind of hypothesis test.

9.1 THE *t* STATISTIC: AN ALTERNATIVE TO *z*

In the previous chapter, we presented the statistical procedures that permit researchers to use a sample mean to test hypotheses about an unknown population mean. These statistical procedures were based on a few basic concepts, which we summarize as follows:

Remember that the expected value of the distribution of sample means is μ, the population mean.

1. A sample mean (M) is expected to approximate its population mean (μ). This permits us to use the sample mean to test a hypothesis about the population mean.
2. The standard error provides a measure of how well a sample mean approximates the population mean. Specifically, the standard error determines how much difference is reasonable to expect between a sample mean (M) and the population mean (μ).

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \sigma_M = \sqrt{\frac{\sigma^2}{n}}$$

3. To quantify our inferences about the population, we compare the obtained sample mean (M) with the hypothesized population mean (μ) by computing a z-score test statistic.

$$z = \frac{M - \mu}{\sigma_M} = \frac{\text{obtained difference between data and hypothesis}}{\text{standard distance between } M \text{ and } \mu}$$

The goal of the hypothesis test is to determine whether the obtained difference between the data and the hypothesis is significantly greater than would be expected by chance. When the z-scores form a normal distribution, we are able to use the unit normal table (in Appendix B) to find the critical region for the hypothesis test.

THE PROBLEM WITH *z*-SCORES

The shortcoming of using a z-score for hypothesis testing is that the z-score formula requires more information than is usually available. Specifically, a z-score requires that we know the value of the population standard deviation (or variance), which is needed to compute the standard error. In most situations, however, the standard deviation for the population is not known. In fact, the whole reason for conducting a hypothesis test is to gain knowledge about an *unknown* population. This situation appears to create a paradox: You want to use a z-score to find out about an unknown population, but you must know about the population before you can compute a z-score. Fortunately, there is a relatively simple solution to this problem. When the variability for the population is not known, we use the sample variability in its place.

INTRODUCING THE *t* STATISTIC

In Chapter 4, the sample variance was developed specifically to provide an unbiased estimate of the corresponding population variance. Recall that the formulas for sample variance and sample standard deviation are as follows:

$$\text{sample variance} = s^2 = \frac{SS}{n - 1} = \frac{SS}{df}$$

$$\text{sample standard deviation} = s = \sqrt{\frac{SS}{n - 1}} = \sqrt{\frac{SS}{df}}$$

The concept of degrees of freedom, $df = n - 1$, was introduced in Chapter 4 (p. 117) and is discussed later in this chapter (p. 287).

Using the sample values, we can now *estimate* the standard error. Recall from Chapters 7 and 8 that the value of the standard error can be computed using either standard deviation or variance:

$$\text{standard error} = \sigma_M = \frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \sigma_M = \sqrt{\frac{\sigma^2}{n}}$$

Now we estimate the standard error by simply substituting the sample variance or standard deviation in place of the unknown population value:

$$\text{estimated standard error} = s_M = \frac{s}{\sqrt{n}} \qquad \text{or} \qquad s_M = \sqrt{\frac{s^2}{n}} \tag{9.1}$$

Notice that the symbol for the *estimated standard error of* M is s_M instead of σ_M, indicating that the estimated value is computed from sample data rather than from the actual population parameter.

DEFINITION

The **estimated standard error** (s_M) is used as an estimate of the real standard error, σ_M, when the value of σ is unknown. It is computed from the sample variance or sample standard deviation and provides an estimate of the standard distance between a sample mean, M, and the population mean, μ.

Finally, you should recognize that we have shown formulas for standard error (actual or estimated) using both the standard deviation and the variance. In the past (Chapters 7 and 8), we concentrated on the formula using the standard deviation. At this point, however, we shift our focus to the formula based on variance. Thus, throughout the remainder of this chapter, and in following chapters, the estimated standard error of M typically is presented and computed using

$$s_M = \sqrt{\frac{s^2}{n}}$$

There are two reasons for making this shift from standard deviation to variance:

1. In Chapter 4 (p. 119) we saw that the sample variance is an *unbiased* statistic; on average, the sample variance (s^2) provides an accurate and unbiased estimate of the population variance (σ^2). Therefore, the most accurate way to estimate the standard error is to use the sample variance to estimate the population variance.
2. In future chapters we encounter other versions of the t statistic that require variance (instead of standard deviation) in the formulas for estimated standard error. To maximize the similarity from one version to another, we use variance in the formula for *all* of the different t statistics. Thus, whenever we present a t statistic, the estimated standard error is computed as

$$\text{estimated standard error} = \sqrt{\frac{\text{sample variance}}{\text{sample size}}}$$

Now we can substitute the estimated standard error in the denominator of the z-score formula. The result is a new test statistic called a *t statistic*:

$$t = \frac{M - \mu}{s_M} \tag{9.2}$$

DEFINITION

The ***t* statistic** is used to test hypotheses about an unknown population mean, μ, when the value of σ is unknown. The formula for the t statistic has the same structure as the z-score formula, except that the t statistic uses the estimated standard error in the denominator.

The only difference between the t formula and the z-score formula is that the z-score uses the actual population variance, σ^2 (or the standard deviation), and the

t formula uses the corresponding sample variance (or standard deviation) when the population value is not known.

$$z = \frac{M - \mu}{\sigma_M} = \frac{M - \mu}{\sqrt{\sigma^2 / n}} \qquad t = \frac{M - \mu}{s_M} = \frac{M - \mu}{\sqrt{s^2 / n}}$$

DEGREES OF FREEDOM AND THE *t* STATISTIC

In this chapter, we have introduced the *t* statistic as a substitute for a *z*-score. The basic difference between these two is that the *t* statistic uses sample variance (s^2) and the *z*-score uses the population variance (σ^2). To determine how well a *t* statistic approximates a *z*-score, we must determine how well the sample variance approximates the population variance.

In Chapter 4, we introduced the concept of degrees of freedom (p. 117). Reviewing briefly, you must know the sample mean before you can compute sample variance. This places a restriction on sample variability such that only $n - 1$ scores in a sample are independent and free to vary. The value $n - 1$ is called the *degrees of freedom* (or *df*) for the sample variance.

$$\text{degrees of freedom} = df = n - 1 \tag{9.3}$$

DEFINITION

Degrees of freedom describe the number of scores in a sample that are independent and free to vary. Because the sample mean places a restriction on the value of one score in the sample, there are $n - 1$ degrees of freedom for a sample with n scores (see Chapter 4).

The greater the value of *df* for a sample, the better the sample variance, s^2, represents the population variance, σ^2, and the better the *t* statistic approximates the *z*-score. This should make sense because the larger the sample (n) is, the better the sample represents its population. Thus, the degrees of freedom associated with s^2 also describe how well *t* represents *z*.

THE *t* DISTRIBUTION

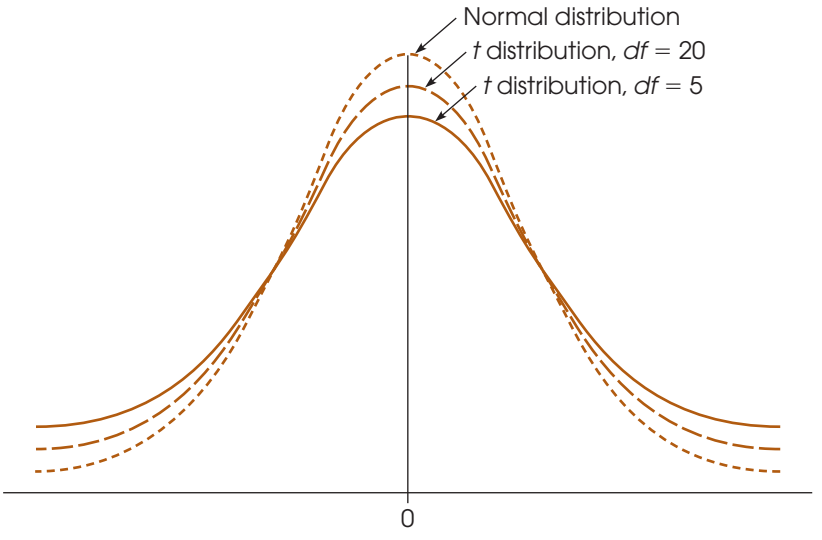
Every sample from a population can be used to compute a *z*-score or a *t* statistic. If you select all of the possible samples of a particular size (n), and compute the *z*-score for each sample mean, then the entire set of *z*-scores form a *z*-score distribution. In the same way, you can compute the *t* statistic for every sample and the entire set of *t* values form a *t distribution*. As we saw in Chapter 7, the distribution of *z*-scores for sample means tends to be a normal distribution. Specifically, if the sample size is large (around $n = 30$ or more) or if the sample is selected from a normal population, then the distribution of sample means is a nearly perfect normal distribution. In these same situations, the *t* distribution approximates a normal distribution, just as a *t* statistic approximates a *z*-score. How well a *t* distribution approximates a normal distributor is determined by degrees of freedom. In general, the greater the sample size (n) is, the larger the degrees of freedom ($n - 1$) are, and the better the *t* distribution approximates the normal distribution (Figure 9.1).

DEFINITION

A ***t* distribution** is the complete set of *t* values computed for every possible random sample for a specific sample size (n) or a specific degrees of freedom (*df*). The *t* distribution approximates the shape of a normal distribution.

FIGURE 9.1

Distributions of the *t* statistic for different values of degrees of freedom are compared to a normal *z*-score distribution. Like the normal distribution, *t* distributions are bell-shaped and symmetrical and have a mean of zero. However, *t* distributions have more variability, indicated by the flatter and more spread-out shape. The larger the value of *df* is, the more closely the *t* distribution approximates a normal distribution.



THE SHAPE OF THE *t* DISTRIBUTION

The exact shape of a *t* distribution changes with degrees of freedom. In fact, statisticians speak of a "family" of *t* distributions. That is, there is a different sampling distribution of *t* (a distribution of all possible sample *t* values) for each possible number of degrees of freedom. As *df* gets very large, the *t* distribution gets closer in shape to a normal *z*-score distribution. A quick glance at Figure 9.1 reveals that distributions of *t* are bell-shaped and symmetrical and have a mean of zero. However, the *t* distribution has more variability than a normal *z* distribution, especially when *df* values are small (see Figure 9.1). The *t* distribution tends to be flatter and more spread out, whereas the normal *z* distribution has more of a central peak.

The reason that the *t* distribution is flatter and more variable than the normal *z*-score distribution becomes clear if you look at the structure of the formulas for *z* and *t*. For both formulas, *z* and *t,* the top of the formula, $M - \mu$, can take on different values because the sample mean (M) varies from one sample to another. For *z*-scores, however, the bottom of the formula does not vary, provided that all of the samples are the same size and are selected from the same population. Specifically, all of the *z*-scores have the same standard error in the denominator, $\sigma_M = \sqrt{\sigma^2/n}$, because the population variance and the sample size are the same for every sample. For *t* statistics, on the other hand, the bottom of the formula varies from one sample to another. Specifically, the sample variance (s^2) changes from one sample to the next, so the estimated standard error also varies, $s_M = \sqrt{s^2/n}$. Thus, only the numerator of the *z*-score formula varies, but both the numerator and the denominator of the *t* statistic vary. As a result, *t* statistics are more variable than are *z*-scores, and the *t* distribution is flatter and more spread out. As sample size and *df* increase, however, the variability in the *t* distribution decreases, and it more closely resembles a normal distribution.

DETERMINING PROPORTIONS AND PROBABILITIES FOR *t* DISTRIBUTIONS

Just as we used the unit normal table to locate proportions associated with *z*-scores, we use a *t* distribution table to find proportions for *t* statistics. The complete *t* distribution table is presented in Appendix B, page 703, and a portion of this table is reproduced in

Table 9.1. The two rows at the top of the table show proportions of the *t* distribution contained in either one or two tails, depending on which row is used. The first column of the table lists degrees of freedom for the *t* statistic. Finally, the numbers in the body of the table are the *t* values that mark the boundary between the tails and the rest of the *t* distribution.

For example, with $df = 3$, exactly 5% of the *t* distribution is located in the tail beyond $t = 2.353$ (Figure 9.2). The process of finding this value is highlighted in Table 9.1. Begin by locating $df = 3$ in the first column of the table. Then locate a proportion of 0.05 (5%) in the one-tail proportion row. When you line up these two values in the table, you should find $t = 2.353$. Similarly, 5% of the *t* distribution is located in the tail beyond $t = -2.353$ (see Figure 9.2). Finally, notice that a total of 10% (or 0.10) is contained in the two tails beyond $t = \pm 2.353$ (check the proportion value in the "two-tails combined" row at the top of the table).

A close inspection of the *t* distribution table in Appendix B demonstrates a point we made earlier: As the value for *df* increases, the *t* distribution becomes more similar to a normal distribution. For example, examine the column containing *t* values for a 0.05 proportion in two tails. You will find that when $df = 1$, the *t* values that separate the extreme 5% (0.05) from the rest of the distribution are $t = \pm 12.706$. As

TABLE 9.1

A portion of the *t*-distribution table. The numbers in the table are the values of *t* that separate the tail from the main body of the distribution. Proportions for one or two tails are listed at the top of the table, and *df* values for *t* are listed in the first column.

	Proportion in One Tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in Two Tails Combined					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707

FIGURE 9.2

The *t* distribution with $df = 3$. Note that 5% of the distribution is located in the tail beyond $t = 2.353$. Also, 5% is in the tail beyond $t = -2.353$. Thus, a total proportion of 10% (0.10) is in the two tails beyond $t = \pm 2.353$.

you read down the column, however, you will find that the critical t values become smaller and smaller, ultimately reaching ± 1.96. You should recognize ± 1.96 as the z-score values that separate the extreme 5% in a normal distribution. Thus, as df increases, the proportions in a t distribution become more like the proportions in a normal distribution. When the sample size (and degrees of freedom) is sufficiently large, the difference between a t distribution and the normal distribution becomes negligible.

Caution: The t distribution table printed in this book has been abridged and does not include entries for every possible df value. For example, the table lists t values for $df = 40$ and for $df = 60$, but does not list any entries for df values between 40 and 60. Occasionally, you will encounter a situation in which your t statistic has a df value that is not listed in the table. In these situations, you should look up the critical t for both of the surrounding df values listed and then use the *larger* value for t. If, for example, you have $df = 53$ (not listed), look up the critical t value for both $df = 40$ and $df = 60$ and *then use the larger t value*. If your sample t statistic is greater than the larger value listed, then you can be certain that the data are in the critical region, and you can confidently reject the null hypothesis.

LEARNING CHECK

1. Under what circumstances is a t statistic used instead of a z-score for a hypothesis test?
2. A sample of $n = 9$ scores has $SS = 288$.
   - **a.** Compute the variance for the sample.
   - **b.** Compute the estimated standard error for the sample mean.
3. In general, a distribution of t statistics is flatter and more spread out than the standard normal distribution. (True or false?)
4. A researcher reports a t statistic with $df = 20$. How many individuals participated in the study?
5. For $df = 15$, find the value(s) of t associated with each of the following:
   - **a.** The top 5% of the distribution.
   - **b.** The middle 95% of the distribution.
   - **c.** The middle 99% of the distribution.

ANSWERS

1. A t statistic is used instead of a z-score when the population standard deviation and variance are not known.
2. **a.** $s^2 = 36$
   b. $s_M = 2$
3. True.
4. $n = 21$
5. **a.** $t = +1.753$
   b. $t = \pm 2.131$
   c. $t = \pm 2.947$

9.2 HYPOTHESIS TESTS WITH THE *t* STATISTIC

In the hypothesis-testing situation, we begin with a population with an unknown mean and an unknown variance, often a population that has received some treatment (Figure 9.3). The goal is to use a sample from the treated population (a treated sample) as the basis for determining whether the treatment has any effect.

As always, the null hypothesis states that the treatment has no effect; specifically, H_0 states that the population mean is unchanged. Thus, the null hypothesis provides a specific value for the unknown population mean. The sample data provide a value for the sample mean. Finally, the variance and estimated standard error are computed from the sample data. When these values are used in the *t* formula, the result becomes

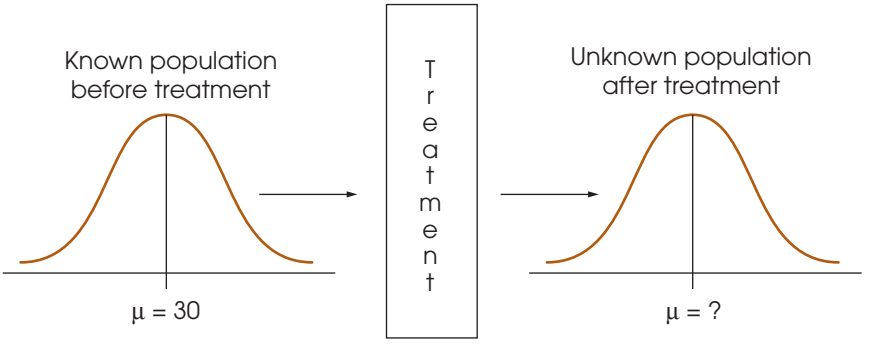
$$t = \frac{\text{sample mean (from the data)} - \text{population mean (hypothesized from } H_0)}{\text{estimated standard error (computed from the sample data)}}$$

As with the *z*-score formula, the *t* statistic forms a ratio. The numerator measures the actual difference between the sample data (M) and the population hypothesis (μ). The estimated standard error in the denominator measures how much difference is reasonable to expect between a sample mean and the population mean. When the obtained difference between the data and the hypothesis (numerator) is much greater than expected (denominator), we obtain a large value for *t* (either large positive or large negative). In this case, we conclude that the data are not consistent with the hypothesis, and our decision is to "reject H_0." On the other hand, when the difference between the data and the hypothesis is small relative to the standard error, we obtain a *t* statistic near zero, and our decision is "fail to reject H_0."

The Unknown Population As mentioned earlier, the hypothesis test often concerns a population that has received a treatment. This situation is shown in Figure 9.3. Note

FIGURE 9.3

The basic experimental situation for using the *t* statistic or the *z*-score is presented. It is assumed that the parameter μ is known for the population before treatment. The purpose of the experiment is to determine whether the treatment has an effect. Note that the population after treatment has unknown values for the mean and the variance. We will use a sample to test a hypothesis about the population mean.



that the value of the mean is known for the population before treatment. The question is whether the treatment influences the scores and causes the mean to change. In this case, the unknown population is the one that exists after the treatment is administered, and the null hypothesis simply states that the value of the mean is not changed by the treatment.

Although the *t* statistic can be used in the "before and after" type of research shown in Figure 9.3, it also permits hypothesis testing in situations for which you do not have a known population mean to serve as a standard. Specifically, the *t* test does not require any prior knowledge about the population mean or the population variance. All you need to compute a *t* statistic is a null hypothesis and a sample from the unknown population. Thus, a *t* test can be used in situations for which the null hypothesis is obtained from a theory, a logical prediction, or just wishful thinking. For example, many surveys contain rating-scale questions to determine how people feel about controversial issues. Participants are presented with a statement and asked to express their opinion on a scale from 1 to 7, with 1 indicating "strongly agree" and 7 indicating "strongly disagree." A score of 4 indicates a neutral position, with no strong opinion one way or the other. In this situation, the null hypothesis would state that there is no preference in the population, H_0: $\mu = 4$. The data from a sample is then used to evaluate the hypothesis. Note that the researcher has no prior knowledge about the population mean and states a hypothesis that is based on logic.

HYPOTHESIS TESTING EXAMPLE

The following research situation demonstrates the procedures of hypothesis testing with the *t* statistic. Note that this is another example of a null hypothesis that is founded in logic rather than prior knowledge of a population mean.

EXAMPLE 9.1

Infants, even newborns, prefer to look at attractive faces compared to less attractive faces (Slater, et al., 1998). In the study, infants from 1 to 6 days old were shown two photographs of women's faces. Previously, a group of adults had rated one of the faces as significantly more attractive than the other. The babies were positioned in front of a screen on which the photographs were presented. The pair of faces remained on the screen until the baby accumulated a total of 20 seconds of looking at one or the other. The number of seconds looking at the attractive face was recorded for each infant. Suppose that the study used a sample of $n = 9$ infants and the data produced an average of $M = 13$ seconds for the attractive face with $SS = 72$. Note that all of the available information comes from the sample. Specifically, we do not know the population mean or the population standard deviation.

STEP 1

State the hypotheses and select an alpha level. Although we have no information about the population of scores, it is possible to form a logical hypothesis about the value of μ. In this case, the null hypothesis states that the infants have no preference for either face. That is, they should average half of the 20 seconds looking at each of the two faces. In symbols, the null hypothesis states

$$H_0: \mu_{\text{attractive}} = 10 \text{ seconds}$$

The alternative hypothesis states that there is a preference and one of the faces is preferred over the other. A directional, one-tailed test would specify which of the two faces is preferred, but the nondirectional alternative hypothesis is expressed as follows:

$$H_1: \mu_{\text{attractive}} \neq 10 \text{ seconds}$$

We set the level of significance at $\alpha = .05$ for two tails.

STEP 2 Locate the critical region. The test statistic is a *t* statistic because the population variance is not known. Therefore, the value for degrees of freedom must be determined before the critical region can be located. For this sample

$$df = n - 1 = 9 - 1 = 8$$

For a two-tailed test at the .05 level of significance and with 8 degrees of freedom, the critical region consists of *t* values greater than +2.306 or less than –2.306. Figure 9.4 depicts the critical region in this *t* distribution.

STEP 3 Calculate the test statistic. The *t* statistic typically requires much more computation than is necessary for a *z*-score. Therefore, we recommend that you divide the calculations into a three-stage process as follows:

a. First, calculate the sample variance. Remember that the population variance is unknown, and you must use the sample value in its place. (This is why we are using a *t* statistic instead of a *z*-score.)

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df} = \frac{72}{8} = 9$$

b. Next, use the sample variance (s^2) and the sample size (n) to compute the estimated standard error. This value is the denominator of the *t* statistic and measures how much difference is reasonable to expect by chance between a sample mean and the corresponding population mean.

$$s_M = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{9}{9}} = \sqrt{1} = 1$$

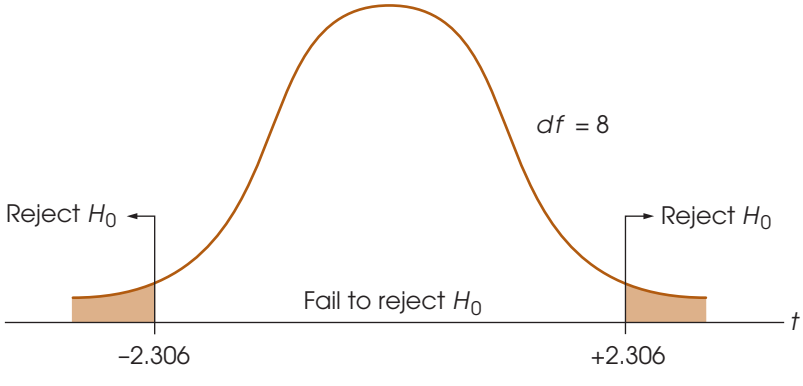
Finally, compute the *t* statistic for the sample data.

$$t = \frac{M - \mu}{s_M} = \frac{13 - 10}{1} = 3.00$$

STEP 4 Make a decision regarding H_0. The obtained *t* statistic of 3.00 falls into the critical region on the right-hand side of the *t* distribution (see Figure 9.4). Our statistical decision is to reject H_0 and conclude that babies do show a preference when given a

FIGURE 9.4

The critical region in the *t* distribution for $\alpha = .05$ and $df = 8$.



choice between an attractive and an unattractive face. Specifically, the average amount of time that the babies spent looking at the attractive face was significantly different from the 10 seconds that would be expected if there were no preference. As indicated by the sample mean, there is a tendency for the babies to spend more time looking at the attractive face.

ASSUMPTIONS OF THE *t* TEST

Two basic assumptions are necessary for hypothesis tests with the *t* statistic.

1. The values in the sample must consist of *independent* observations.

   In everyday terms, two observations are independent if there is no consistent, predictable relationship between the first observation and the second. More precisely, two events (or observations) are independent if the occurrence of the first event has no effect on the probability of the second event. We examined specific examples of independence and non-independence in Box 8.1 (p. 254).

2. The population that is sampled must be normal.

   This assumption is a necessary part of the mathematics underlying the development of the *t* statistic and the *t* distribution table. However, violating this assumption has little practical effect on the results obtained for a *t* statistic, especially when the sample size is relatively large. With very small samples, a normal population distribution is important. With larger samples, this assumption can be violated without affecting the validity of the hypothesis test. If you have reason to suspect that the population distribution is not normal, use a large sample to be safe.

THE INFLUENCE OF SAMPLE SIZE AND SAMPLE VARIANCE

As we noted in Chapter 8 (p. 252), a variety of factors can influence the outcome of a hypothesis test. In particular, the number of scores in the sample and the magnitude of the sample variance both have a large effect on the *t* statistic and thereby influence the statistical decision. The structure of the *t* formula makes these factors easier to understand.

$$t = \frac{M - \mu}{s_M} \qquad \text{where } s_M = \sqrt{\frac{s^2}{n}}$$

Because the estimated standard error, s_M, appears in the denominator of the formula, a larger value for s_M produces a smaller value (closer to zero) for *t*. Thus, any factor that influences the standard error also affects the likelihood of rejecting the null hypothesis and finding a significant treatment effect. The two factors that determine the size of the standard error are the sample variance, s^2, and the sample size, *n*.

The estimated standard error is directly related to the sample variance so that the larger the variance, the larger the error. Thus, large variance means that you are less likely to obtain a significant treatment effect. In general, large variance is bad for inferential statistics. Large variance means that the scores are widely scattered, which makes it difficult to see any consistent patterns or trends in the data. In general, high variance reduces the likelihood of rejecting the null hypothesis.

On the other hand, the estimated standard error is inversely related to the number of scores in the sample. The larger the sample is, the smaller the error is. If all other factors are held constant, large samples tend to produce bigger *t* statistics and therefore are more likely to produce significant results. For example, a 2-point mean difference with

a sample of $n = 4$ may not be convincing evidence of a treatment effect. However, the same 2-point difference with a sample of $n = 100$ is much more compelling.

LEARNING CHECK

1. A sample of $n = 4$ individuals is selected from a population with a mean of $\mu = 40$. A treatment is administered to the individuals in the sample and, after treatment, the sample has a mean of $M = 44$ and a variance of $s^2 = 16$.

a. Is this sample sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.

b. If all other factors are held constant and the sample size is increased to $n = 16$, is the sample sufficient to conclude that the treatment has a significant effect? Again, use a two-tailed test with $\alpha = .05$.

ANSWER

1. a. H_0: $\mu = 40$ even after the treatment. With $n = 4$, the estimated standard error is 2, and $t = 4/2 = 2.00$. With $df = 3$, the critical boundaries are set at $t = \pm 3.182$. Fail to reject H_0 and conclude that the treatment does not have a significant effect.

b. With $n = 16$, the estimated standard error is 1 and $t = 4.00$. With $df = 15$, the critical boundary is ± 2.131. The t value is beyond the critical boundary, so we reject H_0 and conclude that the treatment does have a significant effect.

9.3 MEASURING EFFECT SIZE FOR THE *t* STATISTIC

In Chapter 8, we noted that one criticism of a hypothesis test is that it does not really evaluate the size of the treatment effect. Instead, a hypothesis test simply determines whether the treatment effect is greater than chance, where "chance" is measured by the standard error. In particular, it is possible for a very small treatment effect to be "statistically significant," especially when the sample size is very large. To correct for this problem, it is recommended that the results from a hypothesis test be accompanied by a report of effect size, such as Cohen's *d*.

ESTIMATED COHEN'S *d*

When Cohen's *d* was originally introduced (p. 262), the formula was presented as

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{\mu_{\text{treatment}} - \mu_{\text{no treatment}}}{\sigma}$$

Cohen defined this measure of effect size in terms of the population mean difference and the population standard deviation. However, in most situations the population values are not known and you must substitute the corresponding sample values in their place. When this is done, many researchers prefer to identify the calculated value as an "*estimated* d" or name the value after one of the statisticians who first substituted sample statistics into Cohen's formula (e.g., Glass's g or Hedges's g). For hypothesis tests using the t statistic, the population mean with no treatment is the value specified by the null hypothesis. However, the population mean with treatment and the standard deviation are both unknown. Therefore, we use the mean for the treated sample and the standard deviation for the sample after treatment as

estimates of the unknown parameters. With these substitutions, the formula for estimating Cohen's *d* becomes

$$\text{estimated } d = \frac{\text{mean difference}}{\text{sample standard deviation}} = \frac{M - \mu}{s} \tag{9.4}$$

The numerator measures that magnitude of the treatment effect by finding the difference between the mean for the treated sample and the mean for the untreated population (μ from H_0). The sample standard deviation in the denominator standardizes the mean difference into standard deviation units. Thus, an estimated *d* of 1.00 indicates that the size of the treatment effect is equivalent to one standard deviation. The following example demonstrates how the estimated *d* is used to measure effect size for a hypothesis test using a *t* statistic.

EXAMPLE 9.2

For the infant face-preference study in Example 9.1, the babies averaged $M = 13$ out of 20 seconds looking at the attractive face. If there were no preference (as stated by the null hypothesis), the population mean would be $\mu = 10$ seconds. Thus, the results show a 3-second difference between the mean with a preference ($M = 13$) and the mean with no preference ($\mu = 10$). Also, for this study the sample standard deviation is

$$s = \sqrt{\frac{SS}{df}} = \sqrt{\frac{72}{8}} = \sqrt{9} = 3$$

Thus, Cohen's *d* for this example is estimated to be

$$\text{Cohen's } d = \frac{M - \mu}{s} = \frac{13 - 10}{3} = 1.00$$

According to the standards suggested by Cohen (Table 8.2, p. 264), this is a large treatment effect.

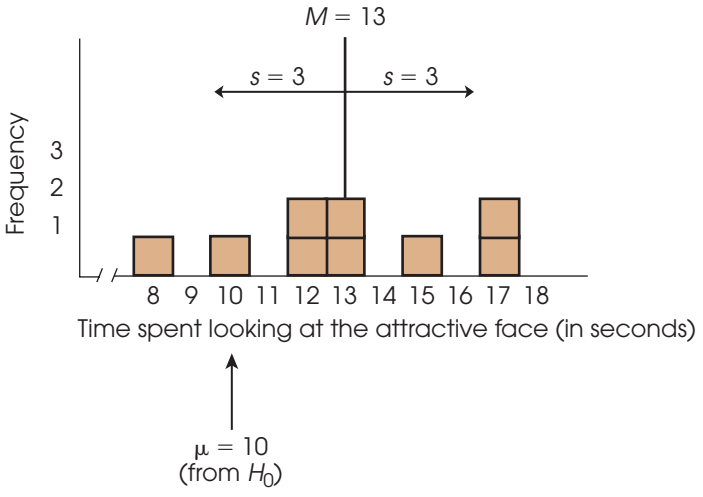
To help you visualize what is measured by Cohen's *d,* we have constructed a set of $n = 9$ scores with a mean of $M = 13$ and a standard deviation of $s = 3$ (the same values as in Examples 9.1 and 9.2). The set of scores is shown in Figure 9.5. Notice that the figure also includes an arrow that locates $\mu = 10$. Recall that $\mu = 10$ is the value specified by the null hypothesis and identifies what the mean ought to be if the treatment has no effect. Clearly, our sample is not centered around $\mu = 10$. Instead, the scores have been shifted to the right so that the sample mean is $M = 13$. This shift, from 10 to 13, is the 3-point mean difference that was caused by the treatment effect. Also notice that the 3-point mean difference is exactly equal to the standard deviation. Thus, the size of the treatment effect is equal to 1 standard deviation. In other words, Cohen's $d = 1.00$.

MEASURING THE PERCENTAGE OF VARIANCE EXPLAINED, r^2

An alternative method for measuring effect size is to determine how much of the variability in the scores is explained by the treatment effect. The concept behind this measure is that the treatment causes the scores to increase (or decrease), which means that the treatment is causing the scores to vary. If we can measure how much of the variability is explained by the treatment, we can obtain a measure of the size of the treatment effect.

FIGURE 9.5

The sample distribution for the scores that were used in Examples 9.1 and 9.2. The population mean, $\mu = 10$ seconds, is the value that would be expected if attractiveness has no effect on the infants' behavior. Note that the sample mean is displaced away from $\mu = 10$ by a distance equal to one standard deviation.



To demonstrate this concept we use the data from the hypothesis test in Example 9.1. Recall that the null hypothesis stated that the treatment (the attractiveness of the faces) has no effect on the infants' behavior. According to the null hypothesis, the infants should show no preference between the two photographs, and therefore should spend an average of $\mu = 10$ out of 20 seconds looking at the attractive face.

However, if you look at the data in Figure 9.5, the scores are not centered around $\mu = 10$. Instead, the scores are shifted to the right so that they are centered around the sample mean, $M = 13$. This shift is the treatment effect. To measure the size of the treatment effect, we calculate deviations from the mean and the sum of squared deviations, *SS*, in two different ways.

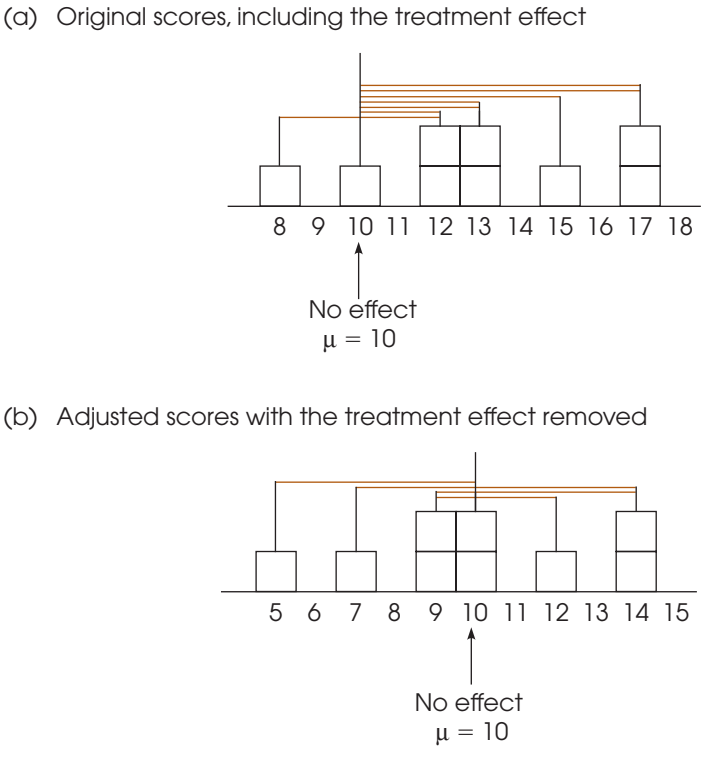
Figure 9.6(a) shows the original set of scores. For each score, the deviation from $\mu = 10$ is shown as a colored line. Recall that $\mu = 10$ comes from the null hypothesis and represents the population mean if the treatment has no effect. Note that almost all of the scores are located on the right-hand side of $\mu = 10$. This shift to the right is the treatment effect. Specifically, the preference for the attractive face has caused the infants to spend more time looking at the attractive photograph, which means that their scores are generally greater than 10. Thus, the treatment has pushed the scores away from $\mu = 10$ and has increased the size of the deviations.

Next, we see what happens if the treatment effect is removed. In this example, the treatment has a 3-point effect (the average increases from $\mu = 10$ to $M = 13$). To remove the treatment effect, we simply subtract 3 points from each score. The adjusted scores are shown in Figure 9.6(b) and, once again, the deviations from $\mu = 10$ are shown as colored lines. First, notice that the adjusted scores are centered at $\mu = 10$, indicating that there is no treatment effect. Also notice that the deviations, the colored lines, are noticeably smaller when the treatment effect is removed.

To measure how much the variability is reduced when the treatment effect is removed, we compute the sum of squared deviations, *SS*, for each set of scores. The left-hand columns of Table 9.2 show the calculations for the original scores [Figure 9.6(a)], and the right-hand columns show the calculations for the adjusted scores [Figure 9.6(b)]. Note that the total variability, including the treatment effect, is $SS = 153$. However, when the treatment effect is removed, the variability is reduced to $SS = 72$. The difference

FIGURE 9.6

Deviations from $\mu = 10$ (no treatment effect) for the scores in Example 9.1. The colored lines in part (a) show the deviations for the original scores, including the treatment effect. In part (b) the colored lines show the deviations for the adjusted scores after the treatment effect has been removed.



between these two values, 153 – 72 = 81 points, is the amount of variability that is accounted for by the treatment effect. This value is usually reported as a proportion or percentage of the total variability:

$$\frac{\text{variability accounted for}}{\text{total variability}} = \frac{81}{153} = 0.5294 \quad (52.94\%)$$

TABLE 9.2

Calculation of *SS*, the sum of squared deviations, for the data in Figure 9.6. The first three columns show the calculations for the original scores, including the treatment effect. The last three columns show the calculations for the adjusted scores after the treatment effect has been removed.

Calculation of *SS* including the treatment effect			Calculation of *SS* after the treatment effect is removed		
Score	Deviation from $\mu = 10$	Squared Deviation	Adjusted Score	Deviation from $\mu = 10$	Squared Deviation
8	−2	4	8 − 3 = 5	−5	25
10	0	0	10 − 3 = 7	−3	9
12	2	4	12 − 3 = 9	−1	1
12	2	4	12 − 3 = 9	−1	1
13	3	9	13 − 3 = 10	0	0
13	3	9	13 − 3 = 10	0	0
15	5	25	15 − 3 = 12	2	4
17	7	49	17 − 3 = 14	4	16
17	7	49	17 − 3 = 14	4	16
		$SS = 153$			$SS = 72$

Thus, removing the treatment effect reduces the variability by 52.94%. This value is called the *percentage of variance accounted for by the treatment* and is identified as r^2.

Rather than computing r^2 directly by comparing two different calculations for *SS*, the value can be found from a single equation based on the outcome of the *t* test.

$$r^2 = \frac{t^2}{t^2 + df} \tag{9.5}$$

The letter *r* is the traditional symbol used for a correlation, and the concept of r^2 is discussed again when we consider correlations in Chapter 15. Also, in the context of *t* statistics, the percentage of variance that we are calling r^2 is often identified by the Greek letter omega squared (ω^2).

For the hypothesis test in Example 9.1, we obtained $t = 3.00$ with $df = 8$. These values produce

$$r^2 = \frac{3^2}{3^2 + 8} = \frac{9}{17} = 0.5294 \quad (52.94\%)$$

Note that this is exactly the same value we obtained with the direct calculation of the percentage of variability accounted for by the treatment.

Interpreting r^2 In addition to developing the Cohen's *d* measure of effect size, Cohen (1988) also proposed criteria for evaluating the size of a treatment effect that is measured by r^2. The criteria were actually suggested for evaluating the size of a correlation, *r*, but are easily extended to apply to r^2. Cohen's standards for interpreting r^2 are shown in Table 9.3.

According to these standards, the data we constructed for Examples 9.1 and 9.2 show a very large effect size with $r^2 = .5294$.

As a final note, we should remind you that, although sample size affects the hypothesis test, this factor has little or no effect on measures of effect size. In particular, estimates of Cohen's *d* are not influenced at all by sample size, and measures of r^2 are only slightly affected by changes in the size of the sample. The sample variance, on the other hand, influences hypothesis tests and measures of effect size. Specifically, high variance reduces both the likelihood of rejecting the null hypothesis and measures of effect size.

CONFIDENCE INTERVALS FOR ESTIMATING μ

An alternative technique for describing the size of a treatment effect is to compute an estimate of the population mean after treatment. For example, if the mean before treatment is known to be $\mu = 80$ and the mean after treatment is estimated to be $\mu = 86$, then we can conclude that the size of the treatment effect is around 6 points.

Estimating an unknown population mean involves constructing a *confidence interval*. A confidence interval is based on the observation that a sample mean tends to provide a reasonably accurate estimate of the population mean. The fact that a sample

TABLE 9.3

Criteria for interpreting the value of r^2 as proposed by Cohen (1988).

Percentage of Variance Explained, r^2	
$r^2 = 0.01$	Small effect
$r^2 = 0.09$	Medium effect
$r^2 = 0.25$	Large effect

mean tends to be near to the population mean implies that the population mean should be near to the sample mean. For example, if we obtain a sample mean of $M = 86$, we can be reasonably confident that the population mean is around 86. Thus, a confidence interval consists of an interval of values around a sample mean, and we can be reasonably confident that the unknown population mean is located somewhere in the interval.

DEFINITION

A **confidence interval** is an interval, or range of values, centered around a sample statistic. The logic behind a confidence interval is that a sample statistic, such as a sample mean, should be relatively near to the corresponding population parameter. Therefore, we can confidently estimate that the value of the parameter should be located in the interval.

CONSTRUCTING A CONFIDENCE INTERVAL

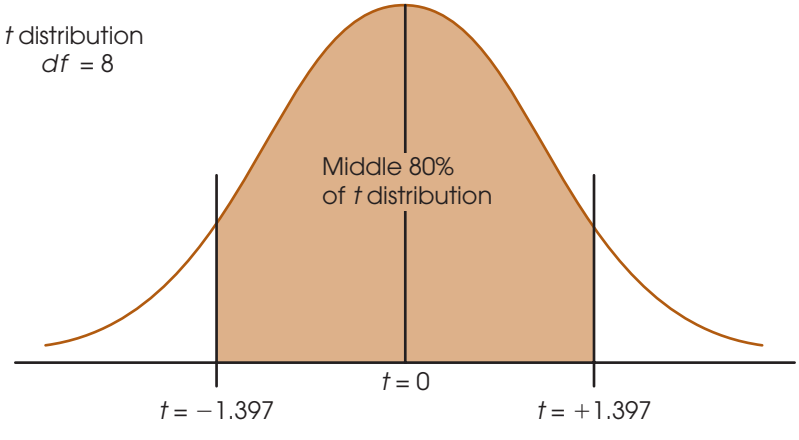
The construction of a confidence interval begins with the observation that every sample mean has a corresponding *t* value defined by the equation

$$t = \frac{M - \mu}{s_M}$$

Although the values for M and s_M are available from the sample data, we cannot use the equation to calculate *t* because the value for μ is unknown. Instead of calculating the *t* value, for a confidence interval we *estimate* the *t* value. For example, if the sample has $n = 9$ scores, then the *t* statistic has $df = 8$, and the distribution of all possible *t* values can be pictured as in Figure 9.7. Notice that the *t* values pile up around $t = 0$, so we can estimate that the *t* value for our sample should have a value around 0. Furthermore, the *t* distribution table lists a variety of different *t* values that correspond to specific proportions of the *t* distribution. With $df = 8$, for example, 80% of the *t* values are located between $t = +1.397$ and $t = -1.397$. To obtain these values, simply look up a two-tailed proportion of 0.20 (20%) for $df = 8$. Because 80% of all of the possible *t* values are located between ± 1.397, we can be 80% confident that our sample mean corresponds to a *t* value in this interval. Similarly, we can be 95% confident that the mean for a sample of $n = 9$ scores corresponds to a *t* value between +2.306 and –2.306. Notice that we are able to estimate the value of *t* with a specific level of

FIGURE 9.7

The distribution of *t* statistics with $df = 8$. The *t* values pile up around $t = 0$ and 80% of all of the possible values are located between $t = -1.397$ and $t = +1.397$.



confidence. To construct a confidence interval for μ, we plug the estimated *t* value into the *t* equation, and then we can calculate the value of μ.

Before we demonstrate the process of constructing a confidence interval for an unknown population mean, we simplify the calculations by regrouping the terms in the *t* equation. Because the goal is to compute the value of μ, we use simple algebra to solve the equation for μ. The result is

$$\mu = M \pm ts_M \tag{9.6}$$

The process of using this equation to construct a confidence interval is demonstrated in the following example.

EXAMPLE 9.3

Example 9.1 describes a study in which infants displayed a preference for the more attractive face by looking at it, instead of the less attractive face, for the majority of a 20-second viewing period. Specifically, a sample of $n = 9$ infants spent an average of $M = 13$ seconds out of a 20-second period looking at the more attractive face. The data produced an estimated standard error of $s_M = 1$. We use this sample to construct a confidence interval to estimate the mean amount of time that the population of infants spends looking at the more attractive face. That is, we construct an interval of values that is likely to contain the unknown population mean.

Again, the estimation formula is

$$\mu = M \pm t(s_M)$$

To have 80% in the middle, there must be 20% (or .20) in the tails. To find the *t* values, look under two tails, .20 in the *t* table.

In the equation, the value of $M = 13$ and $s_M = 1$ are obtained from the sample data. The next step is to select a level of confidence that determines the value of *t* in the equation. The most commonly used confidence level is probably 95%, but values of 80%, 90%, and 99% are also common. For this example, we use a confidence level of 80%, which means that we construct the confidence interval so that we are 80% confident that the population mean is actually contained in the interval. Because we are using a confidence level of 80%, the resulting interval is called the *80% confidence interval for μ*.

To obtain the value for *t* in the equation, we simply estimate that the *t* statistic for our sample is located somewhere in the middle 80% of the *t* distribution. With $df = n - 1 = 8$, the middle 80% of the distribution is bounded by *t* values of +1.397 and –1.397 (see Figure 9.7). Using the sample data and the estimated range of *t* values, we obtain

$$\mu = M \pm t(s_M) = 13 \pm 1.397(1.00) = 13 \pm 1.397$$

At one end of the interval, we obtain $\mu = 13 + 1.397 = 14.397$, and at the other end we obtain $\mu = 13 - 1.397 = 11.603$. Our conclusion is that the average time looking at the more attractive fact for the population of infants is between $\mu = 11.603$ seconds and $\mu = 14.397$ seconds, and we are 80% confident that the true population mean is located within this interval. The confidence comes from the fact that the calculation was based on only one assumption. Specifically, we assumed that the *t* statistic was located between +1.397 and –1.397, and we are 80% confident that this assumption is correct because 80% of all of the possible *t* values are located in this interval. Finally, note that the confidence interval is constructed around the sample mean. As a result, the sample mean, $M = 13$, is located exactly in the center of the interval.

FACTORS AFFECTING THE WIDTH OF A CONFIDENCE INTERVAL

Two characteristics of the confidence interval should be noted. First, notice what happens to the width of the interval when you change the level of confidence (the percent confidence). To gain more confidence in your estimate, you must increase the width of the interval. Conversely, to have a smaller, more precise interval, you must give up confidence. In the estimation formula, the percentage of confidence influences the value of t. A larger level of confidence (the percentage), produces a larger t value and a wider interval. This relationship can be seen in Figure 9.7. In the figure, we identified the middle 80% of the t distribution to find an 80% confidence interval. It should be obvious that if we were to increase the confidence level to 95%, it would be necessary to increase the range of t values, and thereby increase the width of the interval.

Second, note what happens to the width of the interval if you change the sample size. This time the basic rule is as follows: The bigger the sample (n), the smaller the interval. This relationship is straightforward if you consider the sample size as a measure of the amount of information. A bigger sample gives you more information about the population and allows you to make a more precise estimate (a narrower interval). The sample size controls the magnitude of the standard error in the estimation formula. As the sample size increases, the standard error decreases, and the interval gets smaller. Because confidence intervals are influenced by sample size, they do not provide an unqualified measure of absolute effect size and are not an adequate substitute for Cohen's d or r^2. Nonetheless, they can be used in a research report to provide a description of the size of the treatment effect.

LEARNING CHECK

1. If all other factors are held constant, an 80% confidence interval is wider than a 90% confidence interval. (True or false?)

2. If all other factors are held constant, a confidence interval computed from a sample of $n = 25$ is wider than a confidence interval computed from a sample of $n = 100$. (True or false?)

ANSWERS

1. False. Greater confidence requires a wider interval.

2. True. The smaller sample produces a wider interval.

IN THE LITERATURE
REPORTING THE RESULTS OF A *t* TEST

In Chapter 8, we noted the conventional style for reporting the results of a hypothesis test, according to APA format. First, recall that a scientific report typically uses the term *significant* to indicate that the null hypothesis has been rejected and the term *not significant* to indicate failure to reject H_0. Additionally, there is a prescribed format for reporting the calculated value of the test statistic, degrees of freedom, and alpha level for a t test. This format parallels the style introduced in Chapter 8 (p. 251).

In Example 9.1 we calculated a t statistic of 3.00 with $df = 8$, and we decided to reject H_0 with alpha set at .05. Using the same data from Example 9.1, we obtained $r^2 = 0.5294$ (52.94%) for the percentage of variance explained by the treatment effect. In a scientific report, this information is conveyed in a concise statement, as follows:

> The infants spent an average of $M = 13$ out of 20 seconds looking at the attractive face, with $SD = 3.00$. Statistical analysis indicates that the time spent looking at the attractive face was significantly greater than would be expected if there were no preference, $t(8) = 3.00$, $p < .05$, $r^2 = 0.5294$.

The first statement reports the descriptive statistics, the mean ($M = 13$) and the standard deviation ($SD = 3$), as previously described (Chapter 4, p. 123). The next statement provides the results of the inferential statistical analysis. Note that the degrees of freedom are reported in parentheses immediately after the symbol *t*. The value for the obtained *t* statistic follows (3.00), and next is the probability of committing a Type I error (less than 5%). Finally, the effect size is reported, $r^2 = 52.94\%$. If the 80% confidence interval from Example 9.3 were included in the report as a description of effect size, it would be added after the results of the hypothesis test as follows:

$$t(8) = 3.00,\ p < .05,\ 80\%\ \text{CI}\ [11.603, 14.397].$$

Often, researchers use a computer to perform a hypothesis test like the one in Example 9.1. In addition to calculating the mean, standard deviation, and the *t* statistic for the data, the computer usually calculates and reports the *exact probability* (or α level) associated with the *t* value. In Example 9.1 we determined that any *t* value beyond ± 2.306 has a probability of less than .05 (see Figure 9.4). Thus, the obtained *t* value, $t = 3.00$, is reported as being very unlikely, $p < .05$. A computer printout, however, would have included an exact probability for our specific *t* value.

The statement $p < .05$ was explained in Chapter 8, page 251.

Whenever a specific probability value is available, you are encouraged to use it in a research report. For example, the computer analysis of these data reports an exact *p* value of $p = .017$, and the research report would state "$t(8) = 3.00$, $p = .017$" instead of using the less specific "$p < .05$." As one final caution, we note that occasionally a *t* value is so extreme that the computer reports $p = 0.000$. The zero value does not mean that the probability is literally zero; instead, it means that the computer has rounded off the probability value to three decimal places and obtained a result of 0.000. In this situation, you do not know the exact probability value, but you can report $p < .001$.

LEARNING CHECK

1. A sample of $n = 16$ individuals is selected from a population with a mean of $\mu = 80$. A treatment is administered to the sample and, after treatment, the sample mean is found to be $M = 86$ with a standard deviation of $s = 8$.

- **a.** Does the sample provide sufficient evidence to conclude that the treatment has a significant effect? Test with $\alpha = .05$.
- **b.** Compute Cohen's *d* and r^2 to measure the effect size.
- **c.** Find the 95% confidence interval for the population mean after treatment.

2. How does sample size influence the outcome of a hypothesis test and measures of effect size? How does the standard deviation influence the outcome of a hypothesis test and measures of effect size?

ANSWERS

1. a. The estimated standard error is 2 points and the data produce $t = 6/2 = 3.00$. With $df = 15$, the critical values are $t = \pm 2.131$, so the decision is to reject H_0 and conclude that there is a significant treatment effect.

b. For these data, $d = 6/8 = 0.75$ and $r^2 = 9/24 = 0.375$ or 37.5%.

c. For 95% confidence and $df = 15$, use $t = \pm 2.131$. The confidence interval is $\mu = 86 \pm 2.131(2)$ and extends from 81.738 to 90.262.

2. Increasing sample size increases the likelihood of rejecting the null hypothesis but has little or no effect on measures of effect size. Increasing the sample variance reduces the likelihood of rejecting the null hypothesis and reduces measures of effect size.

9.4 DIRECTIONAL HYPOTHESES AND ONE-TAILED TESTS

As noted in Chapter 8, the nondirectional (two-tailed) test is more commonly used than the directional (one-tailed) alternative. On the other hand, a directional test may be used in some research situations, such as exploratory investigations or pilot studies or when there is *a priori* justification (for example, a theory or previous findings). The following example demonstrates a directional hypothesis test with a *t* statistic, using the same experimental situation presented in Example 9.1.

EXAMPLE 9.4

The research question is whether attractiveness affects the behavior of infants looking at photographs of women's faces. The researcher is expecting the infants to prefer the more attractive face. Therefore, the researcher predicts that the infants will spend more than half of the 20-second period looking at the attractive face. For this example, we use the same sample data that were used in the original hypothesis test in Example 9.1. Specifically, the researcher tested a sample of $n = 9$ infants and obtained a mean of $M = 13$ seconds looking at the attractive face with $SS = 72$.

STEP 1 State the hypotheses, and select an alpha level. With most directional tests, it is usually easier to state the hypothesis in words, including the directional prediction, and then convert the words into symbols. For this example, the researcher is predicting that attractiveness will cause the infants to increase the amount of time they spend looking at the attractive face; that is, more than half of the 20 seconds should be spent looking at the attractive face. In general, the null hypothesis states that the predicted effect will not happen. For this study, the null hypothesis states that the infants will not spend more than half of the 20 seconds looking at the attractive face. In symbols,

H_0: $\mu_{\text{attractive}} \leq 10$ seconds (Not more than half of the 20 seconds looking at the attractive face)

Similarly, the alternative states that the treatment will work. In this case, H_1 states that the infants will spend more than half of the time looking at the attractive face. In symbols,

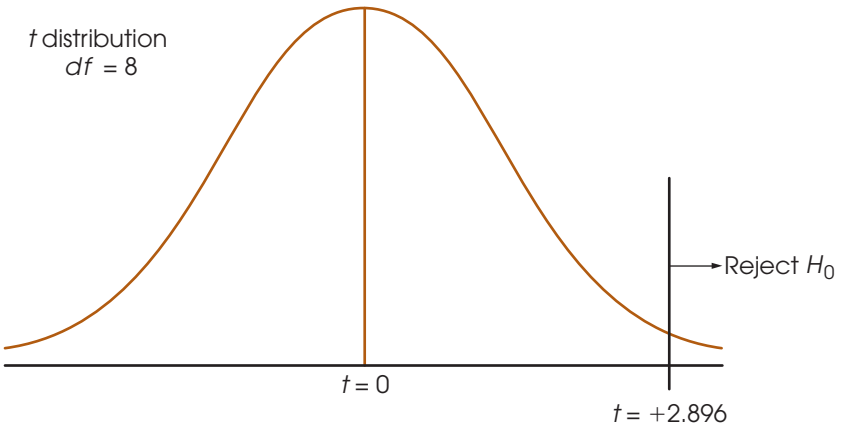
H_1: $\mu_{\text{attractive}} > 10$ seconds (More than half of the 20 seconds looking at the attractive face)

This time, we set the level of significance at $\alpha = .01$.

STEP 2 Locate the critical region. In this example, the researcher is predicting that the sample mean (M) will be greater than 10 seconds. Thus, if the infants average more than 10 seconds looking at the attractive face, the data will provide support for the researcher's prediction and will tend to refute the null hypothesis. Also note that a sample mean greater than 10 will produce a positive value for the *t* statistic. Thus, the critical region for the one-tailed test will consist of positive *t* values located in the right-hand tail of the distribution. To find the critical value, you must look in the *t* distribution table using the one-tail proportions. With a sample of $n = 9$, the *t* statistic has $df = 8$; using $\alpha = .01$, you should find a critical value of $t = 2.896$. Therefore, if we obtain a *t* statistic greater than 2.896, we will reject the null hypothesis and conclude that the infants show a significant preference for the attractive face. Figure 9.8 shows the one-tailed critical region for this test.

FIGURE 9.8

The one-tailed critical region for the hypothesis test in Example 9.4 with $df = 8$ and $\alpha = .01$.



STEP 3 Calculate the test statistic. The computation of the t statistic is the same for either a one-tailed or a two-tailed test. Earlier (in Example 9.1), we found that the data for this experiment produce a test statistic of $t = 3.00$.

STEP 4 Make a decision. The test statistic is in the critical region, so we reject H_0. In terms of the experimental variables, we have decided that the infants show a preference and spend significantly more time looking at the attractive face than they do looking at the unattractive face. In a research report, the results would be presented as follows:

> The time spent looking at the attractive face was significantly greater than would be expected if there were no preference, $t(8) = 3.00$, $p < .01$, one tailed.

Note that the report clearly acknowledges that a one-tailed test was used.

THE CRITICAL REGION FOR A ONE-TAILED TEST

In step 2 of Example 9.4, we determined that the critical region is in the right-hand tail of the distribution. However, it is possible to divide this step into two stages that eliminate the need to determine which tail (right or left) should contain the critical region. The first stage in this process is simply to determine whether the sample mean is in the direction predicted by the original research question. For this example, the researcher predicted that the infants would prefer the attractive face and spend more time looking at it. Specifically, the researcher expects the infants to spend more than 10 out of 20 seconds focused on the attractive face. The obtained sample mean, $M = 13$ seconds, is in the correct direction. This first stage eliminates the need to determine whether the critical region is in the left- or right-hand tail. Because we already have determined that the effect is in the correct direction, the sign of the t statistic (+ or –) no longer matters. The second stage of the process is to determine whether the effect is large enough to be significant. For this example, the requirement is that the sample produces a t statistic greater than 2.896. If the magnitude of the t statistic, independent of its sign, is greater than 2.896, then the result is significant and H_0 is rejected.

LEARNING CHECK

1. A new over-the-counter cold medication includes a warning label stating that it "may cause drowsiness." A researcher would like to evaluate this effect. It is known that under regular circumstances the distribution of reaction times is normal with $\mu = 200$. A sample of $n = 9$ participants is obtained. Each person is given the new cold medication, and, 1 hour later, reaction time is measured for each individual. The average reaction time for this sample is $M = 206$ with $SS = 648$. The researcher would like to use a hypothesis test with $\alpha = .05$ to evaluate the effect of the medication.

a. Use a two-tailed test with $\alpha = .05$ to determine whether the medication has a significant effect on reaction time.

b. Write a sentences that demonstrates how the outcome of the hypothesis test would appear in a research report.

c. Use a one-tailed test with $\alpha = .05$ to determine whether the medication produces a significant increase in reaction time.

d. Write a sentence that demonstrates how the outcome of the one-tailed hypothesis test would appear in a research report.

ANSWER

1. a. For the two-tailed test, H_0: $\mu = 200$. The sample variance is 81, the estimated standard error is 3, and $t = 6/3 = 2.00$. With $df = 8$, the critical boundaries are ± 2.306. Fail to reject the null hypothesis.

b. The result indicates that the medication does not have a significant effect on reaction time, $t(8) = 2.00$, $p > .05$.

c. For a one-tailed test, H_0: $\mu \leq 200$ (no increase). The data product $t = 6/3 = 2.00$. With $df = 8$, the critical boundary is 1.860. Reject the null hypothesis.

d. The results indicate that the medication produces a significant increase in reaction time, $t(8) = 2.00$, $p < .05$, one tailed.

SUMMARY

1. The *t* statistic is used instead of a *z*-score for hypothesis testing when the population standard deviation (or variance) is unknown.

2. To compute the *t* statistic, you must first calculate the sample variance (or standard deviation) as a substitute for the unknown population value.

$$\text{sample variance} = s^2 = \frac{SS}{df}$$

Next, the standard error is *estimated* by substituting s^2 in the formula for standard error. The estimated standard error is calculated in the following manner:

$$\text{estimated standard error} = s_M = \sqrt{\frac{s^2}{n}}$$

Finally, a *t* statistic is computed using the estimated standard error. The *t* statistic is used as a substitute for a *z*-score that cannot be computed when the population variance or standard deviation is unknown.

$$t = \frac{M - \mu}{s_M}$$

The structure of the *t* formula is similar to that of the *z*-score in that

$$z \text{ or } t = \frac{\text{sample mean} - \text{population mean}}{\text{(estimated) standard error}}$$

For a hypothesis test, you hypothesize a value for the unknown population mean and plug the hypothesized value into the equation along with the sample mean and

the estimated standard error, which are computed from the sample data. If the hypothesized mean produces an extreme value for t, then you conclude that the hypothesis was wrong.

4. The t distribution is an approximation of the normal z distribution. To evaluate a t statistic that is obtained for a sample mean, the critical region must be located in a t distribution. There is a family of t distributions, with the exact shape of a particular distribution of t values depending on degrees of freedom $(n - 1)$. Therefore, the critical t values depend on the value for df associated with the t test. As df increases, the shape of the t distribution approaches a normal distribution.

5. When a t statistic is used for a hypothesis test, Cohen's d can be computed to measure effect size. In this situation, the sample standard deviation is used in the formula to obtain an estimated value for d:

$$\text{estimated } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{M - \mu}{s}$$

6. A second measure of effect size is r^2, which measures the percentage of the variability that is accounted for by the treatment effect. This value is computed as follows:

$$r^2 = \frac{t^2}{t^2 + df}$$

7. An alternative method for describing the size of a treatment effect is to use a confidence interval for μ. A confidence interval is a range of values that estimates the unknown population mean. The confidence interval uses the t equation, solved for the unknown mean:

$$\mu = M \pm t(s_M)$$

First, select a level of confidence and then look up the corresponding t values to use in the equation. For example, for 95% confidence, use the range of t values that determine the middle 95% of the distribution.

KEY TERMS

estimated standard error (286)
t statistic (286)
degrees of freedom (287)
t distribution (287)
estimated d (295)
percentage of variance accounted for by the treatment (r^2) (299)
confidence interval (300)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 9 on the book companion website. The website also provides access to a workshop entitled *Single-Sample* t *Test,* which reviews the concepts and logic of hypothesis testing with the t statistic.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform the ***t* Test** presented in this chapter.

Data Entry

Enter all of the scores from the sample in one column of the data editor, probably VAR00001.

Data Analysis

1. Click Analyze on the tool bar, select **Compare Means,** and click on **One-Sample T Test.**
2. Highlight the column label for the set of scores (VAR0001) in the left box and click the arrow to move it into the **Test Variable(s)** box.
3. In the **Test Value** box at the bottom of the One-Sample *t* Test window, enter the hypothesized value for the population mean from the null hypothesis. *Note:* The value is automatically set at zero until you type in a new value.
4. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95%, but you can select **Options** and change the percentage.
5. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the infants-and-attractive-faces study in Example 9.1 and the program output is shown in Figure 9.9. The output includes a table of sample statistics with the mean, standard deviation, and standard error for the sample mean. A second table shows the results of the hypothesis test, including the values for *t*, *df*, and the level of significance (the *p* value for the test), as well as the mean difference from the hypothesized value of $\mu = 10$ and a 95% confidence interval for the mean difference. To obtain a 95% confidence interval for the mean, simply add $\mu = 10$ points to the values in the table.

FOCUS ON PROBLEM SOLVING

1. The first problem we confront in analyzing data is determining the appropriate statistical test. Remember that you can use a *z*-score for the test statistic only when the value for σ is known. If the value for σ is not provided, then you must use the *t* statistic.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	9	13.0000	3.00000	1.00000

One-Sample Test

	Test Value = 10					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
VAR00001	3.000	8	.017	3.00000	.6940	5.3060

FIGURE 9.9

The SPSS output for the hypothesis test presented in Example 9.1.

2. For the *t* test, the sample variance is used to find the value for the estimated standard error. Remember to use $n - 1$ in the denominator when computing the sample variance (see Chapter 4). When computing estimated standard error, use n in the denominator.

DEMONSTRATION 9.1

A HYPOTHESIS TEST WITH THE *t* STATISTIC

A psychologist has prepared an "Optimism Test" that is administered yearly to graduating college seniors. The test measures how each graduating class feels about its future—the higher the score, the more optimistic the class. Last year's class had a mean score of $\mu = 15$. A sample of $n = 9$ seniors from this year's class was selected and tested. The scores for these seniors are 7, 12, 11, 15, 7, 8, 15, 9, and 6, which produce a sample mean of $M = 10$ with $SS = 94$.

On the basis of this sample, can the psychologist conclude that this year's class has a different level of optimism than last year's class?

Note that this hypothesis test uses a *t* statistic because the population variance (σ^2) is not known.

STEP 1 **State the hypotheses, and select an alpha level.** The statements for the null hypothesis and the alternative hypothesis follow the same form for the *t* statistic and the *z*-score test.

H_0: $\mu = 15$ (There is no change.)

H_1: $\mu \neq 15$ (This year's mean is different.)

For this demonstration, we use $\alpha = .05$, two tails.

STEP 2 **Locate the critical region.** With a sample of $n = 9$ students, the t statistic has $df = n - 1 = 8$. For a two-tailed test with $\alpha = .05$ and $df = 8$, the critical t values are $t = \pm 2.306$. These critical t values define the boundaries of the critical region. The obtained t value must be more extreme than either of these critical values to reject H_0.

STEP 3 **Compute the test statistic.** As we have noted, it is easier to separate the calculation of the t statistic into three stages.

Sample variance.

$$s^2 = \frac{SS}{n-1} = \frac{94}{8} = 11.75$$

Estimated standard error. The estimated standard error for these data is

$$s_M = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{11.75}{9}} = 1.14$$

The t statistic. Now that we have the estimated standard error and the sample mean, we can compute the t statistic. For this demonstration,

$$t = \frac{M - \mu}{s_M} = \frac{10 - 15}{1.14} = \frac{-5}{1.14} = -4.39$$

STEP 4 **Make a decision about H_0, and state a conclusion.** The t statistic we obtained ($t = -4.39$) is in the critical region. Thus, our sample data are unusual enough to reject the null hypothesis at the .05 level of significance. We can conclude that there is a significant difference in level of optimism between this year's and last year's graduating classes, $t(8) = -4.39$, $p < .05$, two-tailed.

DEMONSTRATION 9.2

EFFECT SIZE: ESTIMATING COHEN'S *d* AND COMPUTING r^2

We will estimate Cohen's d for the same data used for the hypothesis test in Demonstration 9.1. The mean optimism score for the sample from this year's class was 5 points lower than the mean from last year ($M = 10$ versus $\mu = 15$). In Demonstration 9.1, we computed a sample variance of $s^2 = 11.75$, so the standard deviation is $\sqrt{11.75} = 3.43$. With these values,

$$\text{estimated } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{5}{3.43} = 1.46$$

To calculate the percentage of variance explained by the treatment effect, r^2, we need the value of t and the df value from the hypothesis test. In Demonstration 9.1, we obtained $t = -4.39$ with $df = 8$. Using these values in Equation 9.5, we obtain

$$r^2 = \frac{t^2}{t^2 + df} = \frac{(-4.39)^2}{(-4.39)^2 + 8} = \frac{19.27}{27.27} = 0.71$$

PROBLEMS

1. Under what circumstances is a t statistic used instead of a z-score for a hypothesis test?

2. A sample of $n = 25$ scores has a mean of $M = 83$ and a standard deviation of $s = 15$.
- **a.** Explain what is measured by the sample standard deviation.
- **b.** Compute the estimated standard error for the sample mean and explain what is measured by the standard error.

3. Find the estimated standard error for the sample mean for each of the following samples.
- **a.** $n = 4$ with $SS = 48$
- **b.** $n = 6$ with $SS = 270$
- **c.** $n = 12$ with $SS = 132$

4. Explain why t distributions tend to be flatter and more spread out than the normal distribution.

5. Find the t values that form the boundaries of the critical region for a two-tailed test with $\alpha = .05$ for each of the following sample sizes:
- **a.** $n = 6$
- **b.** $n = 12$
- **c.** $n = 24$

6. The following sample of $n = 6$ scores was obtained from a population with unknown parameters.

Scores: 7, 1, 6, 3, 6, 7
- **a.** Compute the sample mean and standard deviation. (Note that these are descriptive values that summarize the sample data.)
- **b.** Compute the estimated standard error for M. (Note that this is an inferential value that describes how accurately the sample mean represents the unknown population mean.)

7. The following sample was obtained from a population with unknown parameters.

Scores: 6, 12, 0, 3, 4
- **a.** Compute the sample mean and standard deviation. (Note that these are descriptive values that summarize the sample data.)
- **b.** Compute the estimated standard error for M. (Note that this is an inferential value that describes how accurately the sample mean represents the unknown population mean.)

8. To evaluate the effect of a treatment, a sample is obtained from a population with a mean of $\mu = 75$, and the treatment is administered to the individuals in the sample. After treatment, the sample mean is found to be $M = 79.6$ with a standard deviation of $s = 12$.
- **a.** If the sample consists of $n = 16$ individuals, are the data sufficient to conclude that the treatment has a significant effect using a two-tailed test with $\alpha = .05$?
- **b.** If the sample consists of $n = 36$ individuals, are the data sufficient to conclude that the treatment has a significant effect using a two-tailed test with $\alpha = .05$?
- **c.** Comparing your answer for parts a and b, how does the size of the sample influence the outcome of a hypothesis test?

9. To evaluate the effect of a treatment, a sample of $n = 9$ is obtained from a population with a mean of $\mu = 40$, and the treatment is administered to the individuals in the sample. After treatment, the sample mean is found to be $M = 33$.
- **a.** If the sample has a standard deviation of $s = 9$, are the data sufficient to conclude that the treatment has a significant effect using a two-tailed test with $\alpha = .05$?
- **b.** If the sample standard deviation is $s = 15$, are the data sufficient to conclude that the treatment has a significant effect using a two-tailed test with $\alpha = .05$?
- **c.** Comparing your answer for parts a and b, how does the variability of the scores in the sample influence the outcome of a hypothesis test?

10. A random sample of $n = 16$ individuals is selected from a population with $\mu = 70$, and a treatment is administered to each individual in the sample. After treatment, the sample mean is found to be $M = 76$ with $SS = 960$.
- **a.** How much difference is there between the mean for the treated sample and the mean for the original population? (*Note:* In a hypothesis test, this value forms the numerator of the t statistic.)
- **b.** How much difference is expected just by chance between the sample mean and its population mean? That is, find the standard error for M. (*Note:* In a hypothesis test, this value is the denominator of the t statistic.)
- **c.** Based on the sample data, does the treatment have a significant effect? Use a two-tailed test with $\alpha = .05$.

11. The spotlight effect refers to overestimating the extent to which others notice your appearance or behavior, especially when you commit a social faux pas. Effectively, you feel as if you are suddenly standing in a spotlight with everyone looking. In one demonstration of this phenomenon, Gilovich, Medvec, and Savitsky (2000) asked college students to put on a Barry Manilow T-shirt that fellow students had previously judged to be embarrassing. The participants were

then led into a room in which other students were already participating in an experiment. After a few minutes, the participant was led back out of the room and was allowed to remove the shirt. Later, each participant was asked to estimate how many people in the room had noticed the shirt. The individuals who were in the room were also asked whether they noticed the shirt. In the study, the participants significantly overestimated the actual number of people who had noticed.

a. In a similar study using a sample of $n = 9$ participants, the individuals who wore the shirt produced an average estimate of $M = 6.4$ with $SS = 162$. The average number who said they noticed was 3.1. Is the estimate from the participants significantly different from the actual number? Test the null hypothesis that the true mean is $\mu = 3.1$ using a two-tailed test with $\alpha = .05$.

b. Is the estimate from the participants significantly higher than the actual number ($\mu = 3.1$)? Use a one-tailed test with $\alpha = .05$.

12. Many animals, including humans, tend to avoid direct eye contact and even patterns that look like eyes. Some insects, including moths, have evolved eye-spot patterns on their wings to help ward off predators. Scaife (1976) reports a study examining how eye-spot patterns affect the behavior of birds. In the study, the birds were tested in a box with two chambers and were free to move from one chamber to another. In one chamber, two large eye-spots were painted on one wall. The other chamber had plain walls. The researcher recorded the amount of time each bird spent in the plain chamber during a 60-minute session. Suppose the study produced a mean of $M = 37$ minutes in the plain chamber with $SS = 288$ for a sample of $n = 9$ birds. (*Note:* If the eye-spots have no effect, then the birds should spend an average of $\mu = 30$ minutes in each chamber.)

a. Is this sample sufficient to conclude that the eye-spots have a significant influence on the birds' behavior? Use a two-tailed test with $\alpha = .05$.

b. Compute the estimated Cohen's d to measure the size of the treatment effect.

c. Construct the 95% confidence interval to estimate the mean amount of time spent on the plain side for the population of birds.

13. Standardized measures seem to indicate that the average level of anxiety has increased gradually over the past 50 years (Twenge, 2000). In the 1950s, the average score on the Child Manifest Anxiety Scale was $\mu = 15.1$. A sample of $n = 16$ of today's children produces a mean score of $M = 23.3$ with $SS = 240$.

a. Based on the sample, has there been a significant change in the average level of anxiety since the 1950s? Use a two-tailed test with $\alpha = .01$.

b. Make a 90% confidence interval estimate of today's population mean level of anxiety.

c. Write a sentence that demonstrates how the outcome of the hypothesis test and the confidence interval would appear in a research report.

14. The librarian at the local elementary school claims that, on average, the books in the library are more than 20 years old. To test this claim, a student takes a sample of $n = 30$ books and records the publication date for each. The sample produces an average age of $M = 23.8$ years with a variance of $s^2 = 67.5$. Use this sample to conduct a one-tailed test with $\alpha = .01$ to determine whether the average age of the library books is significantly greater than 20 years ($\mu > 20$).

15. For several years researchers have noticed that there appears to be a regular, year-by-year increase in the average IQ for the general population. This phenomenon is called the Flynn effect after the researcher who first reported it (Flynn, 1984, 1999), and it means that psychologists must continuously update IQ tests to keep the population mean at $\mu = 100$. To evaluate the size of the effect, a researcher obtained a 10-year-old IQ test that was standardized to produce a mean IQ of $\mu = 100$ for the population 10 years ago. The test was then given to a sample of $n = 64$ of today's 20-year-old adults. The average score for the sample was $M = 107$ with a standard deviation of $s = 12$.

a. Based on the sample, is the average IQ for today's population significantly different from the average 10 years ago, when the test would have produces a mean of $\mu = 100$? Use a two-tailed test with $\alpha = .01$.

b. Make an 80% confidence interval estimate of today's population mean IQ for the 10-year-old test.

16. In a classic study of infant attachment, Harlow (1959) placed infant monkeys in cages with two artificial surrogate mothers. One "mother" was made from bare wire mesh and contained a baby bottle from which the infants could feed. The other mother was made from soft terry cloth and did not provide any access to food. Harlow observed the infant monkeys and recorded how much time per day was spent with each mother. In a typical day, the infants spent a total of 18 hours clinging to one of the two mothers. If there were no preference between the two, you would expect the time to be divided evenly, with an average of $\mu = 9$ hours for each of the mothers. However, the typical monkey spent around 15 hours per day with the terry-cloth mother, indicating a strong preference for the soft, cuddly mother. Suppose a sample of $n = 9$ infant monkeys averaged $M = 15.3$ hours per day with $SS = 216$ with the terry-cloth mother. Is this result sufficient to conclude that the monkeys spent significantly more time

with the softer mother than would be expected if there were no preference? Use a two-tailed test with $\alpha = .05$.

17. Belsky, Weinraub, Owen, and Kelly (2001) reported on the effects of preschool childcare on the development of young children. One result suggests that children who spend more time away from their mothers are more likely to show behavioral problems in kindergarten. Using a standardized scale, the average rating of behavioral problems for kindergarten children is $\mu = 35$. A sample of $n = 16$ kindergarten children who had spent at least 20 hours per week in childcare during the previous year produced a mean score of $M = 42.7$ with a standard deviation of $s = 6$.

- **a.** Are the data sufficient to conclude that children with a history of childcare show significantly more behavioral problems than the average kindergarten child? Use a one-tailed test with $\alpha = .01$.
- **b.** Compute r^2, the percentage of variance accounted for, to measure the size of the preschool effect.
- **c.** Write a sentence showing how the outcome of the hypothesis test and the measure of effect size would appear in a research report.

18. Other research examining the effects of preschool childcare has found that children who spent time in day care, especially high-quality day care, perform better on math and language tests than children who stay home with their mothers (Broberg, Wessels, Lamb, & Hwang, 1997). Typical results, for example, show that a sample of $n = 25$ children who attended day care before starting school had an average score of $M = 87$ with $SS = 1536$ on a standardized math test for which the population mean is $\mu = 81$.

- **a.** Is this sample sufficient to conclude that the children with a history of preschool day care are significantly different from the general population? Use a two-tailed test with $\alpha = .01$.
- **b.** Compute Cohen's d to measure the size of the preschool effect.
- **c.** Write a sentence showing how the outcome of the hypothesis test and the measure of effect size would appear in a research report.

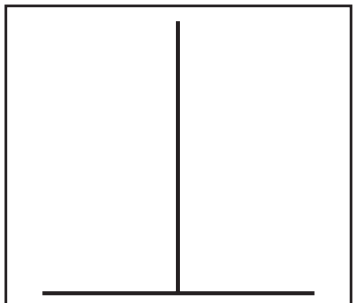
19. A random sample of $n = 25$ scores is obtained from a population with a mean of $\mu = 45$. A treatment is administered to the individuals in the sample and, after treatment, the sample mean is $M = 48$.

- **a.** Assuming that the sample standard deviation is $s = 6$ compute r^2 and the estimated Cohen's d to measure the size of the treatment effect.
- **b.** Assuming that the sample standard deviation is $s = 15$, compute r^2 and the estimated Cohen's d to measure the size of the treatment effect.
- **c.** Comparing your answers from parts a and b, how does the variability of the scores in the sample influence the measures of effect size?

20. A random sample is obtained from a population with a mean of $\mu = 70$. A treatment is administered to the individuals in the sample and, after treatment, the sample mean is $M = 78$ with a standard deviation of $s = 20$.

- **a.** Assuming that the sample consists of $n = 25$ scores, compute r^2 and the estimated Cohen's d to measure the size of treatment effect.
- **b.** Assuming that the sample consists of $n = 16$ scores, compute r^2 and the estimated Cohen's d to measure the size of treatment effect.
- **c.** Comparing your answers from parts a and b, how does the number of scores in the sample influence the measures of effect size?

21. An example of the vertical-horizontal illusion is shown in the figure below. Although the two lines are exactly the same length, the vertical line appears to be much longer. To examine the strength of this illusion, a researcher prepared an example in which both lines were exactly 10 inches long. The example was shown to individual participants who were told that the horizontal line was 10 inches long and then were asked to estimate the length of the vertical line. For a sample of $n = 25$ participants, the average estimate was $M = 12.2$ inches with a standard deviation of $s = 1.00$.



An example of the vertical-horizontal illusion

- **a.** Use a one-tailed hypothesis test with $\alpha = .01$ to demonstrate that the individuals in the sample significantly overestimate the true length of the line. (*Note:* Accurate estimation would produce a mean of $\mu = 10$ inches.)
- **b.** Calculate the estimated d and r^2, the percentage of variance accounted for, to measure the size of this effect.
- **c.** Construct a 95% confidence interval for the population mean estimated length of the vertical line.

22. In studies examining the effect of humor on interpersonal attractions, McGee and Shevlin (2009) found that an individual's sense of humor had a

significant effect on how the individual was perceived by others. In one part of the study, female college students were given brief descriptions of a potential romantic partner. The fictitious male was described positively as being single, ambitious, and having good job prospects. For one group of participants, the description also said that he had a great sense of humor. For another group, it said that he had no sense of humor. After reading the description, each participant was asked to rate the attractiveness of the man on a seven-point scale from 1 (very attractive) to 7 (very unattractive). A score of 4 indicates a neutral rating.

a. The females who read the "great sense of humor" description gave the potential partner an average attractiveness score of $M = 4.53$ with a standard deviation of $s = 1.04$. If the sample consisted of $n = 16$ participants, is the average rating significantly higher than neutral ($\mu = 4$)? Use a one-tailed test with $\alpha = .05$.

b. The females who read the description saying "no sense of humor" gave the potential partner an average attractiveness score of $M = 3.30$ with a standard deviation of $s = 1.18$. If the sample consisted of $n = 16$ participants, is the average rating significantly lower than neutral ($\mu = 4$)? Use a one-tailed test with $\alpha = .05$.

23. A psychologist would like to determine whether there is a relationship between depression and aging. It is known that the general population averages $\mu = 40$ on a standardized depression test. The psychologist obtains a sample of $n = 9$ individuals who are all more than 70 years old. The depression scores for this sample are as follows: 37, 50, 43, 41, 39, 45, 49, 44, 48.

a. On the basis of this sample, is depression for elderly people significantly different from depression in the general population? Use a two-tailed test with $\alpha = .05$.

b. Compute the estimated Cohen's *d* to measure the size of the difference.

c. Write a sentence showing how the outcome of the hypothesis test and the measure of effect size would appear in a research report.

CHAPTER

10

The *t* Test for Two Independent Samples

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Sample variance (Chapter 4)
- Standard error formulas (Chapter 7)
- The *t* statistic (Chapter 9)
  - Distribution of *t* values
  - *df* for the *t* statistic
  - Estimated standard error

Preview

In a classic study in the area of problem solving, Katona (1940) compared the effectiveness of two methods of instruction. One group of participants was shown the exact, step-by-step procedure for solving a problem, and then these participants were required to memorize the solution. This method was called *learning by memorization* (later called the *expository* method). Participants in a second group were encouraged to study the problem and find the solution on their own. Although these participants were given helpful hints and clues, the exact solution was never explained. This method was called *learning by understanding* (later called the *discovery* method).

Katona's experiment included the problem shown in Figure 10.1. This figure shows a pattern of five squares made of matchsticks. The problem is to change the pattern into exactly four squares by moving only three matches. (All matches must be used, none can be removed, and all the squares must be the same size.) Two groups of participants learned the solution to this problem. One group learned by understanding, and the other group learned by memorization. After 3 weeks, both groups returned to be tested again. The two groups did equally well on the matchstick problem they had learned earlier. But when they were given two new problems (similar to the matchstick problem), the *understanding* group performed much better than the *memorization* group.

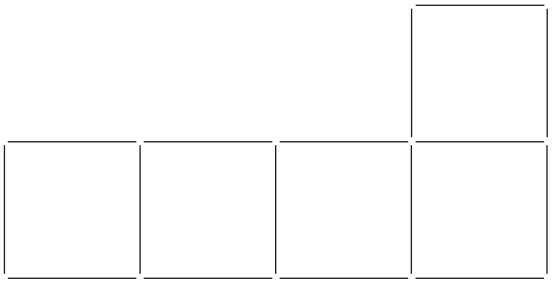


FIGURE 10.1

A pattern of five squares made of matchsticks. The problem is to change the pattern into exactly four squares by moving only three matchsticks.

The Problem: Although the data a show a mean difference between the two groups in Katona's study, you cannot automatically conclude that the difference was caused by the method they used to solve the first problem. Specifically, the two groups consist of different people with different backgrounds, different skills, different IQs, and so on. Because the two different groups consist of different individuals, you should expect them to have different scores and different means. This issue was first presented in Chapter 1 when we introduced the concept of sampling error (see Figure 1.2 on p. 9). Thus, there are two possible explanations for the difference between the two groups.

1. It is possible that there really is a difference between the two treatment conditions so that the method of understanding produces better learning than the method of memorization.
2. It is possible that there is no difference between the two treatment conditions and the mean difference obtained in the experiment is simply the result of sampling error.

A hypothesis test is necessary to determine which of the two explanations is most plausible. However, the hypothesis tests we have examined thus far are intended to evaluate the data from only one sample. In this study there are two separate samples.

The Solution: In this chapter we introduce the *independent-measures* t *test,* which is a hypothesis test that uses two separate samples to evaluate the mean difference between two treatment conditions or between two different populations. Like the *t* test introduced in Chapter 9, the independent-measures *t* test uses the sample variance to compute an estimated standard error. This test, however, combines the variance from the two separate samples to evaluate the difference between two separate sample means.

Incidentally, if you still have not discovered the solution to the matchstick problem, keep trying. According to Katona's results, it would be a very poor teaching strategy for us to give you the answer to the matchstick problem. If you still have not discovered the solution, however, check Appendix C at the beginning of the Chapter 10 problem solutions; there we show you how it is done.

10.1 INTRODUCTION TO THE INDEPENDENT-MEASURES DESIGN

Until this point, all the inferential statistics we have considered involve using one sample as the basis for drawing conclusions about one population. Although these *single-sample* techniques are used occasionally in real research, most research studies require the comparison of two (or more) sets of data. For example, a social psychologist may want to compare men and women in terms of their political attitudes, an educational psychologist may want to compare two methods for teaching mathematics, or a clinical psychologist may want to evaluate a therapy technique by comparing depression scores for patients before therapy with their scores after therapy. In each case, the research question concerns a mean difference between two sets of data.

There are two general research designs that can be used to obtain the two sets of data to be compared:

1. The two sets of data could come from two completely separate groups of participants. For example, the study could involve a sample of men compared with a sample of women. Or the study could compare grades for one group of freshmen who are given laptop computers with grades for a second group who are not given computers.
2. The two sets of data could come from the same group of participants. For example, the researcher could obtain one set of scores by measuring depression for a sample of patients before they begin therapy and then obtain a second set of data by measuring the same individuals after 6 weeks of therapy.

The first research strategy, using completely separate groups, is called an *independent-measures* research design or a *between-subjects* design. These terms emphasize the fact that the design involves separate and independent samples and makes a comparison between two groups of individuals. The structure of an independent-measures research design is shown in Figure 10.2. Notice that the research study uses two separate

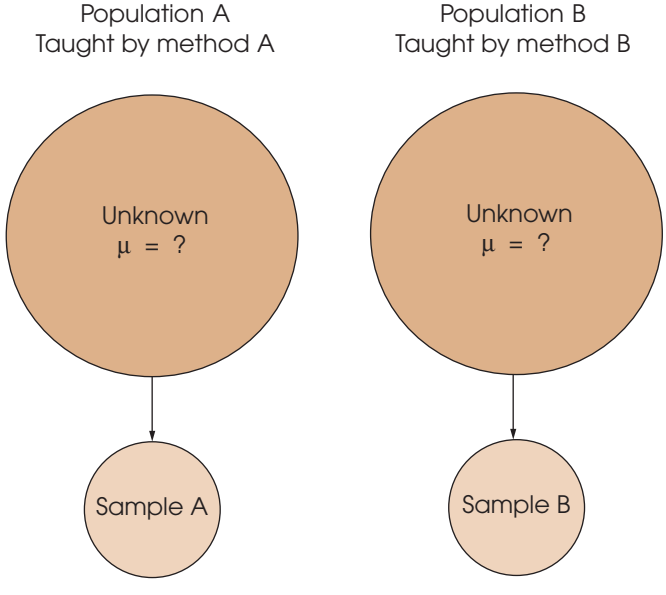


FIGURE 10.2

Do the achievement scores for children taught by method A differ from the scores for children taught by method B? In statistical terms, are the two population means the same or different? Because neither of the two population means is known, it will be necessary to take two samples, one from each population. The first sample provides information about the mean for the first population, and the second sample provides information about the second population.

samples to represent the two different populations (or two different treatments) being compared.

DEFINITION

A research design that uses a separate group of participants for each treatment condition (or for each population) is called an **independent-measures research design** or a **between-subjects research design.**

In this chapter, we examine the statistical techniques used to evaluate the data from an independent-measures design. More precisely, we introduce the hypothesis test that allows researchers to use the data from two separate samples to evaluate the mean difference between two populations or between two treatment conditions.

The second research strategy, in which the two sets of data are obtained from the same group of participants, is called a *repeated-measures* research design or a *within-subjects* design. The statistics for evaluating the results from a repeated-measures design are introduced in Chapter 11. Also, at the end of Chapter 11, we discuss some of the advantages and disadvantages of independent-measures and repeated-measures designs.

10.2 THE *t* STATISTIC FOR AN INDEPENDENT-MEASURES RESEARCH DESIGN

Because an independent-measures study involves two separate samples, we need some special notation to help specify which data go with which sample. This notation involves the use of subscripts, which are small numbers written beside a sample statistic. For example, the number of scores in the first sample would be identified by n_1; for the second sample, the number of scores is n_2. The sample means would be identified by M_1 and M_2. The sums of squares would be SS_1 and SS_2.

THE HYPOTHESIS FOR AN INDEPENDENT-MEASURES TEST

The goal of an independent-measures research study is to evaluate the mean difference between two populations (or between two treatment conditions). Using subscripts to differentiate the two populations, the mean for the first population is μ_1, and the second population mean is μ_2. The difference between means is simply $\mu_1 - \mu_2$. As always, the null hypothesis states that there is no change, no effect, or, in this case, no difference. Thus, in symbols, the null hypothesis for the independent-measures test is

$$H_0: \mu_1 - \mu_2 = 0 \text{ (No difference between the population means)}$$

You should notice that the null hypothesis could also be stated as $\mu_1 = \mu_2$. However, the first version of H_0 produces a specific numerical value (zero) that is used in the calculation of the t statistic. Therefore, we prefer to phrase the null hypothesis in terms of the difference between the two population means.

The alternative hypothesis states that there is a mean difference between the two populations,

$$H_1: \mu_1 - \mu_2 \neq 0 \text{ (There is a mean difference.)}$$

Equivalently, the alternative hypothesis can simply state that the two population means are not equal: $\mu_1 \neq \mu_2$.

THE FORMULAS FOR AN INDEPENDENT-MEASURES HYPOTHESIS TEST

The independent-measures hypothesis test uses another version of the *t* statistic. The formula for this new *t* statistic has the same general structure as the *t* statistic formula that was introduced in Chapter 9. To help distinguish between the two *t* formulas, we refer to the original formula (Chapter 9) as the *single-sample* t *statistic* and we refer to the new formula as the *independent-measures* t *statistic*. Because the new independent-measures *t* includes data from two separate samples and hypotheses about two populations, the formulas may appear to be a bit overpowering. However, the new formulas are easier to understand if you view them in relation to the single-sample *t* formulas from Chapter 9. In particular, there are two points to remember:

1. The basic structure of the *t* statistic is the same for both the independent-measures and the single-sample hypothesis tests. In both cases,

$$t = \frac{\text{sample statistic} - \text{hypothesized population parameter}}{\text{estimated standard error}}$$

2. The independent-measures *t* is basically a *two-sample* t *that doubles all the elements of the single-sample* t *formulas*.

To demonstrate the second point, we examine the two *t* formulas piece by piece.

The overall *t* formula The single-sample *t* uses one sample mean to test a hypothesis about one population mean. The sample mean and the population mean appear in the numerator of the *t* formula, which measures how much difference there is between the sample data and the population hypothesis.

$$t = \frac{\text{sample mean} - \text{population mean}}{\text{estimated standard error}} = \frac{M - \mu}{s_M}$$

The independent-measures *t* uses the difference between *two* sample means to evaluate a hypothesis about the difference between *two* population means. Thus, the independent-measures *t* formula is

$$t = \frac{\text{sample mean difference} - \text{population mean difference}}{\text{estimated standard error}} = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

In this formula, the value of $M_1 - M_2$ is obtained from the sample data and the value for $\mu_1 - \mu_2$ comes from the null hypothesis.

The estimated standard error In each of the *t*-score formulas, the standard error in the denominator measures how accurately the sample statistic represents the population parameter. In the single-sample *t* formula, the standard error measures the amount of error expected for a sample mean and is represented by the symbol s_M. For the independent-measures *t* formula, the standard error measures the amount of error that is expected when you use a sample mean difference ($M_1 - M_2$) to represent a population mean difference ($\mu_1 - \mu_2$). The standard error for the sample mean difference is represented by the symbol $s_{(M_1 - M_2)}$.

Caution: Do not let the notation for standard error confuse you. In general, standard error measures how accurately a statistic represents a parameter. The symbol for standard error takes the form $s_{\text{statistic}}$. When the statistic is a sample mean, *M*, the symbol for standard error is s_M. For the independent-measures test, the statistic is a sample

mean difference $(M_1 - M_2)$, and the symbol for standard error is $s_{(M_1-M_2)}$. In each case, the standard error tells how much discrepancy is reasonable to expect between the sample statistic and the corresponding population parameter.

Interpreting the estimated standard error The *estimated standard error of* $M_1 - M_2$ that appears in the bottom of the independent-measures *t* statistic can be interpreted in two ways. First, the standard error is defined as a measure of the standard, or average, distance between a sample statistic $(M_1 - M_2)$ and the corresponding population parameter $(\mu_1 - \mu_2)$. As always, samples are not expected to be perfectly accurate and the standard error measures how much difference is reasonable to expect between a sample statistic and the population parameter.

$$\begin{array}{c}\text{Sample mean}\\ \text{difference}\\ (M_1 - M_2)\end{array} \xleftrightarrow[\text{(average distance)}]{\text{estimated standard error}} \begin{array}{c}\text{Population mean}\\ \text{difference}\\ (\mu_1 - \mu_2)\end{array}$$

When the null hypothesis is true, however, the population mean difference is zero.

$$\begin{array}{c}\text{Sample mean}\\ \text{difference}\\ (M_1 - M_2)\end{array} \xleftrightarrow[\text{(average distance)}]{\text{estimated standard error}} 0\big(\text{If } H_0 \text{ is true}\big)$$

The standard error is measuring how close the sample mean difference is to zero, which is equivalent to measuring how much difference there is between the two sample means.

$$M_1 \xleftrightarrow[\text{(average distance)}]{\text{estimated standard error}} M_2$$

This produces a second interpretation for the estimated standard error. Specifically, the standard error can be viewed as a measure of how much difference is reasonable to expect between two sample means if the null hypothesis is true.

The second interpretation of the estimated standard error produces a simplified version of the independent-measures *t* statistic.

$$t = \frac{\text{sample mean difference}}{\text{estimated standard error}}$$

$$= \frac{\text{actual difference between } M_1 \text{ and } M_2}{\text{standard difference (If } H_0 \text{ is true) between } M_1 \text{ and } M_2}$$

In this version, the numerator of the *t* statistic measures how much difference *actually exists* between the two sample means, including any difference that is caused by the different treatments. The denominator measures how much difference *should exist* between the two sample means if there is no treatment effect that causes them to be different. A large value for the *t* statistic is evidence for the existence of a treatment effect.

CALCULATING THE ESTIMATED STANDARD ERROR

To develop the formula for $s_{(M_1-M_2)}$, we consider the following three points:

1. Each of the two sample means represents it own population mean, but in each case there is some error.

   M_1 approximates μ_1 with some error.

   M_2 approximates μ_2 with some error.

Thus, there are two sources of error.

2. The amount of error associated with each sample mean is measured by the estimated standard error of M. Using Equation 9.1 (p. 285), the estimated standard error for each sample mean is computed as follows:

$$\text{For } M_1 \quad s_M = \sqrt{\frac{s_1^2}{n_1}} \qquad \text{For } M_2 \quad s_M = \sqrt{\frac{s_2^2}{n_2}}$$

3. For the independent-measures t statistic, we want to know the total amount of error involved in using *two* sample means to approximate *two* population means. To do this, we find the error from each sample separately and then add the two errors together. The resulting formula for standard error is

$$s_{(M_1 - M_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.1}$$

Because the independent-measures t statistic uses two sample means, the formula for the estimated standard error simply combines the error for the first sample mean and the error for the second sample mean (Box 10.1).

POOLED VARIANCE

Although Equation 10.1 accurately presents the concept of standard error for the independent-measures t statistic, this formula is limited to situations in which the

BOX 10.1 THE VARIABILITY OF DIFFERENCE SCORES

It may seem odd that the independent-measures t statistic *adds* together the two sample errors when it *subtracts* to find the difference between the two sample means. The logic behind this apparently unusual procedure is demonstrated here.

We begin with two populations, I and II (Figure 10.3). The scores in population I range from a high of 70 to a low of 50. The scores in population II range from 30 to 20. We use the range as a measure of how spread out (variable) each population is:

For population I, the scores cover a range of 20 points.
For population II, the scores cover a range of 10 points.

If we randomly select one score from population I and one score from population II and compute the difference between these two scores ($X_1 - X_2$), what range of values is possible for these differences? To answer this question, we need to find the biggest possible difference and the smallest possible difference. Look at Figure 10.3; the biggest difference occurs when $X_1 = 70$ and $X_2 = 20$. This is a difference of $X_1 - X_2 = 50$ points. The smallest difference occurs when $X_1 = 50$ and $X_2 = 30$. This is a difference of $X_1 - X_2 = 20$ points. Notice that the differences go from a high of 50 to a low of 20. This is a range of 30 points:

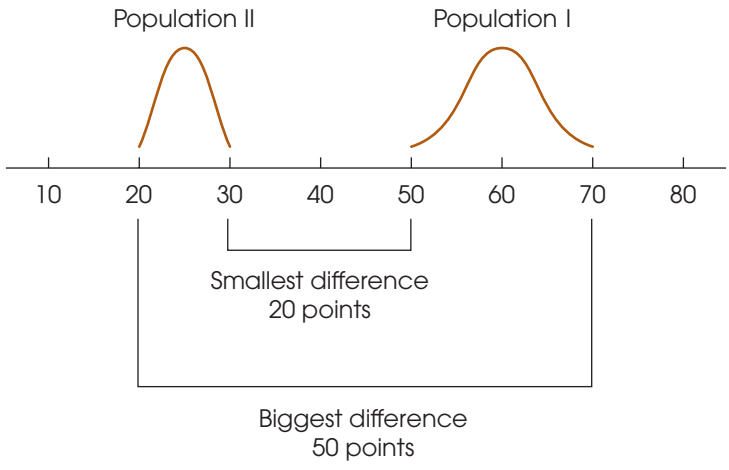
range for population I (X_1 scores) = 20 points
range for population II (X_2 scores) = 10 points
range for the differences ($X_1 - X_2$) = 30 points

The variability for the difference scores is found by *adding* together the variability for each of the two populations.

In the independent-measures t statistics, we compute the variability (standard error) for a sample mean difference. To compute this value, we add together the variability for each of the two sample means.

FIGURE 10.3

Two population distributions. The scores in population I vary from 50 to 70 (a 20-point spread), and the scores in population II range from 20 to 30 (a 10-point spread). If you select one score from each of these two populations, the closest two values are $X_1 = 50$ and $X_2 = 30$. The two values that are farthest apart are $X_1 = 70$ and $X_2 = 20$.



An alternative to computing pooled variance is presented in Box 10.2, p. 339.

two samples are exactly the same size (that is, $n_1 = n_2$). For situations in which the two sample sizes are different, the formula is *biased* and, therefore, inappropriate. The bias comes from the fact that Equation 10.1 treats the two sample variances equally. However, when the sample sizes are different, the two sample variances are not equally good and should not be treated equally. In Chapter 7, we introduced the law of large numbers, which states that statistics obtained from large samples tend to be better (more accurate) estimates of population parameters than statistics obtained from small samples. This same fact holds for sample variances: The variance obtained from a large sample is a more accurate estimate of σ^2 than the variance obtained from a small sample.

One method for correcting the bias in the standard error is to combine the two sample variances into a single value called the *pooled variance*. The pooled variance is obtained by averaging or "pooling" the two sample variances using a procedure that allows the bigger sample to carry more weight in determining the final value.

You should recall that when there is only one sample, the sample variance is computed as

$$s^2 = \frac{SS}{df}$$

For the independent-measures *t* statistic, there are two *SS* values and two *df* values (one from each sample). The values from the two samples are combined to compute what is called the *pooled variance*. The pooled variance is identified by the symbol s_p^2 and is computed as

$$\text{pooled variance} = s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} \tag{10.2}$$

With one sample, the variance is computed as *SS* divided by *df*. With two samples, the pooled variance is computed by combining the two *SS* values and then dividing by the combination of the two *df* values.

As we mentioned earlier, the pooled variance is actually an average of the two sample variances, but the average is computed so that the larger sample carries more weight in determining the final value. The following examples demonstrate this point.

Equal samples sizes We begin with two samples that are exactly the same size. The first sample has $n = 6$ scores with $SS = 50$, and the second sample has $n = 6$ scores with $SS = 30$. Individually, the two sample variances are

$$\text{Variance for sample } 1: s^2 = \frac{SS}{df} = \frac{50}{5} = 10$$

$$\text{Variance for sample } 2: s^2 = \frac{SS}{df} = \frac{30}{5} = 6$$

The pooled variance for these two samples is

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{50 + 30}{5 + 5} = \frac{80}{10} = 8.00$$

Note that the pooled variance is exactly halfway between the two sample variances. Because the two samples are exactly the same size, the pooled variance is simply the average of the two sample variances.

Unequal samples sizes Now consider what happens when the samples are not the same size. This time the first sample has $n = 3$ scores with $SS = 20$, and the second sample has n = 9 scores with SS = 48. Individually, the two sample variances are

$$\text{Variance for sample } 1: s^2 = \frac{SS}{df} = \frac{20}{2} = 10$$

$$\text{Variance for sample } 2: s^2 = \frac{SS}{df} = \frac{48}{8} = 6$$

The pooled variance for these two samples is

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{20 + 48}{2 + 8} = \frac{68}{10} = 6.80$$

This time the pooled variance is not located halfway between the two sample variances. Instead, the pooled value is closer to the variance for the larger sample ($n = 9$ and $s^2 = 6$) than to the variance for the smaller sample ($n = 3$ and $s^2 = 10$). The larger sample carries more weight when the pooled variance is computed.

When computing the pooled variance, the weight for each of the individual sample variances is determined by its degrees of freedom. Because the larger sample has a larger *df* value, it carries more weight when averaging the two variances. This produces an alternative formula for computing pooled variance.

$$\text{pooled variance} = s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} \tag{10.3}$$

For example, if the first sample has $df_1 = 3$ and the second sample has $df_2 = 7$, then the formula instructs you to take 3 of the first sample variance and 7 of the second sample variance for a total of 10 variances. You then divide by 10 to obtain the average. The alternative formula is especially useful if the sample data are summarized as means and variances. Finally, you should note that because the pooled variance is an average of the two sample variances, the value obtained for the pooled variance is always located between the two sample variances.

ESTIMATED STANDARD ERROR

Using the pooled variance in place of the individual sample variances, we can now obtain an unbiased measure of the standard error for a sample mean difference. The resulting formula for the independent-measures estimated standard error is

$$\text{estimated standard error of } M_1 - M_2 = s_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \tag{10.4}$$

Conceptually, this standard error measures how accurately the difference between two sample means represents the difference between the two population means. In a hypothesis test, H_0 specifies that $\mu_1 - \mu_2 = 0$, and the standard error also measures how much difference is expected, on average, between the two sample means. In either case, the formula combines the error for the first sample mean with the error for the second sample mean. Also note that the pooled variance from the two samples is used to compute the standard error for the sample mean difference.

THE FINAL FORMULA AND DEGREES OF FREEDOM

The complete formula for the independent-measures *t* statistic is as follows:

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

$$= \frac{\text{sample mean difference} - \text{population mean difference}}{\text{estimated standard error}} \tag{10.5}$$

In the formula, the estimated standard error in the denominator is calculated using Equation 10.4, and requires calculation of the pooled variance using either Equation 10.2 or 10.3.

The degrees of freedom for the independent-measures *t* statistic are determined by the *df* values for the two separate samples:

$$\begin{aligned} df \text{ for the } t \text{ statistic} &= df \text{ for the first sample} + df \text{ for the second sample} \\ &= df_1 + df_2 \\ &= (n_1 - 1) + (n_2 - 1) \end{aligned} \tag{10.6}$$

Equivalently, the *df* value for the independent-measures *t* statistic can be expressed as

$$df = n_1 + n_2 - 2 \tag{10.7}$$

Note that the *df* formula subtracts 2 points from the total number of scores; 1 point for the first sample and 1 for the second.

The independent-measures *t* statistic is used for hypothesis testing. Specifically, we use the difference between two sample means ($M_1 - M_2$) as the basis for testing hypotheses about the difference between two population means ($\mu_1 - \mu_2$). In this context, the overall structure of the *t* statistic can be reduced to the following:

$$t = \frac{\text{data} - \text{hypothesis}}{\text{error}}$$

This same structure is used for both the single-sample *t* from Chapter 9 and the new independent-measures *t* that was introduced in the preceding pages. Table 10.1 identifies each component of these two *t* statistics and should help reinforce the point that we made earlier in the chapter; that is, the independent-measures *t* statistic simply doubles each aspect of the single-sample *t* statistic.

TABLE 10.1

The basic elements of a *t* statistic for the single-sample *t* and the independent-measures *t*.

	Sample Data	Hypothesized Population Parameter	Estimated Standard Error	Sample Variance
Single-sample *t* statistic	M	μ	$\sqrt{\frac{s^2}{n}}$	$s^2 = \frac{SS}{df}$
Independent-measures *t* statistic	$(M_1 - M_2)$	$(\mu_1 - \mu_2)$	$\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$	$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$

LEARNING CHECK

1. What is the defining characteristic of an independent-measures research study?
2. Explain what is measured by the estimated standard error in the denominator of the independent-measures *t* statistic.
3. One sample from an independent-measures study has $n = 4$ with $SS = 100$. The other sample has $n = 8$ and $SS = 140$.
   a. Compute the pooled variance. (*Note:* Equation 10.2 works well with these data.)
   b. Compute the estimated standard error for the mean difference.
4. One sample from an independent-measures study has $n = 9$ with a variance of $s^2 = 35$. The other sample has $n = 3$ and $s^2 = 40$.
   a. Compute the pooled variance. (*Note:* Equation 10.3 works well with these data.)
   b. Compute the estimated standard error for the mean difference.
5. An independent-measures *t* statistic is used to evaluate the mean difference between two treatments with $n = 8$ in one treatment and $n = 12$ in the other. What is the *df* value for the *t* statistic?

ANSWERS

1. An independent-measures study uses a separate group of participants to represent each of the populations or treatment conditions being compared.
2. The estimated standard error measures how much difference is expected, on average, between a sample mean difference and the population mean difference. In a hypothesis test, $\mu_1 - \mu_2$ is set to zero and the standard error measures how much difference is expected between the two sample means.
3. a. The pooled variance is 240/10 = 24.
   b. The estimated standard error is 3.
4. a. The pooled variance is 36.
   b. The estimated standard error is 4.
5. $df = df_1 + df_2 = 7 + 11 = 18$.

10.3 HYPOTHESIS TESTS AND EFFECT SIZE WITH THE INDEPENDENT-MEASURES *t* STATISTIC

The independent-measures *t* statistic uses the data from two separate samples to help decide whether there is a significant mean difference between two populations or between two treatment conditions. A complete example of a hypothesis test with two independent samples follows.

EXAMPLE 10.1 Research results suggest a relationship between the TV viewing habits of 5-year-old children and their future performance in high school. For example, Anderson, Huston, Wright, and Collins (1998) report that high school students who had regularly watched Sesame Street as children had better grades in high school than their peers who had not watched Sesame Street. Suppose that a researcher intends to examine this phenomenon using a sample of 20 high school students.

The researcher first surveys the students' parents to obtain information on the family's TV-viewing habits during the time that the students were 5 years old. Based on the survey results, the researcher selects a sample of $n = 10$ students with a history of watching Sesame Street and a sample of $n = 10$ students who did not watch the program. The average high school grade is recorded for each student and the data are as follows:

Average High School Grade			
Watched Sesame Street		Did Not Watch Sesame Street	
86	99	90	79
87	97	89	83
91	94	82	86
97	89	83	81
98	92	85	92
$n = 10$		$n = 10$	
$M = 93$		$M = 85$	
$SS = 200$		$SS = 160$	

Note that this is an independent-measures study using two separate samples representing two distinct populations of high school students. The researcher would like to know whether there is a significant difference between the two types of high school student.

STEP 1 State the hypotheses and select the alpha level.

H_0: $\mu_1 - \mu_2 = 0$ (No difference.)

H_1: $\mu_1 - \mu_2 \neq 0$ (There is a difference.)

We set $\alpha = .01$.

Directional hypotheses could be used and would specify whether the students who watched Sesame Street should have higher or lower grades.

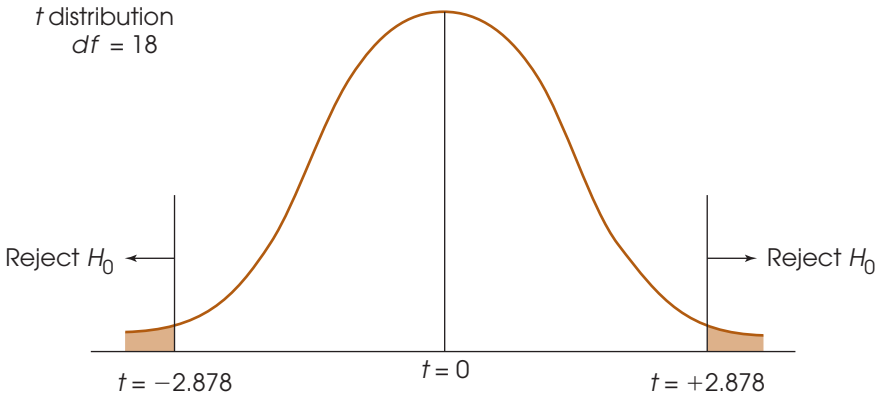
STEP 2 This is an independent-measures design. The *t* statistic for these data has degrees of freedom determined by

$$df = df_1 + df_2$$
$$= (n_1 - 1) + (n_2 - 1)$$
$$= 9 + 9$$
$$= 18$$

The *t* distribution for $df = 18$ is presented in Figure 10.4. For $\alpha = .01$, the critical region consists of the extreme 1% of the distribution and has boundaries of $t = +2.878$ and $t = -2.878$.

FIGURE 10.4

The critical region for the independent-measures hypothesis test in Example 10.1 with $df = 18$ and $\alpha = .01$.



STEP 3 Obtain the data and compute the test statistic. The data are given, so all that remains is to compute the *t* statistic. As with the single-sample *t* test in Chapter 9, we recommend that the calculations be divided into three parts.

First, find the pooled variance for the two samples:

Caution: The pooled variance combines the two samples to obtain a single estimate of variance. In the formula, the two samples are combined in a single fraction.

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

$$= \frac{200 + 160}{9 + 9}$$

$$= \frac{360}{18}$$

$$= 20$$

Second, use the pooled variance to compute the estimated standard error:

Caution: The standard error adds the errors from two separate samples. In the formula, these two errors are added as two separate fractions. In this case, the two errors are equal because the sample sizes are the same.

$$s_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{20}{10} + \frac{20}{10}}$$

$$= \sqrt{2 + 2}$$

$$= \sqrt{4}$$

$$= 2$$

Third, compute the *t* statistic:

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}} = \frac{(93 - 85) - 0}{2}$$

$$= \frac{8}{2}$$

$$= 4$$

STEP 4 Make a decision. The obtained value ($t = 4.00$) is in the critical region. In this example, the obtained sample mean difference is four times greater than would be expected if there were no difference between the two populations. In other words, this result is very unlikely if H_0 is true. Therefore, we reject H_0 and conclude that there is a significant difference between the high school grades for students who watched Sesame Street and those who did not. Specifically, the students who watched Sesame Street had significantly higher grades than those who did not watch the program.

Note that the Sesame Street study in Example 10.1 is an example of nonexperimental research (see Chapter 1, p. 17). Specifically, the researcher did not manipulate the TV programs watched by the children and did not control a variety of variables that could influence high school grades. As a result, we cannot conclude that watching Sesame Street *causes* higher high school grades. In particular, many other, uncontrolled factors, such as the parents' level of education or family economic status, might explain the difference between the two groups. Thus, we do not know exactly why there is a relationship between watching Sesame Street and high school grades, but we do know that a relationship exists.

MEASURING EFFECT SIZE FOR THE INDEPENDENT-MEASURES *t* TEST

As noted in Chapters 8 and 9, a hypothesis test is usually accompanied by a report of effect size to provide an indication of the absolute magnitude of the treatment effect. One technique for measuring effect size is Cohen's *d*, which produces a standardized measure of mean difference. In its general form, Cohen's *d* is defined as

$$d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{\mu_1 - \mu_2}{\sigma}$$

In the context of an independent-measures research study, the difference between the two sample means ($M_1 - M_2$) is used as the best estimate of the mean difference between the two populations, and the pooled standard deviation (the square root of the pooled variance) is used to estimate the population standard deviation. Thus, the formula for estimating Cohen's *d* becomes

$$\text{estimated } d = \frac{\text{estimated mean difference}}{\text{estimated standard deviation}} = \frac{M_1 - M_2}{\sqrt{s_p^2}} \tag{10.8}$$

For the data from Example 10.1, the two sample means are 93 and 85, and the pooled variance is 20. The estimated *d* for these data is

$$d = \frac{M_1 - M_2}{\sqrt{s_p^2}} = \frac{93 - 85}{\sqrt{20}} = \frac{8}{4.7} = 1.79$$

Using the criteria established to evaluate Cohen's *d* (see Table 8.2 on p. 264), this value indicates a very large treatment effect.

The independent-measures *t* test also allows for measuring effect size by computing the percentage of variance accounted for, r^2. As we saw in Chapter 9, r^2 measures how much of the variability in the scores can be explained by the treatment effects. For example, some of the variability in the high school grades from the Sesame Street study can be explained by knowing whether a particular student watched the program; students who watched Sesame Street tend to have higher grades and students who did not watch the show tend to have lower grades. By measuring exactly how much of the variability can be explained, we can obtain a measure of how big the treatment effect

actually is. The calculation of r^2 for the independent-measures t test is exactly the same as it was for the single-sample t test in Chapter 9.

$$r^2 = \frac{t^2}{t^2 + df} \tag{10.9}$$

For the data in Example 10.1, we obtained $t = 4.00$ with $df = 18$. These values produce an r^2 of

$$r^2 = \frac{4^2}{4^2 + 18} = \frac{16}{16 + 18} = \frac{16}{34} = 0.47$$

According to the standards used to evaluate r^2 (see Table 9.3 on p. 299), this value also indicates a very large treatment effect.

Although the value of r^2 is usually obtained by using Equation 10.9, it is possible to determine the percentage of variability directly by computing SS values for the set of scores. The following example demonstrates this process using the data from the Sesame Street study in Example 10.1.

EXAMPLE 10.2

The Sesame Street study described in Example 10.1 compared high school grades for two groups of students; one group who had watched Sesame Street when they were children and one group who had not watched the program. If we assume that the null hypothesis is true and that there is no difference between the two populations of students, then there should be no systematic difference between the two samples. In this case, the two samples can be combined to form a single set of $n = 20$ scores with an overall mean of $M = 89$. The two samples are shown as a single distribution in Figure 10.5(a).

For this example, however, the conclusion from the hypothesis test is that there is a real difference between the two groups. The students who watched Sesame Street have a mean score of $M = 93$, which is 4 points above the overall average. Similarly, the students who did not watch the program had a mean score of $M = 85$, 4 points below the overall average. Thus, the Sesame Street effect causes one group of scores to move toward the right of the distribution, away from the middle, and causes the other group to move toward the left, away from the middle. The result is that the Sesame Street effect causes the scores to spread out and increases the variability.

To determine how much the treatment effect has increased the variability, we remove the treatment effect and examine the resulting scores. To remove the effect, we add 4 points to the score for each student who did not watch Sesame Street and we subtract 4 points from the score for each student who did watch. This adjustment causes both groups to have a mean of $M = 89$, so there is no longer any mean difference between the two groups. The adjusted scores are shown in Figure 10.5(b).

It should be clear that the adjusted scores in Figure 10.5(b) are less variable (more closely clustered) than the original scores in Figure 10.5(a). That is, removing the treatment effect has reduced the variability. To determine exactly how much the treatment influences variability, we have computed SS, the sum of squared deviations, for each set of scores. For the scores in Figure 10.5(a), including the treatment effect, we obtain $SS = 680$. When the treatment effect is removed, in Figure 10.5(b), the variability is reduced to $SS = 360$. The difference between these two values is 320 points. Thus, the treatment effect accounts for 320 points of the total variability in the original scores. When expressed as a proportion of the total variability, we obtain

$$\frac{\text{variability explained by the treatment}}{\text{total variability}} = \frac{320}{680} = 0.47 = 47\%$$

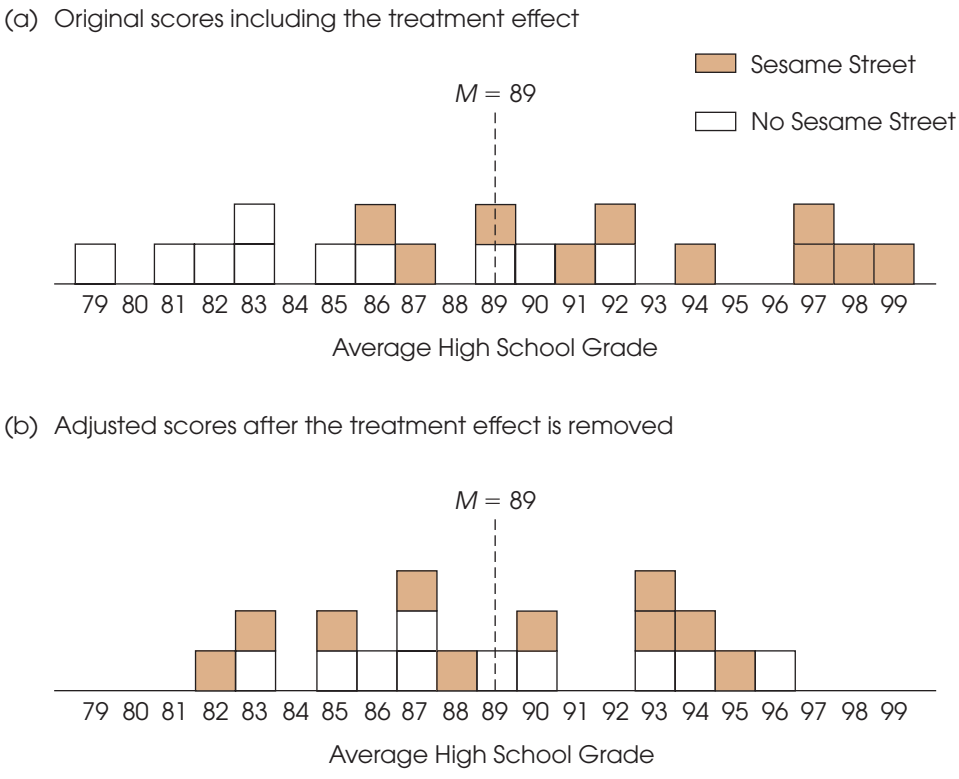


FIGURE 10.5

The two groups of scores from Example 10.1 combined into a single distribution. The original scores, including the treatment effect, are shown in part (a). Part (b) shows the adjusted scores, after the treatment effect has been removed.

You should recognize that this is exactly the same value we obtained for r^2 using Equation 10.9.

CONFIDENCE INTERVALS FOR ESTIMATING $\mu_1 - \mu_2$

As noted in Chapter 9, it is possible to compute a confidence interval as an alternative method for measuring and describing the size of the treatment effect. For the single-sample *t*, we used a single sample mean, *M*, to estimate a single population mean. For the independent-measures *t*, we use a sample mean difference, $M_1 - M_2$, to estimate the population mean difference, $\mu_1 - \mu_2$. In this case, the confidence interval literally estimates the size of the population mean difference between the two populations or treatment conditions.

As with the single-sample *t*, the first step is to solve the *t* equation for the unknown parameter. For the independent-measures *t* statistic, we obtain

$$\mu_1 - \mu_2 = M_1 - M_2 \pm t s_{(M_1 - M_2)} \tag{10.10}$$

In the equation, the values for $M_1 - M_2$ and for $s_{(M_1 - M_2)}$ are obtained from the sample data. Although the value for the *t* statistic is unknown, we can use the degrees of freedom for the *t* statistic and the *t* distribution table to estimate the *t* value. Using the

estimated *t* and the known values from the sample, we can then compute the value of $\mu_1 - \mu_2$. The following example demonstrates the process of constructing a confidence interval for a population mean difference.

EXAMPLE 10.3

Earlier we presented a research study comparing high school grades for students who had watched Sesame Street as children with the grades for students who had not watched the program (p. 326). The results of the hypothesis test indicated a significant mean difference between the two populations of students. Now, we construct a 95% confidence interval to estimate the size of the population mean difference.

The data from the study produced a mean grade of $M = 93$ for the Sesame Street group and a mean of $M = 85$ for the no-Sesame Street group, and the estimated standard error for the mean difference was $s_{(M_1 - M_2)} = 2$. With $n = 10$ scores in each sample, the independent-measures *t* statistic has $df = 18$. To have 95% confidence, we simply estimate that the *t* statistic for the sample mean difference is located somewhere in the middle 95% of all the possible *t* values. According to the *t* distribution table, with $df = 18$, 95% of the *t* values are located between t = +2.101 and $t = -2.101$. Using these values in the estimation equation, we obtain

$$\begin{aligned}\mu_1 - \mu_2 &= M_1 - M_2 \pm ts_{(M_1 - M_2)} \\ &= 93 - 85 \pm 2.101(2) \\ &= 8 \pm 4.202\end{aligned}$$

This produces an interval of values ranging from 8 − 4.202 = 3.798 to 8 + 4.202 = 12.202. Thus, our conclusion is that students who watched Sesame Street have higher grades that those who did not, and the mean difference between the two populations is somewhere between 3.798 points and 12.202 points. Furthermore, we are 95% confident that the true mean difference is in this interval because the only value estimated during the calculations was the *t* statistic, and we are 95% confident that the *t* value is located in the middle 95% of the distribution. Finally note that the confidence interval is constructed around the sample mean difference. As a result, the sample mean difference, $M_1 - M_2 = 93 - 83 = 8$ points, is located exactly in the center of the interval.

As with the confidence interval for the single-sample *t* (p. 302), the confidence interval for an independent-measures *t* is influenced by a variety of factors other than the actual size of the treatment effect. In particular, the width of the interval depends on the percentage of confidence used so that a larger percentage produces a wider interval. Also, the width of the interval depends on the sample size, so that a larger sample produces a narrower interval. Because the interval width is related to sample size, the confidence interval is not a pure measure of effect size like Cohen's *d* or r^2.

CONFIDENCE INTERVALS AND HYPOTHESIS TESTS

In addition to describing the size of a treatment effect, estimation can be used to get an indication of the *significance* of the effect. Example 10.3 presented an independent-measures research study examining the effect on high school grades of having watched Sesame Street as a child. Based on the results of the study, the 95% confidence interval estimated that the population mean difference for the two groups of students was between

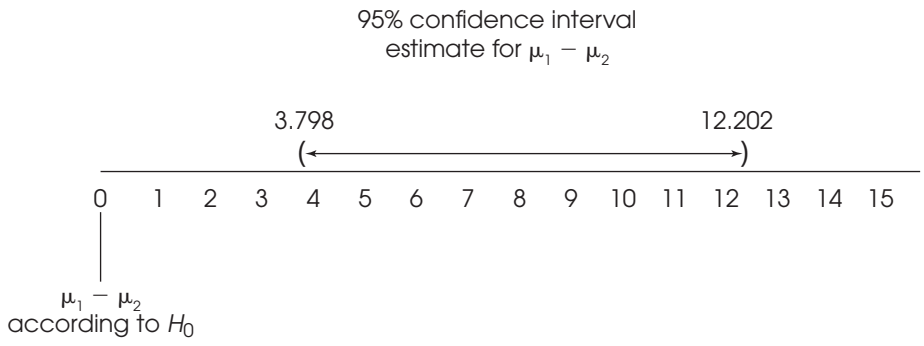


FIGURE 10.6

The 95% confidence interval for the population mean difference ($\mu_1 - \mu_2$) from Example 10.3. Note that $\mu_1 - \mu_2 = 0$ is excluded from the confidence interval, indicating that a zero difference is not an acceptable value (H_0 would be rejected in a hypothesis test with $\alpha = .05$).

The hypothesis test for these data was conducted in Example 10.1 (p. 326) and the decision was to reject H_0.

3.798 and 12.202 points. The confidence interval estimate is shown in Figure 10.6. In addition to the confidence interval for $\mu_1 - \mu_2$, we have marked the spot where the mean difference is equal to zero. You should recognize that a mean difference of zero is exactly what would be predicted by the null hypothesis if we were doing a hypothesis test. You also should realize that a zero difference ($\mu_1 - \mu_2 = 0$) is *outside* of the 95% confidence interval. In other words, $\mu_1 - \mu_2 = 0$ is not an acceptable value if we want 95% confidence in our estimate. To conclude that a value of zero is *not acceptable* with 95% confidence is equivalent to concluding that a value of zero is *rejected* with 95% confidence. This conclusion is equivalent to rejecting H_0 with $\alpha = .05$. On the other hand, if a mean difference of zero were included within the 95% confidence interval, then we would have to conclude that $\mu_1 - \mu_2 = 0$ is an acceptable value, which is the same as failing to reject H_0.

IN THE LITERATURE

REPORTING THE RESULTS OF AN INDEPENDENT-MEASURES *t* TEST

A research report typically presents the descriptive statistics followed by the results of the hypothesis test and measures of effect size (inferential statistics). In Chapter 4 (p. 123), we demonstrated how the mean and the standard deviation are reported in APA format. In Chapter 9 (p. 302), we illustrated the APA style for reporting the results of a *t* test. Now we use the APA format to report the results of Example 10.1, an independent-measures *t* test. A concise statement might read as follows:

> The students who watched Sesame Street as children had higher high school grades ($M = 93$, $SD = 4.71$) than the students who did not watch the program ($M = 85$, $SD = 4.22$). The mean difference was significant, $t\,(18) = 4.00$, $p < .01$, $d = 1.79$.

You should note that standard deviation is not a step in the computations for the independent-measures *t* test, yet it is useful when providing descriptive statistics for each treatment group. It is easily computed when doing the *t* test because you need *SS* and *df* for both groups to determine the pooled variance. Note that the format for reporting *t* is exactly the same as that described in Chapter 9 (p. 302) and that the measure of effect size is reported immediately after the results of the hypothesis test.

Also, as we noted in Chapter 9, if an exact probability is available from a computer analysis, it should be reported. For the data in Example 10.1, the computer analysis reports a probability value of $p = .001$ for $t = 4.00$ with $df = 18$. In the research report, this value would be included as follows:

The difference was significant, $t(18) = 4.00$, $p = .001$, $d = 1.79$.

Finally, if a confidence interval is reported to describe effect size, it appears immediately after the results from the hypothesis test. For the Sesame Street examples (Example 10.1 and Example 10.3), the report would be as follows:

The difference was significant, $t(18) = 4.00$, $p = .001$, 95% CI [3.798, 12.202].

LEARNING CHECK

1. An educational psychologist would like to determine whether access to computers has an effect on grades for high school students. One group of $n = 16$ students has home room each day in a computer classroom in which each student has a computer. A comparison group of $n = 16$ students has home room in a traditional classroom. At the end of the school year, the average grade is recorded for each student. The data are as follows:

Computer	Traditional
$M = 86$	$M = 82.5$
$SS = 1005$	$SS = 1155$

a. Is there a significant difference between the two groups? Use a two-tailed test with $\alpha = .05$.

b. Compute Cohen's d to measure the size of the difference.

c. Write a sentence that demonstrates how the outcome of the hypothesis test and the measure of effect size would appear in a research report.

d. Compute the 90% confidence interval for the population mean difference between a computer classroom and a regular classroom.

2. A researcher report states that there is a significant difference between treatments for an independent-measures design with $t(28) = 2.27$.

a. How many individuals participated in the research study? (*Hint:* Start with the df value.)

b. Should the report state that $p > .05$ or $p < .05$?

ANSWERS

1. a. The pooled variance is 72, the standard error is 3, and $t = 1.17$. With a critical value of $t = 2.042$, fail to reject the null hypothesis.

b. Cohen's $d = 3.5/\sqrt{72} = 0.412$

c. The results show no significant difference in grades for students with computers compared to students without computers, $t(30) = 1.17$, $p > .05$, $d = 0.412$.

d. With $df = 30$ and 90% confidence, the t values for the confidence interval are ± 1.697. The interval is $\mu_1 - \mu_2 = 3.5 \pm 1.697(3)$. Thus, the population mean difference is estimated to be between -1.591 and 8.591. The fact that zero is an acceptable value (inside the interval) is consistent with the decision that there is no significant difference between the two population means.

2. a. The $df = 28$, so the total number of participants is 30.

b. A significant result is indicated by $p < .05$.

DIRECTIONAL HYPOTHESES AND ONE-TAILED TESTS

When planning an independent-measures study, a researcher usually has some expectation or specific prediction for the outcome. For the Sesame Street study in Example 10.1, the researcher clearly expects the students who watched Sesame Street to have higher grades than the students who did not watch. This kind of directional prediction can be incorporated into the statement of the hypotheses, resulting in a directional, or one-tailed, test. Recall from Chapter 8 that one-tailed tests can lead to rejecting H_0 when the mean difference is relatively small compared to the magnitude required by a two-tailed test. As a result, one-tailed tests should be used when clearly justified by theory or previous findings. The following example demonstrates the procedure for stating hypotheses and locating the critical region for a one-tailed test using the independent-measures *t* statistic.

EXAMPLE 10.4

We use the same research situation that was described in Example 10.1. The researcher is using an independent-measures design to examine the relationship between watching educational TV as a child and academic performance as a high school student. The prediction is that high school students who watched Sesame Street regularly as 5-year-old children have higher grades.

STEP 1 State the hypotheses and select the alpha level. As always, the null hypothesis says that there is no effect, and the alternative hypothesis says that there is an effect. For this example, the predicted effect is that the students who watched Sesame Street have higher grades. Thus, the two hypotheses are as follows.

H_0: $\mu_{\text{Sesame Street}} \leq \mu_{\text{No Sesame Street}}$ (Grades are not higher with Sesame Street)

H_1: $\mu_{\text{Sesame Street}} > \mu_{\text{No Sesame Street}}$ (Grades are higher with Sesame Street)

Note that it is usually easier to state the hypotheses in words before you try to write them in symbols. Also, it usually is easier to begin with the alternative hypothesis (H_1), which states that the treatment works as predicted. Also note that the equal sign goes in the null hypothesis, indicating *no difference* between the two treatment conditions. The idea of zero difference is the essence of the null hypothesis, and the numerical value of zero is used for $(\mu_1 - \mu_2)$ during the calculation of the *t* statistic. For this test we use $\alpha = .01$.

STEP 2 Locate the critical region. For a directional test, the critical region is located entirely in one tail of the distribution. Rather than trying to determine which tail, positive or negative, is the correct location, we suggest that you identify the criteria for the critical region in a two-step process as follows. First, look at the data and determine whether the sample mean difference is in the direction that was predicted. If the answer is no, then the data obviously do not support the predicted treatment effect, and you can stop the analysis. On the other hand, if the difference is in the predicted direction, then the second step is to determine whether the difference is large enough to be significant. To test for significance, simply find the one-tailed critical value in the *t* distribution table. If the calculated *t* statistic is more extreme (either positive or negative) than the critical value, then the difference is significant.

For this example, the students who watched Sesame Street had higher grades, as predicted. With $df = 18$, the one-tailed critical value for $\alpha = .01$ is $t = 2.552$.

STEP 3 Collect the data and calculate the test statistic. The details of the calculations were shown in Example 10.1. The data produce a *t* statistic of $t = 4.00$.

STEP 4 Make a decision. The t statistic of $t = 4.00$ is well beyond the critical boundary of $t = 2.552$. Therefore, we reject the null hypothesis and conclude that grades for students who watched Sesame Street are significantly higher than grades for students who did not watch the program. In a research report, the one-tailed test would be clearly noted:

> Grades were significantly higher for students who watched Sesame Street, $t(18) = 4.00$, $p < .01$, one tailed.

THE ROLE OF SAMPLE VARIANCE AND SAMPLE SIZE IN THE INDEPENDENT-MEASURES *t* TEST

In Chapter 9 (p. 294), we identified several factors that can influence the outcome of a hypothesis test. Two factors that play important roles are the variability of the scores and the size of the samples. Both factors influence the magnitude of the estimated standard error in the denominator of the t statistic. However, the standard error is directly related to sample variance (larger variance leads to larger error) but it is inversely related to sample size (larger size leads to smaller error). As a result, larger variance produces a smaller value for the t statistic (closer to zero) and reduces the likelihood of finding a significant result. By contrast, a larger sample produces a larger value for the t statistic (farther from zero) and increases the likelihood of rejecting H_0.

Although variance and sample size both influence the hypothesis test, only variance has a large influence on measures of effect size such as Cohen's d and r^2; larger variance produces smaller measures of effect size. Sample size, on the other hand, has no effect on the value of Cohen's d and only a small influence on r^2.

The following example provides a visual demonstration of how large sample variance can obscure a mean difference between samples and lower the likelihood of rejecting H_0 for an independent-measures study.

EXAMPLE 10.5 We use the data in Figure 10.7 to demonstrate the influence of sample variance. The figure shows the results from a research study comparing two treatments. Notice that the study uses two separate samples, each with $n = 9$, and there is a 5-point mean difference

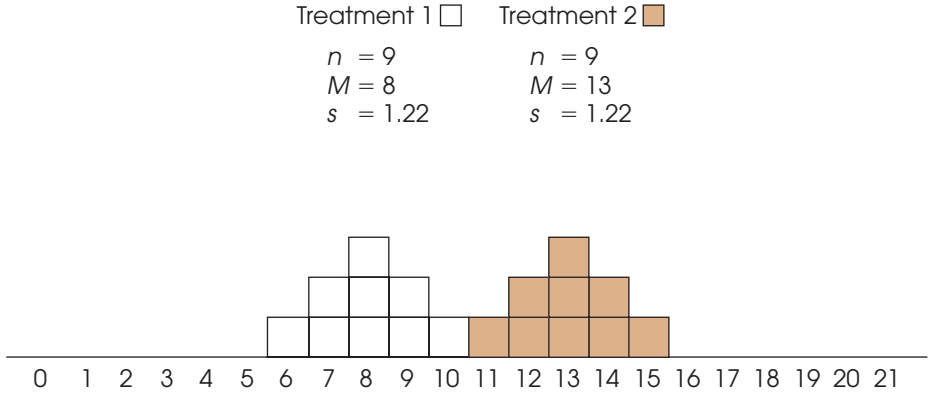


FIGURE 10.7

Two sample distributions representing two different treatments. These data show a significant difference between treatments, $t(16) = 8.62$, $p < .01$, and both measures of effect size indicate a very large treatment effect, $d = 4.10$ and $r^2 = 0.82$.

between the two samples: $M = 8$ for treatment 1 and $M = 13$ for treatment 2. Also notice that there is a clear difference between the two distributions; the scores for treatment 2 are clearly higher than the scores for treatment 1.

For the hypothesis test, the data produce a pooled variance of 1.50 and an estimated standard error of 0.58. The t statistic is

$$t = \frac{\text{mean difference}}{\text{estimated standard error}} = \frac{5}{0.58} = 8.62$$

With $df = 16$, this value is far into the critical region (for $\alpha = .05$ or $\alpha = .01$), so we reject the null hypothesis and conclude that there is a significant difference between the two treatments.

Now consider the effect of increasing sample variance. Figure 10.8 shows the results from a second research study comparing two treatments. Notice that there are still $n = 9$ scores in each sample, and the two sample means are still $M = 8$ and $M = 13$. However, the sample variances have been greatly increased: Each sample now has $s^2 = 44.25$ as compared with $s^2 = 1.5$ for the data in Figure 10.7. Notice that the increased variance means that the scores are now spread out over a wider range, with the result that the two samples are mixed together without any clear distinction between them.

The absence of a clear difference between the two samples is supported by the hypothesis test. The pooled variance is 44.25, the estimated standard error is 3.14, and the independent-measures t statistic is

$$t = \frac{\text{mean difference}}{\text{estimated standard error}} = \frac{5}{3.14} = 1.59$$

With $df = 16$ and $\alpha = .05$, this value is not in the critical region. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference between the two treatments. Although there is still a 5-point difference between sample means (as in Figure 10.7), the 5-point difference is not significant with the increased variance. In general, large sample variance can obscure any mean difference that exists in the data and reduces the likelihood of obtaining a significant difference in a hypothesis test.

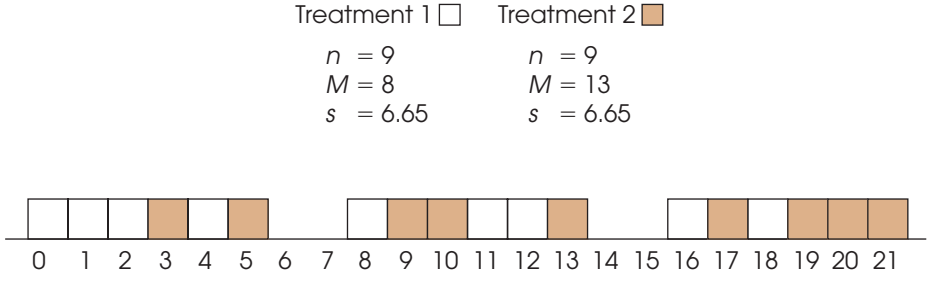


FIGURE 10.8

Two sample distributions representing two different treatments. These data show exactly the same mean difference as the scores in Figure 10.7; however, the variance has been greatly increased. With the increased variance, there is no longer a significant difference between treatments, $t(16) = 1.59$, $p > .05$, and both measures of effect size are substantially reduced, $d = 0.75$ and $r^2 = 0.14$.

Finally, we should note that the problems associated with high variance often can be minimized by transforming the original scores to ranks and then conducting an alternative statistical analysis known as the *Mann-Whitney test,* which is designed specifically for ordinal data. The Mann-Whitney test is presented in Appendix E, which also discusses the general purpose and process of converting numerical scores into ranks. The Mann-Whitney test also can be used if the data violate one of the assumptions for the independent-measures *t* test outlined in the next section.

10.4 ASSUMPTIONS UNDERLYING THE INDEPENDENT-MEASURES *t* FORMULA

There are three assumptions that should be satisfied before you use the independent-measures *t* formula for hypothesis testing:

1. The observations within each sample must be independent (see p. 254).
2. The two populations from which the samples are selected must be normal.
3. The two populations from which the samples are selected must have equal variances.

The first two assumptions should be familiar from the single-sample *t* hypothesis test presented in Chapter 9. As before, the normality assumption is the less important of the two, especially with large samples. When there is reason to suspect that the populations are far from normal, you should compensate by ensuring that the samples are relatively large.

Remember: Adding a constant to (or subtracting a constant from) each score does not change the standard deviation.

The third assumption is referred to as *homogeneity of variance* and states that the two populations being compared must have the same variance. You may recall a similar assumption for the *z*-score hypothesis test in Chapter 8. For that test, we assumed that the effect of the treatment was to add a constant amount to (or subtract a constant amount from) each individual score. As a result, the population standard deviation after treatment was the same as it had been before treatment. We now are making essentially the same assumption, but phrasing it in terms of variances.

Recall that the pooled variance in the *t*-statistic formula is obtained by averaging together the two sample variances. It makes sense to average these two values only if they both are estimating the same population variance—that is, if the homogeneity of variance assumption is satisfied. If the two sample variances are estimating different population variances, then the average is meaningless. (*Note*: If two people are asked to estimate the same thing—for example, what your IQ is—it is reasonable to average the two estimates. However, it is not meaningful to average estimates of two different things. If one person estimates your IQ and another estimates the number of grapes in a pound, it is meaningless to average the two numbers.)

Homogeneity of variance is most important when there is a large discrepancy between the sample sizes. With equal (or nearly equal) sample sizes, this assumption is less critical, but still important. Violating the homogeneity of variance assumption can negate any meaningful interpretation of the data from an independent-measures experiment. Specifically, when you compute the *t* statistic in a hypothesis test, all of the numbers in the formula come from the data except for the population mean difference, which you get from H_0. Thus, you are sure of all of the numbers in the formula except one. If you obtain an extreme result for the *t* statistic (a value in the critical region), then you conclude that the hypothesized value was wrong. But consider what happens when

you violate the homogeneity of variance assumption. In this case, you have two questionable values in the formula (the hypothesized population value and the meaningless average of the two variances). Now if you obtain an extreme *t* statistic, you do not know which of these two values is responsible. Specifically, you cannot reject the hypothesis because it may have been the pooled variance that produced the extreme *t* statistic. Without satisfying the homogeneity of variance requirement, you cannot accurately interpret a *t* statistic, and the hypothesis test becomes meaningless.

HARTLEY'S *F*-MAX TEST

How do you know whether the homogeneity of variance assumption is satisfied? One simple test involves just looking at the two sample variances. Logically, if the two population variances are equal, then the two sample variances should be very similar. When the two sample variances are reasonably close, you can be reasonably confident that the homogeneity assumption has been satisfied and proceed with the test. However, if one sample variance is more than three or four times larger than the other, then there is reason for concern. A more objective procedure involves a statistical test to evaluate the homogeneity assumption. Although there are many different statistical methods for determining whether the homogeneity of variance assumption has been satisfied, Hartley's *F*-max test is one of the simplest to compute and to understand. An additional advantage is that this test can also be used to check homogeneity of variance with more than two independent samples. Later, in Chapter 12, we examine statistical methods for comparing several different samples, and Hartley's test is useful again. The following example demonstrates the *F*-max test for two independent samples.

EXAMPLE 10.6

The *F*-max test is based on the principle that a sample variance provides an unbiased estimate of the population variance. The null hypothesis for this test states that the population variances are equal, therefore, the sample variances should be very similar. The procedure for using the *F*-max test is as follows:

1. Compute the sample variance, $s^2 = \dfrac{SS}{df}$, for each of the separate samples.
2. Select the largest and the smallest of these sample variances and compute

$$F\text{-max} = \frac{s^2(\text{largest})}{s^2(\text{smallest})}$$

   A relatively large value for *F*-max indicates a large difference between the sample variances. In this case, the data suggest that the population variances are different and that the homogeneity assumption has been violated. On the other hand, a small value of *F*-max (near 1.00) indicates that the sample variances are similar and that the homogeneity assumption is reasonable.
3. The *F*-max value computed for the sample data is compared with the critical value found in Table B.3 (Appendix B). If the sample value is larger than the table value, then you conclude that the variances are different and that the homogeneity assumption is not valid.

   To locate the critical value in the table, you need to know:

   a. k = number of separate samples. (For the independent-measures *t* test, $k = 2$.)

   b. $df = n - 1$ for each sample variance. The Hartley test assumes that all samples are the same size.

   c. The alpha level. The table provides critical values for $\alpha = .05$ and $\alpha = .01$. Generally a test for homogeneity would use the larger alpha level.

Suppose, for example, that two independent samples each have $n = 10$ with sample variances of 12.34 and 9.15. For these data,

$$F\text{-max} = \frac{s^2(\text{largest})}{s^2(\text{smallest})} = \frac{12.34}{9.15} = 1.35$$

With $\alpha = .05$, $k = 2$, and $df = n - 1 = 9$, the critical value from the table is 4.03. Because the obtained F-max is smaller than this critical value, you conclude that the data do not provide evidence that the homogeneity of variance assumption has been violated.

The goal for most hypothesis tests is to reject the null hypothesis to demonstrate a significant difference or a significant treatment effect. However, when testing for homogeneity of variance, the preferred outcome is to fail to reject H_0. Failing to reject H_0 with the F-max test means that there is no significant difference between the two population variances and the homogeneity assumption is satisfied. In this case, you may proceed with the independent-measures t test using pooled variance.

If the F-max test rejects the hypothesis of equal variances, or if you simply suspect that the homogeneity of variance assumption is not justified, you should not compute an independent-measures t statistic using pooled variance. However, there is an alternative formula for the t statistic that does not pool the two sample variances and does not require the homogeneity assumption. The alternative formula is presented in Box 10.2.

BOX 10.2 AN ALTERNATIVE TO POOLED VARIANCE

Computing the independent-measures t statistic using pooled variance requires that the data satisfy the homogeneity of variance assumption. Specifically, the two distributions from which the samples are obtained must have equal variances. To avoid this assumption, many statisticians recommend an alternative formula for computing the independent-measures t statistic that does not require pooled variance or the homogeneity assumption. The alternative procedure consists of two steps:

1. The standard error is computed using the two separate sample variances as in Equation 10.1.
2. The value of degrees of freedom for the t statistic is adjusted using the following equation:

$$df = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \quad \text{where } V_1 = \frac{s_1^2}{n_1} \text{ and } V_2 = \frac{s_2^2}{n_2}$$

Decimal values for df should be rounded down to the next lower integer.

The adjustment to degrees of freedom lowers the value of df, which pushes the boundaries for the critical region farther out. Thus, the adjustment makes the test more demanding and therefore corrects for the same bias problem that the pooled variance attempts to avoid.

Note: Many computer programs that perform statistical analysis (such as SPSS) report two versions of the independent-measures t statistic; one using pooled variance (with equal variances assumed) and one using the adjustment shown here (with equal variances not assumed).

LEARNING CHECK

1. A researcher is using an independent-measures design to evaluate the difference between two treatment conditions with $n = 8$ in each treatment. The first treatment produces $M = 63$ with a variance of $s^2 = 18$, and the second treatment has $M = 58$ with $s^2 = 14$.
   - **a.** Use a one-tailed test with $\alpha = .05$ to determine whether the scores in the first treatment are significantly greater than the scores in the second. (*Note*: Because the two samples are the same size, the pooled variance is simply the average of the two sample variances.)
   - **b.** Predict how the value for the *t* statistic would be affected if the two sample variances were increased to $s^2 = 68$ and $s^2 = 60$. Compute the new *t* to confirm your answer.
   - **c.** Predict how the value for the *t* statistic for the original samples would be affected if each sample had $n = 32$ scores (instead of $n = 8$). Compute the new *t* to confirm your answer.
2. The homogeneity of variance assumption requires that the two sample variances be equal. (True or false?)
3. When you are using an *F*-max test to evaluate the homogeneity of variance assumption, you usually do not want to find a significant difference between the variances. (True or false?)

ANSWERS

1. **a.** The pooled variance is 16, the estimated standard error is 2, and $t(14) = 2.50$. With a one-tailed critical value of 1.761, reject the null hypothesis. Scores in the first treatment are significantly higher than scores in the second.
   - **b.** Increasing the variance should lower the value of *t*. The new pooled variance is 64, the estimated standard error is 4, and $t(14) = 1.25$.
   - **c.** Increasing the sample sizes should increase the value of *t*. The pooled variance is still 16, but the new standard error is 1, and $t(62) = 5.00$.
2. False. The assumption is that the two *population* variances are equal.
3. True. If there is a significant difference between the two variances, you cannot do the *t* test with pooled variance.

SUMMARY

1. The independent-measures *t* statistic uses the data from two separate samples to draw inferences about the mean difference between two populations or between two different treatment conditions.
2. The formula for the independent-measures *t* statistic has the same structure as the original *z*-score or the single-sample *t*:

$$t = \frac{\text{sample statistic} - \text{population parameter}}{\text{estimated standard error}}$$

For the independent-measures *t*, the sample statistic is the sample mean difference $(M_1 - M_2)$. The population parameter is the population mean difference, $(\mu_1 - \mu_2)$. The estimated standard error for the sample mean difference is computed by combining the errors for the two sample means. The resulting formula is

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

where the estimated standard error is

$$s_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

The pooled variance in the formula, s_p^2, is the weighted mean of the two sample variances:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

This t statistic has degrees of freedom determined by the sum of the df values for the two samples:

$$df = df_1 + df_2$$
$$= (n_1 - 1) + (n_2 - 1)$$

3. For hypothesis testing, the null hypothesis states that there is no difference between the two population means:

$$H_0: \mu_1 = \mu_2 \qquad \text{or} \qquad \mu_1 - \mu_2 = 0$$

4. When a hypothesis test with an independent-measures t statistic indicates a significant difference, you should also compute a measure of the effect size. One measure of effect size is Cohen's d, which is a standardized measure of the mean difference. For the independent-measures t statistic, Cohen's d is estimated as follows:

$$\text{estimated } d = \frac{M_1 - M_2}{\sqrt{s_p^2}}$$

A second common measure of effect size is the percentage of variance accounted for by the treatment effect. This measure is identified by r^2 and is computed as

$$r^2 = \frac{t^2}{t^2 + df}$$

5. An alternative method for describing the size of the treatment effect is to construct a confidence interval for the population mean difference, $\mu_1 - \mu_2$. The confidence interval uses the independent-measures t equation, solved for the unknown mean difference:

$$\mu_1 - \mu_2 = M_1 - M_2 \pm ts_{(M_1 - M_2)}$$

First, select a level of confidence and then look up the corresponding t values. For example, for 95% confidence, use the range of t values that determine the middle 95% of the distribution. The t values are then used in the equation along with the values for the sample mean difference and the standard error, which are computed from the sample data.

6. Appropriate use and interpretation of the t statistic using pooled variance require that the data satisfy the homogeneity of variance assumption. This assumption stipulates that the two populations have equal variances. An informal test of the assumption can be made by verifying that the two sample variances are approximately equal. Hartley's F-max test provides a statistical technique for determining whether the data satisfy the homogeneity assumption. An alternative technique that avoids pooling variances and eliminates the need for the homogeneity assumption is presented in Box 10.2.

KEY TERMS

independent-measures research design (318)
between-subjects research design (318)
repeated-measures research design (318)
within-subjects research design (318)
independent-measures t statistic (319)
estimated standard error of $M_1 - M_2$ (320)
pooled variance (322)
Mann-Whitney test (337)
homogeneity of variance (337)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 10 on the book companion website. The website also provides access to a workshop entitled *Independent vs. Repeated t-tests,* which compares the t test presented in this chapter with the repeated-measures test presented in Chapter 11.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Independent-Measures *t* Test** presented in this chapter.

Data Entry

1. The scores are entered in what is called *stacked format,* which means that all of the scores from *both samples* are entered in one column of the data editor (probably VAR00001). Enter the scores for sample #2 directly beneath the scores from sample #1 with no gaps or extra spaces.
2. Values are then entered into a second column (VAR00002) to identify the sample or treatment condition corresponding to each of the scores. For example, enter a 1 beside each score from sample #1 and enter a 2 beside each score from sample #2.

Data Analysis

1. Click **Analyze** on the tool bar, select **Compare Means,** and click on **Independent-Samples *t* Test.**
2. Highlight the column label for the set of scores (VAR0001) in the left box and click the arrow to move it into the **Test Variable(s)** box.
3. Highlight the label from the column containing the sample numbers (VAR0002) in the left box and click the arrow to move it into the **Group Variable** box.
4. Click on **Define Groups.**
5. Assuming that you used the numbers 1 and 2 to identify the two sets of scores, enter the values 1 and 2 into the appropriate group boxes.
6. Click **Continue.**
7. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95% but you can select **Options** and change the percentage.
8. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the Sesame Street study in Example 10.1 and the program output is shown in Figure 10.9. The output includes a table of sample statistics with the mean, standard deviation, and standard error of the mean for each group. A second table, which is split into two sections in Figure 10.9, begins with the results of Levene's test for homogeneity of variance. This test should *not* be significant (you do not want the two variances to be different), so you want the reported Sig. value to be greater than .05. Next, the results of the independent-measures *t* test are presented using two different assumptions. The top row shows the outcome assuming equal variances, using the pooled variance to compute *t*. The second row does not assume equal variances and computes the *t* statistic using the alternative method presented in Box 10.2. Each row reports the calculated *t* value, the degrees of freedom, the level of significance (the *p* value for the test), the size of the mean difference and the standard error for the mean difference (the denominator of the *t* statistic). Finally, the output includes a 95% confidence interval for the mean difference.

Group Statistics

	VAR00002	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	1.00	10	93.0000	4.71405	1.49071
	2.00	10	85.0000	4.21637	1.33333

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
VAR00001	Equal variances assumed	.384	.543	4.000	18
	Equal variances not assumed			4.000	17.780

Independent Samples Test

		t-test for Equality of Means				
		Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
VAR00001	Equal variances assumed	.001	8.00000	2.00000	3.79816	12.20184
	Equal variances not assumed	.001	8.00000	2.00000	3.79443	12.20557

FIGURE 10.9

The SPSS output for the independent-measures hypothesis test in Example 10.1.

FOCUS ON PROBLEM SOLVING

1. As you learn more about different statistical methods, one basic problem is deciding which method is appropriate for a particular set of data. Fortunately, it is easy to identify situations in which the independent-measures *t* statistic is used. First, the data always consist of two separate samples (two *n*s, two *M*s, two *SS*s, and so on). Second, this *t* statistic is always used to answer questions about a mean difference: On the average, is one group different (better, faster, smarter) than the other group? If you examine the data and identify the type of question that a researcher is asking, you should be able to decide whether an independent-measures *t* is appropriate.
2. When computing an independent-measures *t* statistic from sample data, we suggest that you routinely divide the formula into separate stages rather than trying to do all of the calculations at once. First, find the pooled variance. Second, compute the standard error. Third, compute the *t* statistic.
3. One of the most common errors for students involves confusing the formulas for pooled variance and standard error. When computing pooled variance, you are "pooling" the two samples together into a single variance. This variance is computed as a *single fraction,* with two *SS* values in the numerator and two *df* values in the denominator. When computing the standard error, you are adding the error from the first sample and the error from the second sample. These two separate errors are added as *two separate fractions* under the square root symbol.

DEMONSTRATION 10.1

THE INDEPENDENT-MEASURES *t* TEST

In a study of jury behavior, two samples of participants were provided details about a trial in which the defendant was obviously guilty. Although group 2 received the same details as group 1, the second group was also told that some evidence had been withheld from the jury by the judge. Later the participants were asked to recommend a jail sentence. The length of term suggested by each participant is presented here. Is there a significant difference between the two groups in their responses?

Group 1	Group 2
4	3
4	7
3	8
2	5
5	4
1	7
1	6
4	8

for Group 1: $M = 3$ and $SS = 16$

for Group 2: $M = 6$ and $SS = 24$

There are two separate samples in this study. Therefore, the analysis uses the independent-measures *t* test.

STEP 1 State the hypothesis, and select an alpha level.

H_0: $\mu_1 - \mu_2 = 0$ (For the population, knowing that evidence has been withheld has no effect on the suggested sentence.)

H_1: $\mu_1 \neq \mu_2 \neq 0$ (For the population, knowing that evidence has been withheld has an effect on the jury's response.)

We set the level of significance to $\alpha = .05$, two tails.

STEP 2 **Identify the critical region.** For the independent-measures *t* statistic, degrees of freedom are determined by

$$df = n_1 + n_2 - 2$$
$$= 8 + 8 - 2$$
$$= 14$$

The *t* distribution table is consulted, for a two-tailed test with $\alpha = .05$ and $df = 14$. The critical *t* values are +2.145 and −2.145.

STEP 3 **Compute the test statistic.** As usual, we recommend that the calculation of the t statistic be separated into three stages.

Pooled variance: For these data, the pooled variance equals

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{16+24}{7+7} = \frac{40}{14} = 2.86$$

Estimated standard error: Now we can calculate the estimated standard error for mean differences.

$$s_{(M_1-M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{2.86}{8} + \frac{2.86}{8}} = \sqrt{0.358+0.358} = \sqrt{0.716} = 0.85$$

The t *statistic:* Finally, the *t* statistic can be computed.

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1-M_2)}} = \frac{(3-6)-0}{0.85} = \frac{-3}{0.85} = -3.53$$

STEP 4 **Make a decision about H_0, and state a conclusion.** The obtained *t* value of −3.53 falls in the critical region of the left tail (critical $t = \pm 2.145$). Therefore, the null hypothesis is rejected. The participants who were informed about the withheld evidence gave significantly longer sentences, $t(14) = -3.53$, $p < .05$, two tails.

DEMONSTRATION 10.2

EFFECT SIZE FOR THE INDEPENDENT-MEASURES *t*

We estimate Cohen's *d* and compute r^2 for the jury decision data in Demonstration 10.1. For these data, the two sample means are $M_1 = 3$ and $M_2 = 6$, and the pooled variance is 2.86. Therefore, our estimate of Cohen's *d* is

$$\text{estimated } d = \frac{M_1 - M_2}{\sqrt{s_p^2}} = \frac{3 - 6}{\sqrt{2.86}} = \frac{3}{1.69} = 1.78$$

With a *t* value of $t = 3.53$ and $df = 14$, the percentage of variance accounted for is

$$r^2 = \frac{t^2}{t^2 + df} = \frac{(3.53)^2}{(3.53)^2 + 14} = \frac{12.46}{26.46} = 0.47 \text{ (or } 47\%)$$

PROBLEMS

1. Describe the basic characteristics of an independent-measures, or a between-subjects, research study.

2. Describe what is measured by the estimated standard error in the bottom of the independent-measures *t* statistic.

3. If other factors are held constant, explain how each of the following influences the value of the independent-measures *t* statistic and the likelihood of rejecting the null hypothesis:
- **a.** Increasing the number of scores in each sample.
- **b.** Increasing the variance for each sample.

4. Describe the homogeneity of variance assumption and explain why it is important for the independent-measures *t* test.

5. One sample has $SS = 48$ and a second sample has $SS = 32$.
- **a.** If $n = 5$ for both samples, find each of the sample variances and compute the pooled variance. Because the samples are the same size, you should find that the pooled variance is exactly halfway between the two sample variances.
- **b.** Now assume that $n = 5$ for the first sample and $n = 9$ for the second. Again, calculate the two sample variances and the pooled variance. You should find that the pooled variance is closer to the variance for the larger sample.

6. One sample has $SS = 70$ and a second sample has $SS = 42$.
- **a.** If $n = 8$ for both samples, find each of the sample variances, and calculate the pooled variance. Because the samples are the same size, you should find that the pooled variance is exactly halfway between the two sample variances.
- **b.** Now assume that $n = 8$ for the first sample and $n = 4$ for the second. Again, calculate the two sample variances and the pooled variance. You should find that the pooled variance is closer to the variance for the larger sample.

7. As noted on page 320, when the two population means are equal, the estimated standard error for the independent-measures *t* test provides a measure of how much difference to expect between two sample means. For each of the following situations, assume that $\mu_1 = \mu_2$ and calculate how much difference should be expected between the two sample means.
- **a.** One sample has $n = 8$ scores with $SS = 45$ and the second sample has $n = 4$ scores with $SS = 15$.
- **b.** One sample has $n = 8$ scores with $SS = 150$ and the second sample has $n = 4$ scores with $SS = 90$.
- **c.** In part b, the samples have larger variability (bigger *SS* values) than in part a, but the sample sizes are unchanged. How does larger variability affect the size of the standard error for the sample mean difference?

8. Two separate samples, each with $n = 12$ individuals, receive two different treatments. After treatment, the first sample has $SS = 1740$ and the second has $SS = 1560$.
- **a.** Find the pooled variance for the two samples.
- **b.** Compute the estimated standard error for the sample mean difference.
- **c.** If the sample mean difference is 8 points, is this enough to reject the null hypothesis and conclude that there is a significant difference for a two-tailed test at the .05 level?

d. If the sample mean difference is 12 points, is this enough to indicate a significant difference for a two-tailed test at the .05 level?

e. Calculate the percentage of variance accounted for (r^2) to measure the effect size for an 8-point mean difference and for a 12-point mean difference.

9. Two separate samples receive two different treatments. The first sample has $n = 9$ with $SS = 710$, and the second has $n = 6$ with $SS = 460$.

a. Compute the pooled variance for the two samples.

b. Calculate the estimated standard error for the sample mean difference.

c. If the sample mean difference is 10 points, is this enough to reject the null hypothesis using a two-tailed test with $\alpha = .05$?

d. If the sample mean difference is 13 points, is this enough to reject the null hypothesis using a two tailed test with $\alpha = .05$?

10. For each of the following, assume that the two samples are selected from populations with equal means and calculate how much difference should be expected, on average, between the two sample means.

a. Each sample has $n = 5$ scores with $s^2 = 38$ for the first sample and $s^2 = 42$ for the second. (*Note:* Because the two samples are the same size, the pooled variance is equal to the average of the two sample variances.)

b. Each sample has $n = 20$ scores with $s^2 = 38$ for the first sample and $s^2 = 42$ for the second.

c. In part b, the two samples are bigger than in part a, but the variances are unchanged. How does sample size affect the size of the standard error for the sample mean difference?

11. For each of the following, calculate the pooled variance and the estimated standard error for the sample mean difference

a. The first sample has $n = 4$ scores and a variance of $s^2 = 55$, and the second sample has $n = 6$ scores and a variance of $s^2 = 63$.

b. Now the sample variances are increased so that the first sample has $n = 4$ scores and a variance of $s^2 = 220$, and the second sample has $n = 6$ scores and a variance of $s^2 = 252$.

c. Comparing your answers for parts a and b, how does increased variance influence the size of the estimated standard error?

12. A researcher conducts an independent-measures study comparing two treatments and reports the t statistic as $t(30) = 2.085$.

a. How many individuals participated in the entire study?

b. Using a two-tailed test with $\alpha = .05$, is there a significant difference between the two treatments?

c. Compute r^2 to measure the percentage of variance accounted for by the treatment effect.

13. Hallam, Price, and Katsarou (2002) investigated the influence of background noise on classroom performance for children aged 10 to 12. In one part of the study, calming music led to better performance on an arithmetic task compared to a no-music condition. Suppose that a researcher selects one class of $n = 18$ students who listen to calming music each day while working on arithmetic problems. A second class of $n = 18$ serves as a control group with no music. Accuracy scores are measured for each child and the average for students in the music condition is $M = 86.4$ with $SS = 1550$ compared to an average of $M = 78.8$ with $SS = 1204$ for students in the no-music condition.

a. Is there a significant difference between the two music conditions? Use a two-tailed test with $\alpha = .05$.

b. Compute the 90% confidence interval for the population mean difference.

c. Write a sentence demonstrating how the results from the hypothesis test and the confidence interval would appear in a research report.

14. Do you view a chocolate bar as delicious or as fattening? Your attitude may depend on your gender. In a study of American college students, Rozin, Bauer, and Catanese (2003) examined the importance of food as a source of pleasure versus concerns about food associated with weight gain and health. The following results are similar to those obtained in the study. The scores are a measure of concern about the negative aspects of eating.

Males	Females
$n = 9$	$n = 15$
$M = 33$	$M = 42$
$SS = 740$	$SS = 1240$

a. Based on these results, is there a significant difference between the attitudes for males and for females? Use a two-tailed test with $\alpha = .05$.

b. Compute r^2, the percentage of variance accounted for by the gender difference, to measure effect size for this study.

c. Write a sentence demonstrating how the result of the hypothesis test and the measure of effect size would appear in a research report.

15. In a study examining overweight and obese college football players, Mathews and Wagner (2008) found that on average both offensive and defensive linemen exceeded the at-risk criterion for body mass index (BMI). BMI is a ratio of body weight to height squared and is commonly used to classify people as overweight or obese. Any value greater than 30 kg/m^2 is considered to be at risk. In the study, a sample of $n = 17$ offensive linemen averaged $M = 34.4$ with a

standard deviation of $s = 4.0$. A sample of $n = 19$ defensive linemen averaged $M = 31.9$ with $s = 3.5$.

a. Use a single-sample *t* test to determine whether the offensive linemen are significantly above the at-risk criterion for BMI. Use a one-tailed test with $\alpha = .01$.

b. Use a single-sample *t* test to determine whether the defensive linemen are significantly above the at-risk criterion for BMI. Use a one-tailed test with $\alpha = .01$.

c. Use an independent-measures *t* test to determine whether there is a significant difference between the offensive linemen and the defensive linemen. Use a two-tailed test with $\alpha = .01$.

16. Functional foods are those containing nutritional supplements in addition to natural nutrients. Examples include orange juice with calcium and eggs with omega-3. Kolodinsky, et al. (2008) examined attitudes toward functional foods for college students. For American students, the results indicated that females had a more positive attitude toward functional foods and were more likely to purchase them compared to males. In a similar study, a researcher asked students to rate their general attitude toward functional foods on a 7-point scale (higher score is more positive). The results are as follows:

Females	Male
$n = 8$	$n = 12$
$M = 4.69$	$M = 4.43$
$SS = 1.60$	$SS = 2.72$

a. Do the data indicate a significant difference in attitude for males and females? Use a two-tailed test with $\alpha = .05$.

b. Compute r^2, the amount of variance accounted for by the gender difference, to measure effect size.

c. Write a sentence demonstrating how the results of the hypothesis test and the measure of effect size would appear in a research report.

17. In 1974, Loftus and Palmer conducted a classic study demonstrating how the language used to ask a question can influence eyewitness memory. In the study, college students watched a film of an automobile accident and then were asked questions about what they saw. One group was asked, "About how fast were the cars going when they smashed into each other?" Another group was asked the same question except the verb was changed to "hit" instead of "smashed into." The "smashed into" group reported significantly higher estimates of speed than the "hit" group. Suppose a researcher repeats this study with a sample of today's college students and obtains the following results.

Estimated Speed	
Smashed into	Hit
$n = 15$	$n = 15$
$M = 40.8$	$M = 34.0$
$SS = 510$	$SS = 414$

a. Do the results indicate a significantly higher estimated speed for the "smashed into" group? Use a one-tailed test with $\alpha = .01$.

b. Compute the estimated value for Cohen's *d* to measure the size of the effect.

c. Write a sentence demonstrating how the results of the hypothesis test and the measure of effect size would appear in a research report.

18. Numerous studies have found that males report higher self-esteem than females, especially for adolescents (Kling, Hyde, Showers, & Buswell, 1999). Typical results show a mean self-esteem score of $M = 39.0$ with $SS = 60.2$ for a sample of $n = 10$ male adolescents and a mean of $M = 35.4$ with $SS = 69.4$ for a sample of $n = 10$ female adolescents.

a. Do the results indicate that self-esteem is significantly higher for males? Use a one-tailed test with $\alpha = .01$.

b. Use the data to make a 95% confidence interval estimate of the mean difference in self-esteem between male and female adolescents.

c. Write a sentence demonstrating how the results from the hypothesis test and the confidence interval would appear in a research report.

19. A researcher is comparing the effectiveness of two sets of instructions for assembling a child's bike. A sample of eight fathers is obtained. Half of the fathers are given one set of instructions and the other half receives the second set. The researcher measures how much time is needed for each father to assemble the bike. The scores are the number of minutes needed by each participant.

Instruction Set I	Instruction Set II
8	14
4	10
8	6
4	10

a. Is there a significant difference in time for the two sets of instructions? Use a two-tailed test at the .05 level of significance.

b. Calculate the estimated Cohen's *d* and r^2 to measure effect size for this study.

20. When people learn a new task, their performance usually improves when they are tested the next day, but only if they get at least 6 hours of sleep (Stickgold, Whidbee, Schirmer, Patel, & Hobson, 2000). The following data demonstrate this phenomenon. The participants learned a visual discrimination task on one day, and then were tested on the task the following day. Half of the participants were allowed to have at least 6 hours of sleep and the other half were kept awake all night. Is there a significant difference between the two conditions? Use a two-tailed test with $\alpha = .05$.

Performance Scores	
6 Hours of Sleep	No Sleep
$n = 14$	$n = 14$
$M = 72$	$M = 65$
$SS = 932$	$SS = 706$

21. Steven Schmidt (1994) conducted a series of experiments examining the effects of humor on memory. In one study, participants were given a mix of humorous and nonhumorous sentences and significantly more humorous sentences were recalled. However, Schmidt argued that the humorous sentences were not necessarily easier to remember, they were simply preferred when participants had a choice between the two types of sentence. To test this argument, he switched to an independent-measures design in which one group got a set of exclusively humorous sentences and another group got a set of exclusively nonhumorous sentences. The following data are similar to the results from the independent-measures study.

Humorous Sentences				Nonhumorous Sentences			
4	5	2	4	6	3	5	3
6	7	6	6	3	4	2	6
2	5	4	3	4	3	4	4
3	3	5	3	5	2	6	4

Do the results indicate a significant difference in the recall of humorous versus nonhumorous sentences? Use a two-tailed test with $\alpha = .05$.

22. Downs and Abwender (2002) evaluated soccer players and swimmers to determine whether the routine blows to the head experienced by soccer players produced long-term neurological deficits. In the study, neurological tests were administered to mature soccer players and swimmers and the results indicated significant differences. In a similar study, a researcher obtained the following data.

Swimmers	Soccer players
10	7
8	4
7	9
9	3
13	7
7	
6	
12	

a. Are the neurological test scores significantly lower for the soccer players than for the swimmers in the control group? Use a one-tailed test with $\alpha = .05$.
b. Compute the value of r^2 (percentage of variance accounted for) for these data.

23. Research has shown that people are more likely to show dishonest and self-interested behaviors in darkness than in a well-lit environment (Zhong, Bohns, & Gino, 2010). In one experiment, participants were given a set of 20 puzzles and were paid \$0.50 for each one solved in a 5-minute period. However, the participants reported their own performance and there was no obvious method for checking their honesty. Thus, the task provided a clear opportunity to cheat and receive undeserved money. One group of participants was tested in a room with dimmed lighting and a second group was tested in a well-lit room. The reported number of solved puzzles was recorded for each individual. The following data represent results similar to those obtained in the study.

Well-Lit Room	Dimly Lit Room
7	9
8	11
10	13
6	10
8	11
5	9
7	15
12	14
5	10

a. Is there a significant difference in reported performance between the two conditions? Use a two-tailed test with $\alpha = .01$.
b. Compute Cohen's d to estimate the size of the treatment effect.

CHAPTER

11

The *t* Test for Two Related Samples

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Introduction to the *t* statistic (Chapter 9)
  - Estimated standard error
  - Degrees of freedom
  - *t* Distribution
  - Hypothesis tests with the *t* statistic
- Independent-measures design (Chapter 10)

Preview

Swearing is a common, almost reflexive, response to pain. Whether you knock your shin into the edge of a coffee table or smash your thumb with a hammer, most of us respond with a streak of obscenities. One question, however, is whether swearing focuses attention on the pain and, thereby, increases its intensity, or serves as a distraction that reduces pain. To address this issue, Stephens, Atkins, and Kingston (2009) conducted an experiment comparing swearing with other responses to pain. In the study, participants were asked to place one hand in icy cold water for as long as they could bear the pain. Half of the participants were told to repeat their favorite swear word over and over for as long as their hands were in the water. The other half repeated a neutral word. The researchers recorded how long each participant was able to tolerate the ice water. After a brief rest, the two groups switched words and repeated the ice water plunge. Thus, all the participants experienced both conditions (swearing and neutral) with half swearing on their first plunge and half on their second. The results clearly showed that swearing significantly increased the average amount of time that participants could tolerate the pain.

The Problem: In the previous chapter, we introduced a statistical procedure for evaluating the mean difference between two sets of data (the independent-measures *t* statistic). However, the independent-measures *t* statistic is intended for research situations involving two separate and independent samples. You should realize that the two sets of scores in the swearing study are not independent samples. In fact, the same group individuals participated in both of the treatment conditions. What is needed is a new statistical analysis for comparing two means that are both obtained from the same group of participants.

The Solution: In this chapter, we introduce the *repeated-measures* t *statistic,* which is used for hypothesis tests evaluating the mean difference between two sets of scores obtained from the same group of individuals. As you will see, however, this new *t* statistic is very similar to the original *t* statistic that was introduced in Chapter 9.

Finally, we should note that researchers often have a choice when they are planning a research study that compares two different treatment conditions. Specifically, a researcher may choose to use two separate groups of participants, one for each of the treatments, or a researcher may choose to use one group and measure each individual in both of the treatment conditions. Later in this chapter, we take a closer look at the differences between these two research designs and discuss the advantages and disadvantages of each.

11.1 INTRODUCTION TO REPEATED-MEASURES DESIGNS

In the previous chapter, we introduced the independent-measures research design as one strategy for comparing two treatment conditions or two populations. The independent-measures design is characterized by the fact that two separate samples are used to obtain the two sets of scores that are to be compared. In this chapter, we examine an alternative strategy known as a *repeated-measures design,* or a *within-subjects design.* With a repeated-measures design, two separate scores are obtained for each individual in the sample. For example, a group of patients could be measured before therapy and then measured again after therapy. Or, response time could be measured in a driving simulation task for a group of individuals who are first tested when they are sober and then tested again after two alcoholic drinks. In each case, the same variable is being measured twice for the same set of individuals; that is, we are literally repeating measurements on the same sample.

DEFINITION

A **repeated-measures design,** or a **within-subject design,** is one in which the dependent variable is measured two or more times for each individual in a single sample. The same group of subjects is used in all of the treatment conditions.

The main advantage of a repeated-measures study is that it uses exactly the same individuals in all treatment conditions. Thus, there is no risk that the participants in one treatment are substantially different from the participants in another. With an independent-measures design, on the other hand, there is always a risk that the results are biased because the individuals in one sample are systematically different (smarter, faster, more extroverted, and so on) than the individuals in the other sample. At the end of this chapter, we present a more detailed comparison of repeated-measures studies and independent-measures studies, considering the advantages and disadvantages of both types of research.

THE MATCHED-SUBJECTS DESIGN

Occasionally, researchers try to approximate the advantages of a repeated-measures design by using a technique known as *matched subjects*. A matched-subjects design involves two separate samples, but each individual in one sample is matched one-to-one with an individual in the other sample. Typically, the individuals are matched on one or more variables that are considered to be especially important for the study. For example, a researcher studying verbal learning might want to be certain that the two samples are matched in terms of IQ and gender. In this case, a male participant with an IQ of 120 in one sample would be matched with another male with an IQ of 120 in the other sample. Although the participants in one sample are not *identical* to the participants in the other sample, the matched-subjects design at least ensures that the two samples are equivalent (or matched) with respect to some specific variables.

DEFINITION

In a **matched-subjects design,** each individual in one sample is matched with an individual in the other sample. The matching is done so that the two individuals are equivalent (or nearly equivalent) with respect to a specific variable that the researcher would like to control.

Of course, it is possible to match participants on more than one variable. For example, a researcher could match pairs of subjects on age, gender, race, and IQ. In this case, for example, a 22-year-old white female with an IQ of 115 who was in one sample would be matched with another 22-year-old white female with an IQ of 115 in the second sample. The more variables that are used, however, the more difficult it is to find matching pairs. The goal of the matching process is to simulate a repeated-measures design as closely as possible. In a repeated-measures design, the matching is perfect because the same individual is used in both conditions. In a matched-subjects design, however, the best you can get is a degree of match that is limited to the variable(s) that are used for the matching process.

In a repeated-measures design or a matched-subjects design comparing two treatment conditions, the data consist of two sets of scores, which are grouped into sets of two, corresponding to the two scores obtained for each individual or each matched pair of subjects (Table 11.1). Because the scores in one set are directly related, one-to-one, with the scores in the second set, the two research designs are statistically equivalent and share the common name *related-samples* designs (or correlated-samples designs). In this chapter, we focus our discussion on repeated-measures designs because they are overwhelmingly the more common example of related-samples designs. However, you should realize that the statistical techniques used for repeated-measures studies also can be applied directly to data from matched-subjects studies. We should also note that a matched-subjects study occasionally is called a *matched samples design*, but the subjects in the samples must be matched one-to-one before you can use the statistical techniques in this chapter.

TABLE 11.1

An example of the data from a repeated-measures or a matched-subjects study using $n = 5$ participants (or matched pairs).

Participant or Matched Pair	First Score	Second Score	
#1	12	15	←The 2 scores for
#2	10	14	one participant or
#3	15	17	one matched pair
#4	17	17	
#5	12	18	

Now we examine the statistical techniques that allow a researcher to use the sample data from a repeated-measures study to draw inferences about the general population.

11.2 THE *t* STATISTIC FOR A REPEATED-MEASURES RESEARCH DESIGN

The *t* statistic for a repeated-measures design is structurally similar to the other *t* statistics we have examined. As we shall see, it is essentially the same as the single-sample *t* statistic covered in Chapter 9. The major distinction of the related-samples *t* is that it is based on *difference scores* rather than raw scores (*X* values). In this section, we examine difference scores and develop the *t* statistic for related samples.

DIFFERENCE SCORES: THE DATA FOR A REPEATED-MEASURES STUDY

Many over-the-counter cold medications include the warning "may cause drowsiness." Table 11.2 shows an example of data from a study that examines this phenomenon. Note that there is one sample of $n = 4$ participants, and that each individual is measured twice. The first score for each person (X_1) is a measurement of reaction time before the medication was administered. The second score (X_2) measures reaction time 1 hour after taking the medication. Because we are interested in how the medication affects reaction time, we have computed the difference between the first score and the second score for each individual. The *difference scores*, or *D* values, are shown in the last column of the table. Notice that the difference scores measure the amount of change

TABLE 11.2

Reaction-time measurements taken before and after taking an over-the-counter cold medication.

Person	Before Medication (X_1)	After Medication (X_2)	Difference *D*
A	215	210	−5
B	221	242	21
C	196	219	23
D	203	228	25
			$\Sigma D = 64$

Note that M_D is the mean for the sample of *D* scores.

$$M_D = \frac{\Sigma D}{n} = \frac{64}{4} = 16$$

in reaction time for each person. Typically, the difference scores are obtained by subtracting the first score (before treatment) from the second score (after treatment) for each person:

$$\text{difference score} = D = X_2 - X_1 \tag{11.1}$$

Note that the sign of each D score tells you the direction of the change. Person A, for example, shows a decrease in reaction time after taking the medication (a negative change), but person B shows an increase (a positive change).

The sample of difference scores (D values) serves as the sample data for the hypothesis test and all calculations are done using the D scores. To compute the t statistic, for example, we use the number of D scores (n) as well as the sample mean (M_D) and the value of SS for the sample of D scores.

THE HYPOTHESES FOR A RELATED-SAMPLES STUDY

The researcher's goal is to use the sample of difference scores to answer questions about the general population. In particular, the researcher would like to know whether there is any difference between the two treatment conditions for the general population. Note that we are interested in a population of *difference scores*. That is, we would like to know what would happen if every individual in the population were measured in two treatment conditions (X_1 and X_2) and a difference score (D) were computed for everyone. Specifically, we are interested in the mean for the population of difference scores. We identify this population mean difference with the symbol μ_D (using the subscript letter D to indicate that we are dealing with D values rather than X scores).

As always, the null hypothesis states that, for the general population, there is no effect, no change, or no difference. For a repeated-measures study, the null hypothesis states that the mean difference for the general population is zero. In symbols,

$$H_0: \mu_D = 0$$

Again, this hypothesis refers to the mean for the entire population of difference scores. Figure 11.1(a) shows an example of a population of difference scores with a

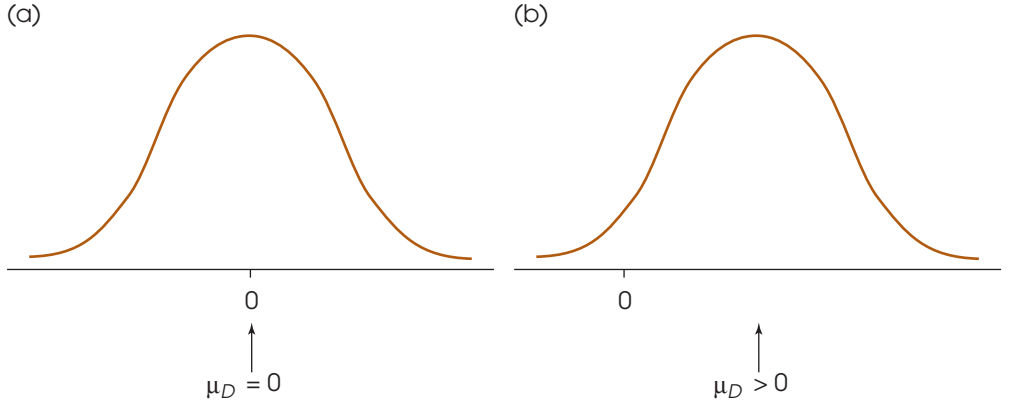


FIGURE 11.1

(a) A population of difference scores for which the mean is $\mu_D = 0$. Note that the typical difference score (D value) is not equal to zero. (b) A population of difference scores for which the mean is greater than zero. Note that most of the difference scores are also greater than zero.

mean of $\mu_D = 0$. Although the population mean is zero, the individual scores in the population are not all equal to zero. Thus, even when the null hypothesis is true, we still expect some individuals to have positive difference scores and some to have negative difference scores. However, the positives and negatives are unsystematic and in the long run balance out to $\mu_D = 0$. Also note that a sample selected from this population probably will not have a mean exactly equal to zero. As always, there will be some error between a sample mean and the population mean, so even if $\mu_D = 0$ (H_0 is true), we do not expect M_D to be exactly equal to zero.

The alternative hypothesis states that there is a treatment effect that causes the scores in one treatment condition to be systematically higher (or lower) than the scores in the other condition. In symbols,

$$H_1: \mu_D \neq 0$$

According to H_1, the difference scores for the individuals in the population tend to be systematically positive (or negative), indicating a consistent, predictable difference between the two treatments.

Figure 11.1(b) shows an example of a population of difference scores with a positive mean difference, $\mu_D > 0$. This time, most of the individuals in the population have difference scores that are greater than zero. A sample selected from this population will contain primarily positive difference scores and will probably have a mean difference that is greater than zero, $M_D > 0$. See Box 11.1 for further discussion of H_0 and H_1.

THE *t* STATISTIC FOR RELATED SAMPLES

Figure 11.2 shows the general situation that exists for a repeated-measures hypothesis test. You may recognize that we are facing essentially the same situation that we encountered in Chapter 9. In particular, we have a population for which the mean and the standard deviation are unknown, and we have a sample that will be used to test a hypothesis about the unknown population. In Chapter 9, we introduced the single-sample *t* statistic, which allowed us to use a sample mean as a basis for testing hypotheses about an unknown

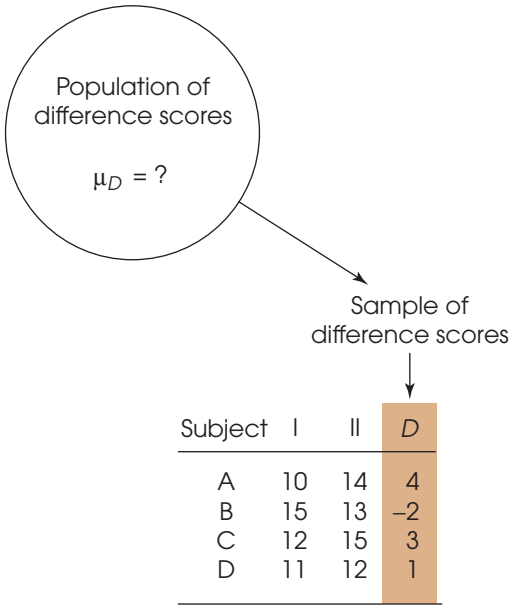
BOX 11.1 ANALOGIES FOR H_0 AND H_1 IN THE REPEATED-MEASURES TEST

An Analogy for H_0: Intelligence is a fairly stable characteristic; that is, you do not get noticeably smarter or dumber from one day to the next. However, if we gave you an IQ test every day for a week, we probably would get seven different numbers. The day-to-day changes in your IQ score are caused by random factors such as your health, your mood, and your luck at guessing answers you do not know. Some days your IQ score is slightly higher, and some days it is slightly lower. On average, the day-to-day changes in IQ should balance out to zero. This is the situation that is predicted by the null hypothesis for a repeated-measures test. According to H_0, any changes that occur either for an individual or for a sample are just due to chance, and in the long run, they will average out to zero.

An Analogy for H_1: On the other hand, suppose that we evaluate your performance on a new video game by measuring your score every day for a week. Again, we probably will find small differences in your scores from one day to the next, just as we did with the IQ scores. However, the day-to-day changes in your game score will not be random. Instead, there should be a general trend toward higher scores as you gain more experience with the new game. Thus, most of the day-to-day changes should show an increase. This is the situation that is predicted by the alternative hypothesis for the repeated-measures test. According to H_1, the changes that occur are systematic and predictable and will not average out to zero.

FIGURE 11.2

A sample of $n = 4$ people is selected from the population. Each individual is measured twice, once in treatment I and once in treatment II, and a difference score, *D,* is computed for each individual. This sample of difference scores is intended to represent the population. Note that we are using a sample of difference scores to represent a population of difference scores. Note that the mean for the population of difference scores is unknown. The null hypothesis states that for the general population there is no consistent or systematic difference between the two treatments, so the population mean difference is $\mu_D = 0$.



Subject	I	II	*D*
A	10	14	4
B	15	13	–2
C	12	15	3
D	11	12	1

population mean. This *t*-statistic formula is used again here to develop the repeated-measures *t* test. To refresh your memory, the single-sample *t* statistic (Chapter 9) is defined by the formula

$$t = \frac{M - \mu}{s_M}$$

In this formula, the sample mean, M, is calculated from the data, and the value for the population mean, μ, is obtained from the null hypothesis. The estimated standard error, s_M, is also calculated from the data and provides a measure of how much difference it is reasonable to expect between a sample mean and the population mean.

For the repeated-measures design, the sample data are difference scores and are identified by the letter *D,* rather than *X*. Therefore, we use *D*s in the formula to emphasize that we are dealing with difference scores instead of *X* values. Also, the population mean that is of interest to us is the population mean difference (the mean amount of change for the entire population), and we identify this parameter with the symbol μ_D. With these simple changes, the *t* formula for the repeated-measures design becomes

As noted earlier, the repeated-measures *t* formula is also used for matched-subjects designs.

$$t = \frac{M_D - \mu_D}{s_{M_D}} \tag{11.2}$$

In this formula, the *estimated standard error for* M_D, s_{M_D}, is computed in exactly the same way as it is computed for the single-sample *t* statistic. To calculate the estimated standard error, the first step is to compute the variance (or the standard deviation) for the sample of *D* scores.

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df} \quad \text{or} \quad s = \sqrt{\frac{SS}{df}}$$

The estimated standard error is then computed using the sample variance (or sample standard deviation) and the sample size, n.

$$s_{M_D} = \sqrt{\frac{s^2}{n}} \quad \text{or} \quad s_{M_D} = \frac{s}{\sqrt{n}} \tag{11.3}$$

Notice that all of the calculations are done using the difference scores (the D scores) and that there is only one D score for each subject. With a sample of n subjects, there are exactly n D scores, and the t statistic has $df = n - 1$. Remember that n refers to the number of D scores, not the number of X scores in the original data.

You should also note that the *repeated-measures* t *statistic* is conceptually similar to the t statistics that we have previously examined:

$$t = \frac{\text{sample statistic} - \text{population parameter}}{\text{estimated standard error}}$$

In this case, the sample data are represented by the sample mean of the difference scores (M_D), the population parameter is the value predicted by H_0 ($\mu_D = 0$), and the estimated standard error is computed from the sample data using Equation 11.3.

LEARNING CHECK

1. For a research study comparing two treatment conditions, what characteristic differentiates a repeated-measures design from an independent-measures design?
2. Describe the data used to compute the sample mean and the sample variance for the repeated-measures t statistic.
3. In words and in symbols, what is the null hypothesis for a repeated-measures t test?

ANSWERS

1. For a repeated-measures design, the same group of individuals is tested in both of the treatments. An independent-measures design uses a separate group for each treatment.
2. The two scores obtained for each individual are used to compute a difference score. The sample of difference scores is used to compute the mean and variance.
3. The null hypothesis states that, for the general population, the average difference between the two conditions is zero. In symbols, $\mu_D = 0$.

11.3 HYPOTHESIS TESTS AND EFFECT SIZE FOR THE REPEATED-MEASURES DESIGN

In a repeated-measures study, each individual is measured in two different treatment conditions and we are interested in whether there is a systematic difference between the scores in the first treatment condition and the scores in the second treatment condition. A difference score (D value) is computed for each person and the hypothesis test uses the difference scores from the sample to evaluate the overall mean difference, μ_D, for the entire population. The hypothesis test with the repeated-measures t statistic

follows the same four-step process that we have used for other tests. The complete hypothesis-testing procedure is demonstrated in Example 11.1.

EXAMPLE 11.1

Research indicates that the color red increases men's attraction to women (Elliot & Niesta, 2008). In the original study, men were shown women's photographs presented on either a white or a red background. Photographs presented on red were rated significantly more attractive than the same photographs mounted on white. In a similar study, a researcher prepares a set of 30 women's photographs, with 15 mounted on a white background and 15 mounted on red. One picture is identified as the test photograph, and appears twice in the set, once on white and once on red. Each male participant looks through the entire set of photographs and rates the attractiveness of each woman on a 12-point scale. Table 11.3 summarizes the ratings of the test photograph for a sample of $n = 9$ men. Are the ratings for the test photograph significantly different when it is presented on a red background compared to a white background?

STEP 1 State the hypotheses, and select the alpha level.

H_0: $\mu_D = 0$ (There is no difference between the two colors.)

H_1: $\mu_D \neq 0$ (There is a change.)

For this test, we use $\alpha = .01$.

STEP 2 Locate the critical region. For this example, $n = 9$, so the t statistic has $df = n - 1 = 8$. For $\alpha = .01$, the critical value listed in the t distribution table is ± 3.355. The critical region is shown in Figure 11.3.

TABLE 11.3

Attractiveness ratings for a woman shown in a photograph presented on a red and a white background.

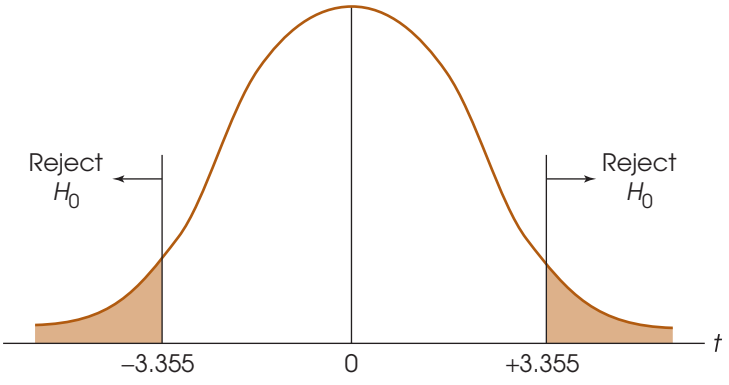
Participant	White Background	Red Background	D	D^2
A	6	9	+3	9
B	8	9	+1	1
C	7	10	+3	9
D	7	11	+4	16
E	8	11	+3	9
F	6	9	+3	9
G	5	11	+6	36
H	10	11	+1	1
I	8	11	+3	9
			$\Sigma D = 27$	$\Sigma D^2 = 99$

$$M_D = \frac{27}{9} = 3.00$$

$$SS = \Sigma D^2 - \frac{(\Sigma D)^2}{n} = 99 - \frac{(27)^2}{9} = 99 - 81 = 18$$

FIGURE 11.3

The critical region for the *t* distribution with $df = 8$ and $\alpha = .01$.



STEP 3 Calculate the *t* statistic. Table 11.3 shows the sample data and the calculations of $M_D = 3.00$ and $SS = 18$. Note that all calculations are done with the difference scores. As we have done with the other *t* statistics, we present the calculation of the *t* statistic as a three-step process.

First, compute the sample variance.

$$s^2 = \frac{SS}{n-1} = \frac{18}{8} = 2.25$$

Next, use the sample variance to compute the estimated standard error.

$$s_{M_D} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{2.25}{9}} = 0.50$$

Finally, use the sample mean (M_D) and the hypothesized population mean (μ_D) along with the estimated standard error to compute the value for the *t* statistic.

$$t = \frac{M_D - \mu_D}{s_{M_D}} = \frac{3.00 - 0}{0.50} = 6.00$$

STEP 4 Make a decision. The *t* value we obtained falls in the critical region (see Figure 11.3). The researcher rejects the null hypothesis and concludes that the background color has a significant effect on the judged attractiveness of the woman in the test photograph.

MEASURING EFFECT SIZE FOR THE REPEATED-MEASURES *t*

As we noted with other hypothesis tests, whenever a treatment effect is found to be statistically significant, it is recommended that you also report a measure of the absolute magnitude of the effect. The most commonly used measures of effect size are Cohen's *d* and r^2, the percentage of variance accounted for. The size of the treatment effect also can be described with a confidence interval estimating the population mean difference, μ_D. Using the data from Example 11.1, we demonstrate how these values are calculated to measure and describe effect size.

Cohen's *d* In Chapters 8 and 9, we introduced Cohen's *d* as a standardized measure of the mean difference between treatments. The standardization simply divides the population mean difference by the standard deviation. For a repeated-measures study, Cohen's *d* is defined as

$$d = \frac{\text{population mean difference}}{\text{standard deviation}} = \frac{\mu_D}{\sigma_D}$$

Because the population mean and standard deviation are unknown, we use the sample values instead. The sample mean, M_D, is the best estimate of the actual mean difference, and the sample standard deviation (square root of sample variance) provides the best estimate of the actual standard deviation. Thus, we are able to estimate the value of *d* as follows:

Because we are measuring the size of the effect and not the direction, it is customary to ignore the minus sign and report Cohen's *d* as a positive value.

$$\text{estimated } d = \frac{\text{sample mean difference}}{\text{sample standard deviation}} = \frac{M_D}{s} \tag{11.4}$$

For the repeated-measures study in Example 11.1, $M_D = 3$ and the sample variance is $s^2 = 2.25$, so the data produce

$$\text{estimated } d = \frac{M_D}{s} = \frac{3.00}{\sqrt{2.25}} = \frac{3.00}{1.5} = 2.00$$

Any value greater than 0.80 is considered to be a large effect, and these data are clearly in that category (see Table 8.2 on p. 264).

The percentage of variance accounted for, r^2 Percentage of variance is computed using the obtained *t* value and the *df* value from the hypothesis test, exactly as was done for the single-sample *t* (see p. 299) and for the independent-measures *t* (see p. 329). For the data in Example 11.1, we obtain

$$r^2 = \frac{t^2}{t^2 + df} = \frac{(6.00)^2}{(6.00)^2 + 8} = \frac{36}{44} = 0.818 \text{ or } 81.8\%$$

For these data, 81.8% of the variance in the scores is explained by the background color for the photograph. More specifically, the color red caused the difference scores to be consistently positive. Thus, the deviations from zero are largely explained by the treatment.

Confidence intervals for estimating μ_D As noted in the previous two chapters, it is possible to compute a confidence interval as an alternative method for measuring and describing the size of the treatment effect. For the repeated-measures *t*, we use a sample mean difference, M_D, to estimate the population mean difference, μ_D. In this case, the confidence interval literally estimates the size of the treatment effect by estimating the population mean difference between the two treatment conditions.

As with the other *t* statistics, the first step is to solve the *t* equation for the unknown parameter. For the repeated-measures *t* statistic, we obtain

$$\mu_D = M_D \pm t s_{M_D} \tag{11.5}$$

In the equation, the values for M_D and for s_{M_D} are obtained from the sample data. Although the value for the *t* statistic is unknown, we can use the degrees of freedom for the *t* statistic and the *t* distribution table to estimate the *t* value. Using the estimated *t* and the known values from the sample, we can then compute the value of μ_D. The following example demonstrates the process of constructing a confidence interval for a population mean difference.

EXAMPLE 11.2

In Example 11.1 we presented a research study demonstrating how men's attractiveness ratings for women are influenced by the color red. In the study, a sample of $n = 9$ men rated a woman shown in a photograph as significantly more attractive when the photo was presented on a red background than when it was on a white background. The mean difference between treatments was $M_D = 3$ points and the estimated standard error for the mean difference was $s_{M_D} = 0.50$. Now, we construct a 95% confidence interval to estimate the size of the population mean difference.

With a sample of $n = 9$ participants, the repeated-measures *t* statistic has $df = 8$. To have 95% confidence, we simply estimate that the *t* statistic for the sample mean difference is located somewhere in the middle 95% of all the possible *t* values. According to the *t* distribution table, with $df = 8$, 95% of the *t* values are located between $t = +2.306$ and $t = -2.306$. Using these values in the estimation equation, together with the values for the sample mean and the standard error, we obtain

$$\mu_D = M_D \pm t s_{M_D}$$

$$= 3 \pm 2.306(0.50)$$

$$= 3 \pm 1.153$$

This produces an interval of values ranging from 3 – 1.153 = 1.847 to 3 + 1.153 = 4.153. Our conclusion is that for general population of men, changing the background color from white to red increases the average attractiveness rating for the woman in the photograph between 1.847 and 4.153 points. We are 95% confident that the true mean difference is in this interval because the only value estimated during the calculations was the *t* statistic, and we are 95% confident that the *t* value is located in the middle 95% of the distribution. Finally note that the confidence interval is constructed around the sample mean difference. As a result, the sample mean difference, $M_D = 3$ points, is located exactly in the center of the interval.

As with the other confidence intervals presented in Chapters 9 and 10, the confidence interval for a repeated-measures *t* is influenced by a variety of factors other than the actual size of the treatment effect. In particular, the width of the interval depends on the percentage of confidence used, so that a larger percentage produces a wider interval. Also, the width of the interval depends on the sample size, so that a larger sample produces a narrower interval. Because the interval width is related to sample size, the confidence interval is not a pure measure of effect size like Cohen's *d* or r^2.

Finally, we should note that the 95% confidence interval computed in Example 11.2 does not include the value $\mu_D = 0$. In other words, we are 95% confident that the population mean difference is not $\mu_D = 0$. This is equivalent to concluding that a null hypothesis specifying that $\mu_D = 0$ would be rejected with a test using $\alpha = .05$. If $\mu_D = 0$ were included in the 95% confidence interval, it would indicate that a hypothesis test would fail to reject H_0 with $\alpha = .05$.

IN THE LITERATURE

REPORTING THE RESULTS OF A REPEATED-MEASURES *t* TEST

As we have seen in Chapters 9 and 10, the APA format for reporting the results of *t* tests consists of a concise statement that incorporates the *t* value, degrees of freedom, and alpha level. One typically includes values for means and standard deviations, either in a statement or a table (Chapter 4). For Example 11.1, we observed a mean difference of $M_D = 3.00$ with $s = 1.50$. Also, we obtained a *t* statistic of $t = 6.00$ with $df = 8$, and our decision was to reject the null hypothesis at the .01 level of significance. Finally, we measured effect size by computing the percentage of variance explained and obtained $r^2 = 0.818$. A published report of this study might summarize the results as follows:

> Changing the background color from white to red increased the attractiveness rating of the woman in the photograph by an average of $M = 3.00$ points with $SD = 1.50$. The treatment effect was statistically significant, $t(8) = 6.00$, $p < .01$, $r^2 = 0.818$.

When the hypothesis test is conducted with a computer program, the printout typically includes an exact probability for the level of significance. The *p*-value from the printout is then stated as the level of significance in the research report. However, the data from Example 11.1 produced a significance level of $p = .000$ in the computer printout. In this case, the probability was so small that the computer rounded it off to 3 decimal points and obtained a value of zero. In this situation you do not know the exact probability value and should report $p < .001$.

If the confidence interval from Example 11.2 is reported as a description of effect size together with the results from the hypothesis test, it would appear as follows:

> Changing the background color from white to red significantly increased the attractiveness rating, $t(8) = 6.00$, $p < .001$, 95% CI [1.817, 4.183].

DESCRIPTIVE STATISTICS AND THE HYPOTHESIS TEST

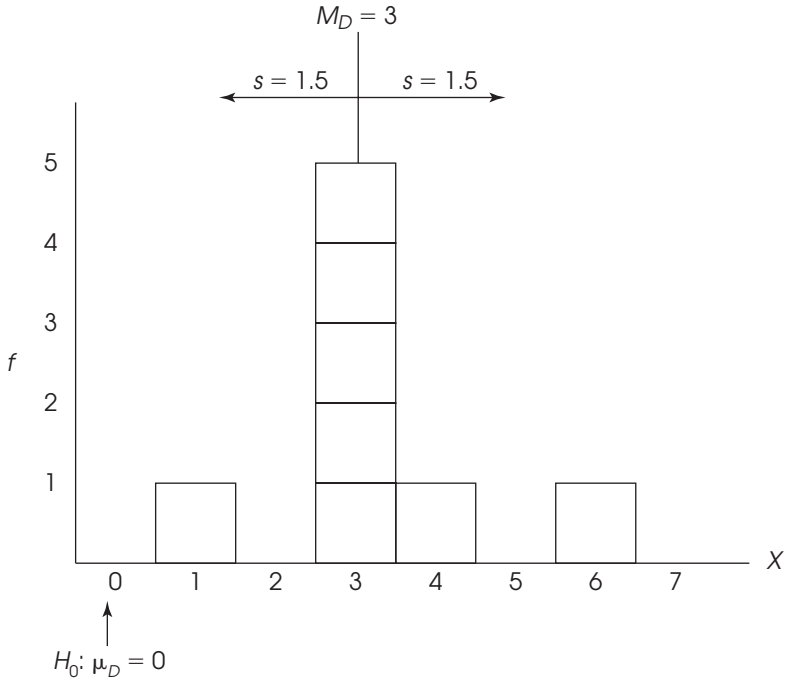
Often, a close look at the sample data from a research study makes it easier to see the size of the treatment effect and to understand the outcome of the hypothesis test. In Example 11.1, we obtained a sample of $n = 9$ men who produce a mean difference of $M_D = 3.00$ with a standard deviation of $s = 1.50$ points. The sample mean and standard deviation describe a set of scores centered at $M_D = 3.00$ with most of the scores located within 1.5 points of the mean. Figure 11.4 shows the actual set of difference scores that were obtained in Example 11.1. In addition to showing the scores in the sample, we have highlighted the position of $\mu_D = 0$; that is, the value specified in the null hypothesis. Notice that the scores in the sample are displaced away from zero. Specifically, the data are not consistent with a population mean of $\mu_D = 0$, which is why we rejected the null hypothesis. In addition, note that the sample mean is located 2 standard deviations above zero. This distance corresponds to the effect size measured by Cohen's $d = 2.00$. For these data, the picture of the sample distribution (see Figure 11.4) should help you to understand the measure of effect size and the outcome of the hypothesis test.

VARIABILITY AS A MEASURE OF CONSISTENCY FOR THE TREATMENT EFFECT

In a repeated-measures study, the variability of the difference scores becomes a relatively concrete and easy-to-understand concept. In particular, the sample variability describes the *consistency* of the treatment effect. For example, if a treatment consistently adds a few points to each individual's score, then the set of difference scores are clustered together with relatively small variability. This is the situation that we

FIGURE 11.4

The sample of difference scores from Example 11.1. The mean is $M_D = 3$ and the standard deviation is $s = 1.5$. The data show a consistent increase in scores (positive differences) and suggest that $\mu_D = 0$ is not a reasonable hypothesis.



observed in Example 11.1 (see Figure 11.4) in which all of the participants produced higher attractiveness ratings for the photograph on a red background. In this situation, with small variability, it is easy to see the treatment effect and it is likely to be significant.

Now consider what happens when the variability is large. Suppose that the red/white study in Example 11.1 produced a sample of $n = 9$ difference scores consisting of –4, –3, –2, +1, +1, +3, +8, +11, and +12. These difference scores also have a mean of $M_D = 3.00$, but now the variability is substantially increased so that $SS = 288$ and the standard deviation is $s = 6.00$. Figure 11.5 shows the new set of difference scores. Again, we have highlighted the position of $\mu_D = 0$, which is the value specified in the null hypothesis. Notice that the high variability means that there is no consistent treatment effect. Some participants rate the photograph as more attractive when it is on a red background (the positive differences) and some rate it higher on a white background (the negative differences). In the hypothesis test, the high variability increases the size of the estimated standard error and results in a hypothesis test that produces $t = 1.50$, which is not in the critical region. With these data, we would fail to reject the null hypothesis and conclude that the color has no effect on the perceived attractiveness of the woman in the photograph.

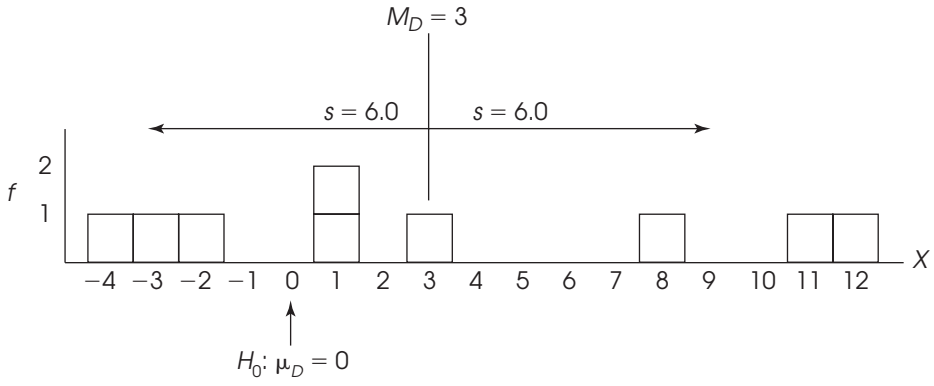
With small variability (see Figure 11.4), the 3-point treatment effect is easy to see and is statistically significant. With large variability (see Figure 11.5), the 3-point effect is not easy to see and is not significant. As we have noted several times in the past, large variability can obscure patterns in the data and reduces the likelihood of finding a significant treatment effect.

DIRECTIONAL HYPOTHESIS AND ONE-TAILED TESTS

In many repeated-measures and matched-subjects studies, the researcher has a specific prediction concerning the direction of the treatment effect. For example, in the study described in Example 11.1, the researcher expects the woman to be judged as more

FIGURE 11.5

A sample of difference scores with a mean difference of $M_D = 3$ and a standard deviation of $s = 6$. The data do not show a consistent increase or decrease in scores. Because there is no consistent treatment effect, $\alpha_D = 0$ is a reasonable hypothesis.



attractive when her photograph is presented on a red background. This kind of directional prediction can be incorporated into the statement of the hypotheses, resulting in a directional, or one-tailed, hypothesis test. The following example demonstrates how the hypotheses and critical region are determined for a directional test.

EXAMPLE 11.3

We reexamine the experiment presented in Example 11.1. The researcher is using a repeated-measures design to investigate the effect of the color red on the perceived attractiveness of a woman. The researcher predicts that the attractiveness ratings for the woman in a photograph will increase when the photograph is presented on a red background compared to a white background.

STEP 1 State the hypotheses and select the alpha level. For this example, the researcher predicts that attractiveness ratings will increase when the photograph is shown on the red background. The null hypothesis, on the other hand says that the attractiveness ratings will not increase but rather will be unchanged or even lowered with the red background. In symbols,

$$H_0: \mu_D \leq 0 \text{ (There is no increase with the color red.)}$$

The alternative hypothesis says that the treatment does work. For this example, H_1 says that the color red will increase the attractiveness ratings.

$$H_1: \mu_D > 0 \text{ (The rating is increased.)}$$

We use $\alpha = .01$.

STEP 2 Locate the critical region. As we demonstrated with the independent-measures t statistic (p. 305), the critical region for a one-tailed test can be located using a two-stage process. Rather than trying to determine which tail of the distribution contains the critical region, you first look at the sample mean difference to verify that it is in the predicted direction. If not, then the treatment clearly did not work as expected and you can stop the test. If the change is in the correct direction, then the question is whether it is large enough to be significant. For this example, change is in the predicted direction (the researcher predicted higher ratings and the sample mean shows an increase.) With $n = 9$, we obtain $df = 8$ and a critical value of $t = 2.896$ for a one-tailed test with $\alpha = .01$. Thus, any t statistic beyond 2.896 (positive or negative) is sufficient to reject the null hypothesis.

STEP 3 Compute the t statistic. We calculated the t statistic in Example 11.1, and obtained $t = 6.00$.

STEP 4 Make a decision. The obtained t statistic is well beyond the critical boundary. Therefore, we reject the null hypothesis and conclude that the color red significantly increased the attractiveness ratings for the woman in the photograph. In a research report, the use of a one-tailed test would be clearly noted as follows:

> Changing the background color from white to red significantly increased the attractiveness rating, $t(8) = 6.00$, $p < .01$, one tailed.

LEARNING CHECK

1. A researcher is investigating the effectiveness of acupuncture treatment for chronic back pain. A sample of $n = 4$ participants is obtained from a pain clinic. Each individual ranks the current level of pain and then begins a 6-week program of acupuncture treatment. At the end of the program, the pain level is rated again and the researcher records the amount of difference between the two ratings. For this sample, pain level decreased by an average of $M = 4.5$ points with $SS = 27$.
   a. Are the data sufficient to conclude that acupuncture has a significant effect on back pain? Use a two-tailed test with $\alpha = .05$.
   b. Can you conclude that acupuncture significantly reduces back pain? Use a one-tailed test with $\alpha = .05$.
2. Compute the effect size using both Cohen's d and r^2 acupuncture study in the previous question.
3. A computer printout for a repeated-measures t test reports a p value of $p = .021$.
   a. Can the researcher claim a significant effect with $\alpha = .01$?
   b. Is the effect significant with $\alpha = .05$?

ANSWERS

1. a. For these data, the sample variance is 9, the standard error is 1.50, and $t = 3.00$. With $df = 3$, the critical values are $t = \pm 3.182$. Fail to reject the null hypothesis.
   b. For a one-tailed test, the critical value is $t = 2.353$. Reject the null hypothesis and conclude that acupuncture treatment significantly reduces pain.
2. $d = 4.5/3 = 1.50$ and $r^2 = 9/12 = 0.75$.
3. a. The exact p value, $p = .021$, is not less than $\alpha = .01$. Therefore, the effect is not significant for $\alpha = .01$ ($p > .01$).
   b. The p value is less than .05, so the effect is significant with $\alpha = .05$.

11.4 USES AND ASSUMPTIONS FOR REPEATED-MEASURES *t* TESTS

REPEATED-MEASURES VERSUS INDEPENDENT-MEASURES DESIGNS

In many research situations, it is possible to use either a repeated-measures design or an independent-measures design to compare two treatment conditions. The independent-measures design would use two separate samples (one in each treatment condition) and the repeated-measures design would use only one sample with the same individuals participating in both treatments. The decision about which design to use is often

made by considering the advantages and disadvantages of the two designs. In general, the repeated-measures design has most of the advantages.

Number of subjects A repeated-measures design typically requires fewer subjects than an independent-measures design. The repeated-measures design uses the subjects more efficiently because each individual is measured in both of the treatment conditions. This can be especially important when there are relatively few subjects available (for example, when you are studying a rare species or individuals in a rare profession).

Study changes over time The repeated-measures design is especially well suited for studying learning, development, or other changes that take place over time. Remember that this design involves measuring individuals at one time and then returning to measure the same individuals at a later time. In this way, a researcher can observe behaviors that change or develop over time.

Individual differences The primary advantage of a repeated-measures design is that it reduces or eliminates problems caused by individual differences. *Individual differences* are characteristics such as age, IQ, gender, and personality that vary from one individual to another. These individual differences can influence the scores obtained in a research study, and they can affect the outcome of a hypothesis test. Consider the data in Table 11.4. The first set of data represents the results from a typical independent-measures study and the second set represents a repeated-measures study. Note that we have identified each participant by name to help demonstrate the effects of individual differences.

For the independent-measures data, note that every score represents a different person. For the repeated-measures study, on the other hand, the same participants are measured in both of the treatment conditions. This difference between the two designs has some important consequences.

1. We have constructed the data so that both research studies have exactly the same scores and they both show the same 5-point mean difference between treatments. In each case, the researcher would like to conclude that the 5-point difference was caused by the treatments. However, with the independent-measures design, there is always the possibility that the participants in treatment 1 have different characteristics than those in treatment 2. For example, the three participants in treatment 1 may be more intelligent than those in treatment 2 and their higher intelligence caused them to have higher scores. Note that this problem disappears with the repeated-measures design. Specifically, with repeated measures there is no possibility that the participants in one treatment are different from those in another treatment because the same participants are used in all of the treatments.

TABLE 11.4

Hypothetical data showing the results from an independent-measures study and a repeated-measures study. The two sets of data use exactly the same numerical scores and they both show the same 5-point mean difference between treatments.

Independent-Measures Study (2 Separate Samples)		Repeated-Measures Study (Same Sample in Both Treatments)		
Treatment 1	Treatment 2	Treatment 1	Treatment 2	*D*
(John) $X = 18$	(Sue) $X = 15$	(John) $X = 18$	(John) $X = 15$	-3
(Mary) $X = 27$	(Tom) $X = 20$	(Mary) $X = 27$	(Mary) $X = 20$	-7
(Bill) $X = 33$	(Dave) $X = 28$	(Bill) $X = 33$	(Bill) $X = 28$	-5
$M = 26$	$M = 21$			$M_D = -5$
$SS = 114$	$SS = 86$			$SS = 8$

2. Although the two sets of data contain exactly the same scores and have exactly the same 5-point mean difference, you should realize that they are very different in terms of the variance used to compute standard error. For the independent-measures study, you calculate the *SS* or variance for the scores in each of the two separate samples. Note that in each sample there are big differences between participants. In treatment 1, for example, Bill has a score of 33 and John's score is only 18. These individual differences produce a relatively large sample variance and a large standard error. For the independent-measures study, the standard error is 5.77, which produces a *t* statistic of $t = 0.87$. For these data, the hypothesis test concludes that there is no significant difference between treatments.

   In the repeated-measures study, the *SS* and variance are computed for the difference scores. If you examine the repeated-measures data in Table 11.4, you will see that the big differences between John and Bill that exist in treatment 1 and in treatment 2 are eliminated when you get to the difference scores. Because the individual differences are eliminated, the variance and standard error are dramatically reduced. For the repeated-measures study, the standard error is 1.15 and the *t* statistic is $t = -4.35$. With the repeated-measures *t,* the data show a significant difference between treatments. Thus, one big advantage of a repeated-measures study is that it reduces variance by removing individual differences, which increases the chances of finding a significant result.

TIME-RELATED FACTORS AND ORDER EFFECTS

The primary disadvantage of a repeated-measures design is that the structure of the design allows for factors other than the treatment effect to cause a participant's score to change from one treatment to the next. Specifically, in a repeated-measures design, each individual is measured in two different treatment conditions, usually *at two different times.* In this situation, outside factors that change over time may be responsible for changes in the participants' scores. For example, a participant's health or mood may change over time and cause a difference in the participant's scores. Outside factors such as the weather can also change and may have an influence on participants' scores. Because a repeated-measures study typically takes place over time, it is possible that time-related factors (other than the two treatments) are responsible for causing changes in the participants' scores.

Also, it is possible that participation in the first treatment influences the individual's score in the second treatment. If the researcher is measuring individual performance, for example, the participants may gain experience during the first treatment condition, and this extra practice may help their performance in the second condition. In this situation, the researcher would find a mean difference between the two conditions; however, the difference would not be caused by the treatments, instead it would caused be by practice effects. Changes in scores that are caused by participation in an earlier treatment are called *order effects* and can distort the mean differences found in repeated-measures research studies.

Counterbalancing One way to deal with time-related factors and order effects is to counterbalance the order of presentation of treatments. That is, the participants are randomly divided into two groups, with one group receiving treatment 1 followed by treatment 2, and the other group receiving treatment 2 followed by treatment 1. The goal of counterbalancing is to distribute any outside effects evenly over the two treatments. For example, if practice effects are a problem, then half of the participants gain experience in treatment 1, which then helps their performance in treatment 2. However, the other half gain experience in treatment 2, which helps their performance in treatment 1. Thus, prior experience helps the two treatments equally.

Finally, if there is reason to expect strong time-related effects or strong order effects, your best strategy is not to use a repeated-measures design. Instead, use independent-measures (or a matched-subjects design) so that each individual participates in only one treatment and is measured only one time.

ASSUMPTIONS OF THE RELATED-SAMPLES *t* TEST

The related-samples *t* statistic requires two basic assumptions:

1. The observations within each treatment condition must be independent (see p. 254). Notice that the assumption of independence refers to the scores *within* each treatment. Inside each treatment, the scores are obtained from different individuals and should be independent of one another.
2. The population distribution of difference scores (*D* values) must be normal.

   As before, the normality assumption is not a cause for concern unless the sample size is relatively small. In the case of severe departures from normality, the validity of the *t* test may be compromised with small samples. However, with relatively large samples ($n > 30$), this assumption can be ignored.

If there is reason to suspect that one of the assumptions for the repeated-measures *t* test has been violated, an alternative analysis known as the *Wilcoxon test* is presented in Appendix E. The Wilcoxon test requires that the original scores be transformed into ranks before evaluating the difference between the two treatment conditions.

LEARNING CHECK

1. What assumptions must be satisfied for repeated-measures *t* tests to be valid?
2. Describe some situations for which a repeated-measures design is well suited.
3. How is a matched-subjects design similar to a repeated-measures design? How do they differ?
4. The data from a research study consist of 10 scores in each of two different treatment conditions. How many individual subjects would be needed to produce these data
   a. For an independent-measures design?
   b. For a repeated-measures design?
   c. For a matched-subjects design?

ANSWERS

1. The observations within a treatment are independent. The population distribution of *D* scores is assumed to be normal.
2. The repeated-measures design is suited to situations in which a particular type of subject is not readily available for study. This design is helpful because it uses fewer subjects (only one sample is needed). Certain questions are addressed more adequately by a repeated-measures design—for example, any time one would like to study changes across time in the same individuals. Also, when individual differences are large, a repeated-measures design is helpful because it reduces the amount of this type of error in the statistical analysis.
3. They are similar in that the role of individual differences in the experiment is reduced. They differ in that there are two samples in a matched-subjects design and only one in a repeated-measures study.
4. a. The independent-measures design would require 20 subjects (two separate samples with $n = 10$ in each).
   b. The repeated-measures design would require 10 subjects (the same 10 individuals are measured in both treatments).
   c. The matched-subjects design would require 20 subjects (10 matched pairs).

SUMMARY

1. In a related-samples research study, the individuals in one treatment condition are directly related, one-to-one, with the individuals in the other treatment condition(s). The most common related-samples study is a repeated-measures design, in which the same sample of individuals is tested in all of the treatment conditions. This design literally repeats measurements on the same subjects. An alternative is a matched-subjects design, in which the individuals in one sample are matched one-to-one with individuals in another sample. The matching is based on a variable relevant to the study.

2. The repeated-measures *t* test begins by computing a difference between the first and second measurements for each subject (or the difference for each matched pair). The difference scores, or *D* scores, are obtained by

$$D = X_2 - X_1$$

   The sample mean, M_D, and sample variance, s^2, are used to summarize and describe the set of difference scores.

3. The formula for the repeated-measures *t* statistic is

$$t = \frac{M_D - \mu_D}{s_{M_D}}$$

   In the formula, the null hypothesis specifies $\mu_D = 0$, and the estimated standard error is computed by

$$s_{M_D} = \sqrt{\frac{s^2}{n}}$$

4. A repeated-measures design may be preferred to an independent-measures study when one wants to observe changes in behavior in the same subjects, as in learning or developmental studies. An important advantage of the repeated-measures design is that it removes or reduces individual differences, which, in turn lowers sample variability and tends to increase the chances for obtaining a significant result.

5. For a repeated-measures design, effect size can be measured using either r^2 (the percentage of variance accounted for) or Cohen's *d* (the standardized mean difference). The value of r^2 is computed the same way for both independent- and repeated-measures designs.

$$r^2 = \frac{t^2}{t^2 + df}$$

   Cohen's *d* is defined as the sample mean difference divided by standard deviation for both repeated- and independent-measures designs. For repeated-measures studies, Cohen's *d* is estimated as

$$\text{estimated } d = \frac{M_D}{s}$$

6. An alternative method for describing the size of the treatment effect is to construct a confidence interval for the population mean difference, μ_D. The confidence interval uses the repeated-measures *t* equation, solved for the unknown mean difference:

$$\mu_D = M_D \pm t s_{M_D}$$

   First, select a level of confidence and then look up the corresponding *t* values. For example, for 95% confidence, use the range of *t* values that determine the middle 95% of the distribution. The *t* values are then used in the equation along with the values for the sample mean difference and the standard error, which are computed from the sample data.

KEY TERMS

repeated-measures design (352)
within-subjects design (352)
matched-subjects design (353)
related-samples design (353)
difference scores (354)
estimated standard error for M_D (357)
repeated-measures *t* statistic (358)
individual differences (367)
order effects (368)
Wilcoxon test (369)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 11 on the book companion website. The website also provides access to a workshop entitled *Independent vs. Repeated t-tests* that compares the *t* test presented in this chapter with the independent-measures test that was presented in Chapter 10.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Repeated-Measures *t* Test** presented in this chapter.

Data Entry

Enter the data into two columns (VAR0001 and VAR0002) in the data editor with the first score for each participant in the first column and the second score in the second column. The two scores for each participant must be in the same row.

Data Analysis

1. Click **Analyze** on the tool bar, select **Compare Means**, and click on **Paired-Samples T Test.**
2. One at a time, highlight the column labels for the two data columns and click the arrow to move them into the **Paired Variables** box.

3. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95%, but you can select **Options** and change the percentage.
4. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the red/white photograph experiment in Example 11.1 and the program output is shown in Figure 11.6. The output includes a table of sample statistics with the mean and standard deviation for each treatment. A second table shows the correlation between the two sets of scores (correlations are presented in Chapter 15). The final table, which is split into two sections in Figure 11.6, shows the results of the hypothesis test, including the mean and standard deviation for the difference scores, the standard error for the mean, a 95% confidence interval for the mean difference, and the values for *t*, *df*, and the level of significance (the *p* value for the test).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	VAR00001	7.2222	9	1.48137	.49379
	VAR00002	10.2222	9	.97183	.32394

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	VAR00001 & VAR00002	9	.309	.419

Paired Samples Test

		Paired Differences		
		Mean	Std. Deviation	Std. Error Mean
Pair 1	VAR00001 - VAR00002	−3.00000	1.50000	.50000

Paired Samples Test

		Paired Differences				
		95% Confidence Interval of the Difference				
		Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	VAR00001 - VAR00002	−4.15300	−1.84700	−6.000	8	.000

FIGURE 11.6

The SPSS output for the repeated-measures hypothesis test in Example 11.1.

FOCUS ON PROBLEM SOLVING

1. Once data have been collected, we must then select the appropriate statistical analysis. How can you tell whether the data call for a repeated-measures *t* test? Look at the experiment carefully. Is there only one sample of subjects? Are the same subjects tested a second time? If your answers are yes to both of these questions, then a repeated-measures *t* test should be done. There is only one situation in which the repeated-measures *t* can be used for data from two samples, and that is for *matched-subjects* studies (p. 353).
2. The repeated-measures *t* test is based on difference scores. In finding difference scores, be sure that you are consistent with your method. That is, you may use either $X_2 - X_1$ or $X_1 - X_2$ to find D scores, but you must use the same method for all subjects.

DEMONSTRATION 11.1

A REPEATED-MEASURES *t* TEST

A major oil company would like to improve its tarnished image following a large oil spill. Its marketing department develops a short television commercial and tests it on a sample of $n = 7$ participants. People's attitudes about the company are measured with a short questionnaire, both before and after viewing the commercial. The data are as follows:

Person	X_1 (Before)	X_2 (After)	D (Difference)
A	15	15	0
B	11	13	+2
C	10	18	+8
D	11	12	+1
E	14	16	+2
F	10	10	0
G	11	19	+8

$\Sigma D = 21$

$M_D = \frac{21}{7} = 3.00$

$SS = 74$

Was there a significant change? Note that participants are being tested twice—once before and once after viewing the commercial. Therefore, we have a repeated-measures design.

STEP 1 **State the hypotheses, and select an alpha level.** The null hypothesis states that the commercial has no effect on people's attitude, or, in symbols,

$$H_0: \mu_D = 0 \text{ (The mean difference is zero.)}$$

The alternative hypothesis states that the commercial does alter attitudes about the company, or

$$H_1: \mu_D \neq 0 \text{ (There is a mean change in attitudes.)}$$

For this demonstration, we use an alpha level of .05 for a two-tailed test.

STEP 2 **Locate the critical region.** Degrees of freedom for the repeated-measures *t* test are obtained by the formula

$$df = n - 1$$

For these data, degrees of freedom equal

$$df = 7 - 1 = 6$$

The *t* distribution table is consulted for a two-tailed test with $\alpha = .05$ for $df = 6$. The critical *t* values for the critical region are $t = \pm 2.447$.

STEP 3 **Compute the test statistic.** Once again, we suggest that the calculation of the *t* statistic be divided into a three-part process.

Variance for the D scores: The variance for the sample of *D* scores is

$$s^2 = \frac{SS}{n-1} = \frac{74}{6} = 12.33$$

Estimated standard error for M_D: The estimated standard error for the sample mean difference is computed as follows:

$$s_{M_D} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{12.33}{7}} = \sqrt{1.76} = 1.33$$

The repeated-measures t statistic: Now we have the information required to calculate the *t* statistic.

$$t = \frac{M_D - \mu_D}{s_{M_D}} = \frac{3-0}{1.33} = 2.26$$

STEP 4 **Make a decision about H_0, and state the conclusion.** The obtained *t* value is not extreme enough to fall in the critical region. Therefore, we fail to reject the null hypothesis. We conclude that there is not enough evidence to conclude that the commercial changes people's attitudes, $t(6) = 2.26$, $p > .05$, two-tailed. (Note that we state that *p* is *greater than* .05 because we failed to reject H_0.)

DEMONSTRATION 11.2

EFFECT SIZE FOR THE REPEATED-MEASURES *t*

We estimate Cohen's *d* and calculate r^2 for the data in Demonstration 11.1. The data produced a sample mean difference of $M_D = 3.00$ with a sample variance of $s^2 = 12.33$. Based on these values, Cohen's *d* is

$$\text{estimated } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{M_D}{s} = \frac{3.00}{\sqrt{12.33}} = \frac{3.00}{3.51} = 0.86$$

The hypothesis test produced $t = 2.26$ with $df = 6$. Based on these values,

$$r^2 = \frac{t^2}{t^2 + df} = \frac{(2.26)^2}{(2.26)^2 + 6} = \frac{5.11}{11.11} = 0.46 \quad (\text{or } 46\%)$$

PROBLEMS

1. For the following studies, indicate whether a repeated-measures *t* test is the appropriate analysis. Explain your answers.

a. A researcher is comparing the amount of time spent playing video games each week for college males versus college females.

b. A researcher is comparing two new designs for cell phones by having a group of high school students send a scripted text message on each model and measuring the difference in speed for each student.

c. A researcher is evaluating the effects of fatigue by testing people in the morning when they are well rested and testing again at midnight when they have been awake for at least 14 hours.

2. Participants enter a research study with unique characteristics that produce different scores from one person to another. For an independent-measures study, these individual differences can cause problems. Briefly explain how these problems are eliminated or reduced with a repeated-measures study.

3. Explain the difference between a matched-subjects design and a repeated-measures design.

4. A researcher conducts an experiment comparing two treatment conditions and obtains data with 10 scores for each treatment condition.

a. If the researcher used an independent-measures design, how many subjects participated in the experiment?

b. If the researcher used a repeated-measures design, how many subjects participated in the experiment?

c. If the researcher used a matched-subjects design, how many subjects participated in the experiment?

5. A sample of $n = 9$ individuals participates in a repeated-measures study that produces a sample mean difference of $M_D = 6.5$ with $SS = 200$ for the difference scores.

a. Calculate the standard deviation for the sample of difference scores. Briefly explain what is measured by the standard deviation.

b. Calculate the estimated standard error for the sample mean difference. Briefly explain what is measured by the estimated standard error.

6. a. A repeated-measures study with a sample of $n = 25$ participants produces a mean difference of $M_D = 3$ with a standard deviation of $s = 4$. Based on the mean and standard deviation, you should be able to visualize (or sketch) the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

b. Now assume that the sample standard deviation is $s = 12$, and once again visualize the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

Explain how the size of the sample standard deviation influences the likelihood of finding a significant mean difference.

7. a. A repeated-measures study with a sample of $n = 9$ participants produces a mean difference of $M_D = 3$ with a standard deviation of $s = 6$. Based on the mean and standard deviation, you should be able to visualize (or sketch) the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

b. Now assume that the sample mean difference is $M_D = 12$, and once again visualize the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

c. Explain how the size of the sample mean difference influences the likelihood of finding a significant mean difference.

8. A sample of difference scores from a repeated-measures experiment has a mean of $M_D = 4$ with a standard deviation of $s = 6$.

a. If $n = 4$, is this sample sufficient to reject the null hypothesis using a two-tailed test with $\alpha = .05$?

b. Would you reject H_0 if $n = 16$? Again, assume a two-tailed test with $\alpha = .05$.

c. Explain how the size of the sample influences the likelihood of finding a significant mean difference.

9. As mentioned in Chapters 2 and 3 (pp. 38 and 81), Steven Schmidt (1994) reported a series of studies examining the effect of humor on memory. In one part of the study, participants were presented with a list containing a mix of humorous and nonhumorous sentences, and were then asked to recall as many sentences as possible. Schmidt recorded the number of humorous and the number of nonhumorous sentences recalled by each individual. Notice that the data consist of two memory scores for each participant. Suppose that a difference score is computed for each individual in a sample of $n = 16$ and the resulting data show that participants recalled an average of $M_D = 3.25$ more humorous sentences than nonhumorous, with $SS = 135$. Are these results sufficient to conclude that humor has a significant effect on memory? Use a two-tailed test with $\alpha = .05$.

10. Research has shown that losing even one night's sleep can have a significant effect on performance of complex tasks such as problem solving (Linde & Bergstroem, 1992). To demonstrate this phenomenon, a sample of $n = 25$ college students was given a problem-solving task at noon on one day and again at noon on the following day. The students were not permitted any sleep between the two tests. For each student, the difference between the first and second score was recorded. For this sample, the students averaged $M_D = 4.7$ points better on the first test with a variance of $s^2 = 64$ for the difference scores.

a. Do the data indicate a significant change in problem-solving ability? Use a two-tailed test with $\alpha = .05$.

b. Compute an estimated Cohen's d to measure the size of the effect.

11. Strack, Martin, and Stepper (1988) reported that people rate cartoons as funnier when holding a pen in their teeth (which forced them to smile) than when holding a pen in their lips (which forced them to frown). A researcher attempted to replicate this result using a sample of $n = 25$ adults between the ages of 40 and 45. For each person, the researcher recorded the difference between the rating obtained while smiling and the rating obtained while frowning. On average the cartoons were rated as funnier when the participants were smiling, with an average difference of $M_D = 1.6$ with $SS = 150$.

a. Do the data indicate that the cartoons are rated significantly funnier when the participants are smiling? Use a one-tailed test with $\alpha = .01$.

b. Compute r^2 to measure the size of the treatment effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

12. How would you react to doing much worse on an exam than you expected? There is some evidence to suggest that most individuals believe that they can cope with this kind of problem better than their fellow students (Igou, 2008). In the study, participants read a scenario of a negative event and were asked to use a 10-point scale to rate how it would affect their immediate well-being (-5 strongly worsen to $+5$ strongly improve). Then they were asked to imagine the event from the perspective of an ordinary fellow student and rate how it would affect that person. The difference between the two ratings was recorded. Suppose that a sample of $n = 25$ participants produced a mean difference of $M_D = 1.28$ points (self rated higher) with a standard deviation of $s = 1.50$ for the difference scores.

a. Is this result sufficient to conclude that there is a significant difference in the ratings for self versus others? Use a two-tailed test with $\alpha = .05$.

b. Compute r^2 and estimate Cohen's d to measure the size of the treatment effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

13. Research results indicate that physically attractive people are also perceived as being more intelligent (Eagly, Ashmore, Makhijani, & Longo, 1991). As a demonstration of this phenomenon, a researcher obtained a set of 10 photographs, 5 showing men who were judged to be attractive and 5 showing men who were judged to be unattractive. The photographs were shown to a sample of $n = 25$ college students and the students were asked to rate the intelligence of the person in the photo on a scale from 1 to 10. For each student, the researcher determined the average rating for the 5 attractive photos and the average for the 5 unattractive photos, and then computed the difference between the two scores. For the entire sample, the average difference was $M_D = 2.7$ (attractive photos rated higher) with $s = 2.00$. Are the data sufficient to conclude that there was a significant difference in perceived intelligence for the two sets of photos? Use a two-tailed test at the .05 level of significance.

14. Researchers have noted a decline in cognitive functioning as people age (Bartus, 1990). However, the results from other research suggest that the antioxidants in foods such as blueberries may reduce and even reverse these age-related declines (Joseph et al., 1999). To examine this phenomenon, suppose that a researcher obtains a sample of $n = 16$ adults who are between the ages of 65 and 75. The researcher uses a standardized test to measure cognitive performance for each individual. The participants then begin a 2-month program in which they receive daily doses of a blueberry supplement. At the end of the 2-month period, the researcher again measures cognitive performance for each participant. The results show an average increase in performance of $M_D = 7.4$ with $SS = 1215$.

a. Does this result support the conclusion that the antioxidant supplement has a significant effect on cognitive performance? Use a two-tailed test with $\alpha = .05$.

b. Construct a 95% confidence interval to estimate the average cognitive performance improvement for the population of older adults.

15. The following data are from a repeated-measures study examining the effect of a treatment by measuring a group of $n = 4$ participants before and after they receive the treatment.

a. Calculate the difference scores and M_D.

b. Compute *SS,* sample variance, and estimated standard error.

c. Is there a significant treatment effect? Use $\alpha = .05$, two tails.

Participant	Before Treatment	After Treatment
A	7	10
B	6	13
C	9	12
D	5	8

16. A researcher for a cereal company wanted to demonstrate the health benefits of eating oatmeal. A sample of 9 volunteers was obtained and each participant ate a fixed diet without any oatmeal for 30 days. At the end of the 30-day period, cholesterol was measured for each individual. Then the participants began a second 30-day period in which they repeated exactly the same diet except that they added 2 cups of oatmeal each day. After the second 30-day period, cholesterol levels were measured again and the researcher recorded the difference between the two scores for each participant. For this sample, cholesterol scores averaged $M_D = 16$ points lower with the oatmeal diet with $SS = 538$ for the difference scores.

a. Are the data sufficient to indicate a significant change in cholesterol level? Use a two-tailed test with $\alpha = .01$.

b. Compute r^2, the percentage of variance accounted for by the treatment, to measure the size of the treatment effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

17. A variety of research results suggest that visual images interfere with visual perception. In one study, Segal and Fusella (1970) had participants watch a screen, looking for brief presentations of a small blue arrow. On some trials, the participants were also asked to form a mental image (for example, imagine a volcano). The results for a sample of $n = 6$, show that participants made an average of $M_D = 4.3$ more errors while forming images than while not forming images. The difference scores had $SS = 63$. Do the data indicate a significant difference between the two conditions? Use a two-tailed test with $\alpha = .05$.

18. One of the primary advantages of a repeated-measures design, compared to independent-measures, is that it reduces the overall variability by removing variance caused by individual differences. The following data are from a research study comparing two treatment conditions.

a. Assume that the data are from an independent-measures study using two separate samples, each with $n = 6$ participants. Compute the pooled variance and the estimated standard error for the mean difference.

b. Now assume that the data are from a repeated-measures study using the same sample of $n = 6$ participants in both treatment conditions. Compute the variance for the sample of difference scores and the estimated standard error for the mean difference. (You should find that the repeated-measures design substantially reduces the variance and the standard error.)

Treatment 1	Treatment 2	Difference
10	13	3
12	12	0
8	10	2
6	10	4
5	6	1
7	9	2
$M = 8$ $SS = 34$	$M = 10$ $SS = 30$	$M_D = 2$ $SS = 10$

19. The previous problem demonstrates that removing individual differences can substantially reduce variance and lower the standard error. However, this benefit only occurs if the individual differences are consistent across treatment conditions. In problem 18, for example, the first two participants (top two rows) consistently had the highest scores in both treatment conditions. Similarly, the last two participants consistently had the lowest scores in both treatments. To construct the following data, we started with the scores in problem 18 and scrambled the scores in treatment 1 to eliminate the consistency of the individual differences.

a. Assume that the data are from an independent-measures study using two separate samples, each with $n = 6$ participants. Compute the pooled variance and the estimated standard error for the mean difference.

b. Now assume that the data are from a repeated-measures study using the same sample of $n = 6$ participants in both treatment conditions. Compute the variance for the sample of difference scores and the estimated standard error for the mean difference. (This time you should find that removing the individual differences does not reduce the variance or the standard error.)

Treatment 1	Treatment 2	Difference
6	13	7
7	12	5
8	10	2
10	10	0
5	6	0
12	9	−3
$M = 8$	$M = 10$	$M_D = 2$
$SS = 34$	$SS = 30$	$SS = 64$

20. A researcher uses a matched-subjects design to investigate whether single people who own pets are generally happier than singles without pets. A mood inventory questionnaire is administered to a group of 20- to 29-year-old non–pet owners and a similar age group of pet owners. The pet owners are matched one to one with the non–pet owners for income, number of close friendships, and general health. The data are as follows:

Matched Pair	Non–Pet Owner	Pet Owner
A	12	14
B	8	7
C	10	13
D	9	9
E	7	13
F	10	12

a. Is there a significant difference in the mood scores for non–pet owners versus pet owners? Test with $\alpha = .05$ for two tails.

b. Construct the 95% confidence interval to estimate the size of the mean difference in mood between the population of pet owners and the population of non–pet owners. (You should find that a mean difference of $\mu_D = 0$ is an acceptable value, which is consistent with the conclusion from the hypothesis test.)

21. There is some evidence suggesting that you are likely to improve your test score if you rethink and change answers on a multiple-choice exam (Johnston, 1975). To examine this phenomenon, a teacher gave the same final exam to two sections of a psychology course. The students in one section were told to turn in their exams immediately after finishing, without changing any of their answers. In the other section, students were encouraged to reconsider each question and to change answers whenever they felt it was appropriate. Before the final exam, the teacher had matched 9 students in the first section with 9 students in the second section based on their midterm grades. For example, a student in the no-change section with an 89 on the midterm exam was matched with student in the change section who also had an 89 on the midterm. The final exam grades for the 9 matched pairs of students are presented in the following table.

a. Do the data indicate a significant difference between the two conditions? Use a two-tailed test with $\alpha = .05$.

b. Construct a 95% confidence interval to estimate the size of the population mean difference.

c. Write a sentence demonstrating how the results of the hypothesis test and the confidence interval would appear in a research report.

Matched Pair	No-Change Section	Change Section
#1	71	86
#2	68	80
#3	91	88
#4	65	74
#5	73	82
#6	81	89
#7	85	85
#8	86	88
#9	65	76

22. The teacher from the previous problem also tried a different approach to answering the question of whether changing answers helps or hurts exam grades. In a separate class, students were encouraged to review their final exams and change any answers they wanted to before turning in their papers. However, the students had to indicate both the original answer and the changed answer for each question. The teacher then graded each exam twice, one using the set of original answers and once with the changes. In the class of $n = 22$ students, the average exam score improved by an average of $M_D = 2.5$ points with the changed answers. The standard deviation for the difference scores was $s = 3.1$. Are the data sufficient to conclude that rethinking and changing answers can significantly improve exam scores? Use a one-tailed test at the .01 level of significance.

23. At the Olympic level of competition, even the smallest factors can make the difference between winning and losing. For example, Pelton (1983) has shown that Olympic marksmen shoot much better if they fire between heartbeats, rather than squeezing the trigger during a heartbeat. The small vibration caused by a heartbeat seems to be sufficient to affect the marksman's aim. The following hypothetical data

demonstrate this phenomenon. A sample of $n = 8$ Olympic marksmen fires a series of rounds while a researcher records heartbeats. For each marksman, a score is recorded for shots fired during heartbeats and for shots fired between heartbeats. Do these data indicate a significant difference? Test with $\alpha = .05$.

Participant	During Heartbeats	Between Heartbeats
A	93	98
B	90	94
C	95	96
D	92	91
E	95	97
F	91	97
G	92	95
H	93	97

24. The Preview section of this chapter presented a repeated-measures research study demonstrating that swearing can help reduce pain (Stephens, Atkins, & Kingston, 2009). In the study, each participant was asked to plunge a hand into icy water and keep it there as long as the pain would allow. In one condition, the participants repeated their favorite curse words while their hands were in the water. In the other condition, the participants repeated a neutral word. Data similar to the results obtained in the study are shown in the following table.

a. Do these data indicate a significant difference in pain tolerance between the two conditions? Use a two-tailed test with $\alpha = .05$.

b. Compute r^2, the percentage of variance accounted for, to measure the size of the treatment effect.

c. Write a sentence demonstrating how the results of the hypothesis test and the measure of effect size would appear in a research report.

	Amount of Time (in Seconds)	
Participant	Swear Words	Neutral Words
1	94	59
2	70	61
3	52	47
4	83	60
5	46	35
6	117	92
7	69	53
8	39	30
9	51	56
10	73	61

PART III

REVIEW

After completing this part, you should be able to perform hypothesis tests and compute confidence intervals using *t* statistics. These include:

1. The single-sample *t* introduced in Chapter 9.
2. The independent-measures *t* introduced in Chapter 10.
3. The repeated-measures *t* introduced in Chapter 11.

In this part, we considered a set of three *t* statistics that are used to draw inferences about the means and mean differences for unknown populations. Because the populations are completely unknown, we rely on sample data to provide all of the necessary information. In particular, each inferential procedure begins by computing sample means and sample variances (or the corresponding *SS* values or standard deviations). Therefore, a good understanding of the definitions and formulas from Chapters 3 and 4 is a critical foundation for this section.

With three different *t* statistics available, the first problem is often deciding which one is appropriate for a specific research situation. Perhaps the best approach is to begin with a close look at the sample data.

1. For the single-sample *t* (Chapter 9), there is only one group of participants and only one score for each individual. With a single sample mean and a single sample variance, the *t* statistic can be used to test a hypothesis about a single unknown population mean or construct a confidence interval to estimate the population mean.
2. For the independent-measures *t*, there are two separate groups of participants who produce two groups of scores. The mean and variance are computed for each group, producing two sample means and two sample variances. After pooling the two variances, the *t* statistic uses the difference between the two sample means to test a hypothesis about the corresponding difference between the two unknown population means or estimate the population mean difference with a confidence interval. The null hypothesis always states that there is no difference between the two population means; $\mu_1 - \mu_2 = 0$.
3. For the repeated-measures *t*, there is only one group of participants but each individual is measured twice, at two different times and/or under two different treatment conditions. The two scores are then used to find a difference score for each person, and the mean and variance are computed for the sample of difference scores. The *t* statistic uses the sample mean difference to test a hypothesis about the corresponding population mean difference or estimate the population mean difference with a confidence interval. The null hypothesis always states that the mean for the population of difference scores is zero; $\mu_D = 0$.

REVIEW EXERCISES

1. People tend to evaluate the quality of their lives relative to others around them. In a demonstration of this phenomenon, Frieswijk, Buunk, Steverink, and Slaets (2004) conducted interviews with frail elderly people. In the interview, each person was compared with fictitious others who were worse off. After the interviews, the participants completed a life-satisfaction survey and reported more satisfaction with their own lives. Following are hypothetical data similar to those obtained in the research study, representing satisfaction scores for a sample of $n = 9$ older people who completed the interview. Assume that the average score on the life-satisfaction scale is $\mu = 20$. The scores for the sample are 18, 23, 24, 22, 19, 27, 23, 26, 25.

- **a.** Calculate the mean and standard deviation for the sample.
- **b.** Are the data sufficient to conclude that the people in this sample are significantly more satisfied than others in the general population? Use a one-tailed test with $\alpha = .05$.
- **c.** Compute Cohen's *d* to estimate the size of the effect.
- **d.** Compute the 90% confidence interval for the mean life-satisfaction score for people who participate in this type of interview.

2. In the problems at the end of Chapter 8, we presented a study indicating that people with visible tattoos are viewed more negatively than people without visible tattoos (Resenhoeft, Villa, & Wiseman, 2008). Suppose that a researcher intends to examine this phenomenon by asking participants to rate the attractiveness of women in a series of ten photographs. For one group of participants, none of the women has any visible tattoos. For a second group, however, the researcher modified one of the photographs by adding a tattoo of a butterfly on the woman's left arm. Using a 7-point rating scale, the $n = 15$ participants who viewed the photograph with no tattoo gave the woman an average rating of $M = 4.9$ with $SS = 15.0$. The $n = 15$ participants who saw the photograph with a tattoo gave the same woman an average rating of $M = 4.2$ with $SS = 18.6$.

- **a.** Does the existence of a tattoo have a significant effect on the attractiveness rating of the woman in the photograph? Use a two-tailed test with $\alpha = .05$.
- **b.** Compute r^2, the percentage of variance accounted for by the treatment, to measure the effect size.
- **c.** Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

3. The stimulant Ritalin has been shown to increase attention span and improve academic performance in children with ADHD (Evans, Pelham, Smith, et al., 2001). To demonstrate the effectiveness of the drug, a researcher selects a sample of $n = 20$ children diagnosed with the disorder and measures each child's attention span before and after taking the drug. The data show an average increase of attention span of $M_D = 4.8$ minutes with a variance of $s^2 = 125$ for the sample of difference scores.

a. Is this result sufficient to conclude that Ritalin significantly improves attention span? Use a one-tailed test with $\alpha = .05$.

b. Compute the 80% confidence interval for the mean change in attention span for the population.

This page intentionally left blank

PART

IV

Analysis of Variance: Tests for Differences Among Two or More Population Means

In Part III we presented a set of *t* statistics that use sample means and mean differences to draw inferences about the corresponding population means and mean differences. However, the *t* statistics are limited to situations that compare no more than two population means. Often, a research question involves the differences among more than two means and, in these situations, *t* tests are not appropriate. In this part we introduce a new hypothesis testing technique known as analysis of variance (ANOVA). ANOVA permits researchers to evaluate the mean differences among *two or more* populations using sample data. We present three different applications of ANOVA that apply to three distinct research situations:

1. Independent-measures designs: Using two or more separate samples to draw an inference about the mean differences between two or more unknown populations.
2. Repeated-measures designs: Using one sample, with each individual tested in two or more different treatment conditions, to draw an inference about the population mean differences among the conditions.
3. Two-factor designs: Allowing two independent variables to change simultaneously within one study to create combinations of treatment conditions involving both variables. The ANOVA then evaluates the mean differences attributed to each variable acting independently and to combinations of the two variables interacting together.

In the next three chapters we continue to examine statistical methods that use sample means as the foundation for drawing inferences about population means. The primary application of these

inferential methods is to help researchers interpret the outcome of their research studies. In a typical study, the goal is to demonstrate a difference between two or more treatment conditions. For example, a researcher hopes to demonstrate that a group of children who are exposed to violent TV programs behave more aggressively than children who are shown nonviolent TV programs. In this situation, the data consist of one sample mean representing the scores in one treatment condition and another sample mean representing the scores from a different treatment. The researcher hopes to find a difference between the sample means and would like to generalize the mean difference to the entire population.

The problem is that sample means can be different even when there are no differences whatsoever among the population means. As you saw in Chapter 1 (see Figure 1.2), two samples can have different means even when they are selected from the same population. Thus, even though a researcher may obtain a sample mean difference in a research study, it does not necessarily indicate that there is a mean difference in the population. As with the *t* tests presented in Part III, a hypothesis test is needed to determine whether the mean differences found in sample data are statistically significant. With more than two sample means, the appropriate hypothesis test is ANOVA.

CHAPTER

12

Introduction to Analysis of Variance

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Variability (Chapter 4)
- Sum of squares
- Sample variance
- Degrees of freedom
- Introduction to hypothesis testing (Chapter 8)
- The logic of hypothesis testing
- Independent-measures *t* statistic (Chapter 10)

Preview

"But I read the chapter four times! How could I possibly have failed the exam?!"

Most of you probably have had the experience of reading a textbook and suddenly realizing that you have no idea of what was said on the past few pages. Although you have been reading the words, your mind has wandered off, and the meaning of the words has never reached memory. In an influential paper on human memory, Craik and Lockhart (1972) proposed a *levels of processing* theory of memory that can account for this phenomenon. In general terms, this theory says that all perceptual and mental processing leaves behind a memory trace. However, the quality of the memory trace depends on the level or the depth of the processing. If you superficially skim the words in a book, your memory also is superficial. On the other hand, when you think about the meaning of the words and try to understand what you are reading, the result is a good, substantial memory that should serve you well on exams. In general, deeper processing results in better memory.

Rogers, Kuiper, and Kirker (1977) conducted an experiment demonstrating the effect of levels of processing. Participants in this experiment were shown lists of words and asked to answer questions about each word. The questions were designed to require different levels of processing, from superficial to deep. In one experimental condition, participants were simply asked to judge the physical characteristics of each printed word ("Is it printed in capital letters or lowercase letters?"). A second condition asked about the sound of each word ("Does it rhyme with 'boat'?"). In a third condition, participants were required to process the meaning of each word ("Does it have the same meaning as 'attractive'?"). The final condition required participants to understand each word and relate its meaning to themselves ("Does this word describe you?"). After going through the complete list, all participants were given a surprise memory test. As you can see in Figure 12.1, deeper processing resulted in better memory. Remember that the participants were not trying to memorize the words; they were simply reading through the list answering questions. However, the more they processed and understood the words, the better they recalled the words on the test.

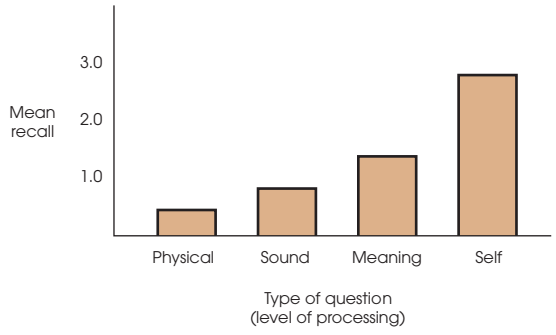


FIGURE 12.1

Mean recall as a function of the level of processing.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of personality and Social Psychology*, 35, 677–688. Copyright (1977) by the American Psychological Association. Adapted by permission of the author.

The Problem: In terms of human memory, the Rogers, Kuiper, and Kirker experiment is notable because it demonstrates the importance of "self" in memory. You are most likely to remember material that is directly related to you. In terms of statistics, however, this study is notable because it compares four different treatment conditions in a single experiment. We now have four different means and need a hypothesis test to evaluate the mean differences. Unfortunately, the *t* tests introduced in Chapter 10 and 11 are limited to comparing only two treatments. A new hypothesis test is needed for this kind of data.

The Solution: In this chapter we introduce a new hypothesis test known as *analysis of variance* that is designed to evaluate the mean differences from research studies producing two or more sample means. Although "two or more" may seem like a small step from "two," this new hypothesis testing procedure provides researchers with a tremendous gain in experimental sophistication. In this chapter, and the two that follow, we examine some of the many applications of analysis of variance.

12.1 INTRODUCTION

Analysis of variance (ANOVA) is a hypothesis-testing procedure that is used to evaluate mean differences between two or more treatments (or populations). As with all inferential procedures, ANOVA uses sample data as the basis for drawing general conclusions about populations. It may appear that ANOVA and *t* tests are simply two different ways of doing exactly the same job: testing for mean differences. In some respects, this is true—both tests use sample data to test hypotheses about population means. However, ANOVA has a tremendous advantage over *t* tests. Specifically, *t* tests are limited to situations in which there are only two treatments to compare. The major advantage of ANOVA is that it can be used to compare *two or more treatments*. Thus, ANOVA provides researchers with much greater flexibility in designing experiments and interpreting results.

Figure 12.2 shows a typical research situation for which ANOVA would be used. Note that the study involves three samples representing three populations. The goal of the analysis is to determine whether the mean differences observed among the samples provide enough evidence to conclude that there are mean differences among the three populations. Specifically, we must decide between two interpretations:

1. There really are no differences between the populations (or treatments). The observed differences between the sample means are caused by random, unsystematic factors (sampling error) that differentiate one sample from another.
2. The populations (or treatments) really do have different means, and these population mean differences are responsible for causing systematic differences between the sample means.

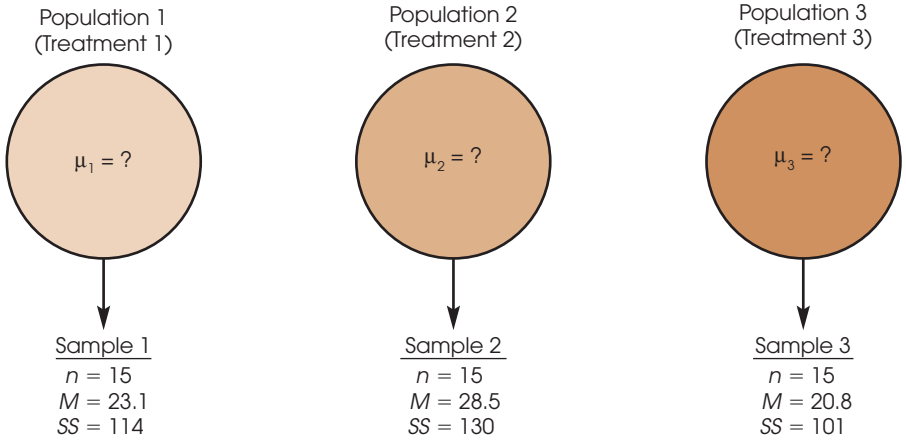
You should recognize that these two interpretations correspond to the two hypotheses (null and alternative) that are part of the general hypothesis-testing procedure.

TERMINOLOGY IN ANOVA

Before we continue, it is necessary to introduce some of the terminology that is used to describe the research situation shown in Figure 12.2. Recall (from Chapter 1) that when a researcher manipulates a variable to create the treatment conditions in an experiment, the variable is called an *independent variable*. For example, Figure 12.2 could represent

FIGURE 12.2

A typical situation in which ANOVA would be used. Three separate samples are obtained to evaluate the mean differences among three populations (or treatments) with unknown means.



a study examining driving performance under three different telephone conditions: driving with no phone, talking on a hands-free phone, and talking on a hand-held phone. Note that the three conditions are created by the researcher. On the other hand, when a researcher uses a nonmanipulated variable to designate groups, the variable is called a *quasi-independent variable*. For example, the three groups in Figure 12.2 could represent 6-year-old, 8-year-old, and 10-year-old children. In the context of ANOVA, an independent variable or a quasi-independent variable is called a *factor*. Thus, Figure 12.2 could represent an experimental study in which the telephone condition is the factor being evaluated or it could represent a nonexperimental study in which age is the factor being examined.

DEFINITION

In ANOVA, the variable (independent or quasi-independent) that designates the groups being compared is called a **factor**.

In addition, the individual groups or treatment conditions that are used to make up a factor are called the *levels* of the factor. For example, a study that examined performance under three different telephone conditions would have three levels of the factor.

DEFINITION

The individual conditions or values that make up a factor are called the **levels** of the factor.

Like the *t* tests presented in Chapters 10 and 11, ANOVA can be used with either an independent-measures or a repeated-measures design. Recall that an independent-measures design means that there is a separate group of participants for each of the treatments (or populations) being compared. In a repeated-measures design, on the other hand, the same group is tested in all of the different treatment conditions. In addition, ANOVA can be used to evaluate the results from a research study that involves more than one factor. For example, a researcher may want to compare two different therapy techniques, examining their immediate effectiveness as well as the persistence of their effectiveness over time. In this situation, the research study could involve two different groups of participants, one for each therapy, and measure each group at several different points in time. The structure of this design is shown in Figure 12.3. Notice that the study uses two factors, one independent-measures factor and one repeated-measures factor:

1. Factor 1: Therapy technique. A separate group is used for each technique (independent measures).
2. Factor 2: Time. Each group is tested at three different times (repeated measures).

In this case, the ANOVA would evaluate mean differences between the two therapies as well as mean differences between the scores obtained at different times. A study that combines two factors, like the one in Figure 12.3, is called a *two-factor design* or a *factorial design*.

The ability to combine different factors and to mix different designs within one study provides researchers with the flexibility to develop studies that address scientific questions that could not be answered by a single design using a single factor.

Although ANOVA can be used in a wide variety of research situations, this chapter introduces ANOVA in its simplest form. Specifically, we consider only *single-factor* designs. That is, we examine studies that have only one independent variable (or only one quasi-independent variable). Second, we consider only *independent-measures* designs; that is, studies that use a separate group of participants for each treatment condition. The basic logic and procedures that are presented in this chapter form the foundation for more complex applications of

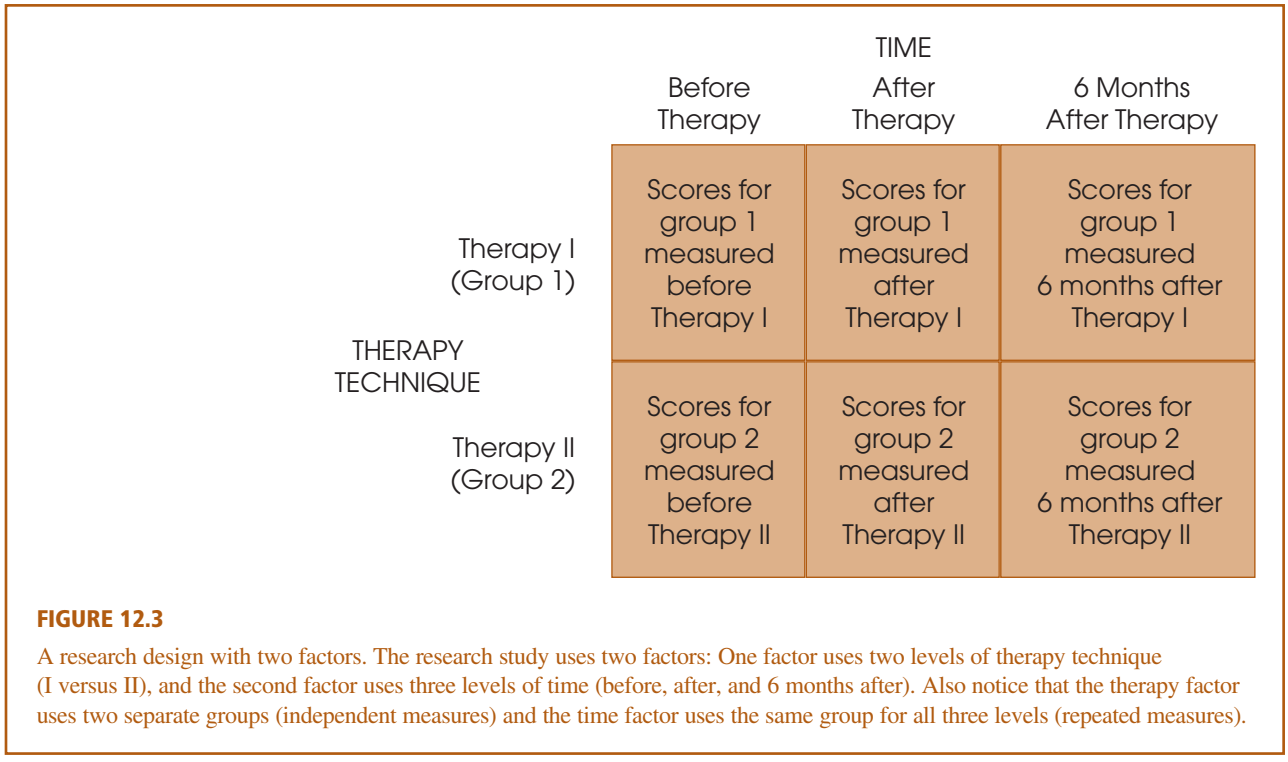


FIGURE 12.3

A research design with two factors. The research study uses two factors: One factor uses two levels of therapy technique (I versus II), and the second factor uses three levels of time (before, after, and 6 months after). Also notice that the therapy factor uses two separate groups (independent measures) and the time factor uses the same group for all three levels (repeated measures).

ANOVA. For example, in Chapter 13, we extend the analysis to single-factor, repeated-measures designs and in Chapter 14, we introduce two-factor designs. But for now, in this chapter, we limit our discussion of ANOVA to *single-factor, independent-measures* research studies.

STATISTICAL HYPOTHESES FOR ANOVA

The following example introduces the statistical hypotheses for ANOVA. Suppose that a researcher examined driving performance under three different telephone conditions: no phone, a hands-free phone, and a hand-held phone. Three samples of participants are selected, one sample for each treatment condition. The purpose of the study is to determine whether using a telephone affects driving performance. In statistical terms, we want to decide between two hypotheses: the null hypothesis (H_0), which states that the telephone condition has no effect, and the alternative hypothesis (H_1), which states that the telephone condition does affect driving. In symbols, the null hypothesis states

$$H_0: \mu_1 = \mu_2 = \mu_3$$

In words, the null hypothesis states that the telephone condition has no effect on driving performance. That is, the population means for the three telephone conditions are all the same. In general, H_0 states that there is no treatment effect.

The alternative hypothesis states that the population means are not all the same:

H_1: There is at least one mean difference among the populations.

In general, H_1 states that the treatment conditions are not all the same; that is, there is a real treatment effect. As always, the hypotheses are stated in terms of population parameters, even though we use sample data to test them.

Notice that we are not stating a specific alternative hypothesis. This is because many different alternatives are possible, and it would be tedious to list them all.

One alternative, for example, is that the first two populations are identical, but that the third is different. Another alternative states that the last two means are the same, but that the first is different. Other alternatives might be

H_1: $\mu_1 \neq \mu_2 \neq \mu_3$ (All three means are different.)

H_1: $\mu_1 = \mu_3$, but μ_2 is different.

We should point out that a researcher typically entertains only one (or at most a few) of these alternative hypotheses. Usually a theory or the outcomes of previous studies dictate a specific prediction concerning the treatment effect. For the sake of simplicity, we state a general alternative hypothesis rather than try to list all of the possible specific alternatives.

THE TEST STATISTIC FOR ANOVA

The test statistic for ANOVA is very similar to the independent-measures t statistic used in Chapter 10. For the t statistic, we first computed the standard error, which measures how much difference is expected between two sample means if there is no treatment effect (that is, if H_0 is true). Then we computed the t statistic with the following structure:

$$t = \frac{\text{obtained difference between two sample means}}{\text{standard error (the difference expected with no treatment effect)}}$$

For ANOVA, however, we want to compare differences among two *or more* sample means. With more than two samples, the concept of "difference between sample means" becomes difficult to define or measure. For example, if there are only two samples and they have means of $M = 20$ and $M = 30$, then there is a 10-point difference between the sample means. Suppose, however, that we add a third sample with a mean of $M = 35$. Now how much difference is there between the sample means? It should be clear that we have a problem. The solution to this problem is to use variance to define and measure the size of the differences among the sample means. Consider the following two sets of sample means:

Set 1	Set 2
$M_1 = 20$	$M_1 = 28$
$M_2 = 30$	$M_2 = 30$
$M_3 = 35$	$M_3 = 31$

If you compute the variance for the three numbers in each set, then the variance is $s^2 = 58.33$ for set 1 and the variance is $s^2 = 2.33$ for set 2. Notice that the two variances provide an accurate representation of the size of the differences. In set 1, there are relatively large differences between sample means and the variance is relatively large. In set 2, the mean differences are small and the variance is small.

Thus, we can use variance to measure sample mean differences when there are two or more samples. The test statistic for ANOVA uses this fact to compute an F-*ratio* with the following structure:

$$F = \frac{\text{variance (differences) between sample means}}{\text{variance (differences) expected with no treatment effect}}$$

Note that the F-ratio has the same basic structure as the t statistic but is based on *variance* instead of sample mean *difference*. The variance in the numerator of the F-ratio provides a single number that measures the differences among all of the sample means. The variance in the denominator of the F-ratio, like the standard error in the

denominator of the *t* statistic, measures the mean differences that would be expected if there were no treatment effect. Thus, the *t* statistic and the *F*-ratio provide the same basic information. In each case, a large value for the test statistic provides evidence that the sample mean differences (numerator) are larger than would be expected if there were no treatment effects (denominator).

TYPE I ERRORS AND MULTIPLE-HYPOTHESIS TESTS

If we already have *t* tests for comparing mean differences, you might wonder why ANOVA is necessary. Why create a whole new hypothesis-testing procedure that simply duplicates what the *t* tests can already do? The answer to this question is based in a concern about Type I errors.

Remember that each time you do a hypothesis test, you select an alpha level that determines the risk of a Type I error. With $\alpha = .05$, for example, there is a 5%, or a 1-in-20, risk of a Type I error. Often a single experiment requires several hypothesis tests to evaluate all the mean differences. However, each test has a risk of a Type I error, and the more tests you do, the more risk there is.

For this reason, researchers often make a distinction between the *testwise alpha level* and the *experimentwise alpha level*. The testwise alpha level is simply the alpha level that you select for each individual hypothesis test. The experimentwise alpha level is the total probability of a Type I error accumulated from all of the separate tests in the experiment. As the number of separate tests increases, so does the experimentwise alpha level.

DEFINITIONS

The **testwise alpha level** is the risk of a Type I error, or alpha level, for an individual hypothesis test.

When an experiment involves several different hypothesis tests, the **experimentwise alpha level** is the total probability of a Type I error that is accumulated from all of the individual tests in the experiment. Typically, the experimentwise alpha level is substantially greater than the value of alpha used for any one of the individual tests.

For example, an experiment involving three treatments would require three separate *t* tests to compare all of the mean differences:

Test 1 compares treatment I with treatment II.

Test 2 compares treatment I with treatment III.

Test 3 compares treatment II with treatment III.

If all tests use $\alpha = .05$, then there is a 5% risk of a Type I error for the first test, a 5% risk for the second test, and another 5% risk for the third test. The three separate tests accumulate to produce a relatively large experimentwise alpha level. The advantage of ANOVA is that it performs all three comparisons simultaneously in one hypothesis test. Thus, no matter how many different means are being compared, ANOVA uses one test with one alpha level to evaluate the mean differences, and thereby avoids the problem of an inflated experimentwise alpha level.

12.2 THE LOGIC OF ANOVA

The formulas and calculations required in ANOVA are somewhat complicated, but the logic that underlies the whole procedure is fairly straightforward. Therefore, this section gives a general picture of ANOVA before we start looking at the details. We introduce the logic of ANOVA with the help of the hypothetical data in Table 12.1.

TABLE 12.1

Hypothetical data from an experiment examining driving performance under three telephone conditions.*

Treatment 1: No Phone (Sample 1)	Treatment 2: Hand-Held (Sample 2)	Treatment 3: Hands-Free (Sample 3)
4	0	1
3	1	2
6	3	2
3	1	0
4	0	0
$M = 4$	$M = 1$	$M = 1$

*Note that there are three separate samples, with $n = 5$ in each sample. The dependent variable is a measure of performance in a driving simulator.

These data represent the results of an independent-measures experiment comparing performance in a driving simulator under three telephone conditions.

One obvious characteristic of the data in Table 12.1 is that the scores are not all the same. In everyday language, the scores are different; in statistical terms, the scores are variable. Our goal is to measure the amount of variability (the size of the differences) and to explain why the scores are different.

The first step is to determine the total variability for the entire set of data. To compute the total variability, we combine all of the scores from all of the separate samples to obtain one general measure of variability for the complete experiment. Once we have measured the total variability, we can begin to break it apart into separate components. The word *analysis* means dividing into smaller parts. Because we are going to analyze variability, the process is called *analysis of variance*. This analysis process divides the total variability into two basic components.

1. **Between-Treatments Variance.** Looking at the data in Table 12.1, we clearly see that much of the variability in the scores results from general differences between treatment conditions. For example, the scores in the no-phone condition tend to be much higher ($M = 4$) than the scores in the hand-held condition ($M = 1$). We calculate the variance between treatments to provide a measure of the overall differences between treatment conditions. Notice that the variance between treatments is really measuring the differences between sample means.
2. **Within-Treatment Variance.** In addition to the general differences between treatment conditions, there is variability within each sample. Looking again at Table 12.1, we see that the scores in the no-phone condition are not all the same; they are variable. The within-treatments variance provides a measure of the variability inside each treatment condition.

Analyzing the total variability into these two components is the heart of ANOVA. We now examine each of the components in more detail.

BETWEEN-TREATMENTS VARIANCE

Remember that calculating variance is simply a method for measuring how big the differences are for a set of numbers. When you see the term *variance,* you can automatically translate it into the term *differences*. Thus, the *between-treatments variance* simply measures how much difference exists between the treatment conditions. There are two possible explanations for these between-treatment differences:

1. The differences between treatments are not caused by any treatment effect but are simply the naturally occurring, random, and unsystematic differences that

exist between one sample and another. That is, the differences are the result of sampling error.

2. The differences between treatments have been caused by the *treatment effects*. For example, if using a telephone really does interfere with driving performance, then scores in the telephone conditions should be systematically lower than scores in the no-phone condition.

Thus, when we compute the between-treatments variance, we are measuring differences that could be caused by a systematic treatment effect or could simply be random and unsystematic mean differences caused by sampling error. To demonstrate that there really is a treatment effect, we must establish that the differences between treatments are bigger than would be expected by sampling error alone. To accomplish this goal, we determine how big the differences are when there is no systematic treatment effect; that is, we measure how much difference (or variance) can be explained by random and unsystematic factors. To measure these differences, we compute the variance within treatments.

WITHIN-TREATMENTS VARIANCE

Inside each treatment condition, we have a set of individuals who all receive exactly the same treatment; that is, the researcher does not do anything that would cause these individuals to have different scores. In Table 12.1, for example, the data show that five individuals were tested while talking on a hand-held phone (sample 2). Although these five individuals all received exactly the same treatment, their scores are different. Why are the scores different? The answer is that there is no specific cause for the differences. Instead, the differences that exist within a treatment represent random and unsystematic differences that occur when there are no treatment effects causing the scores to be different. Thus, the *within-treatments variance* provides a measure of how big the differences are when H_0 is true.

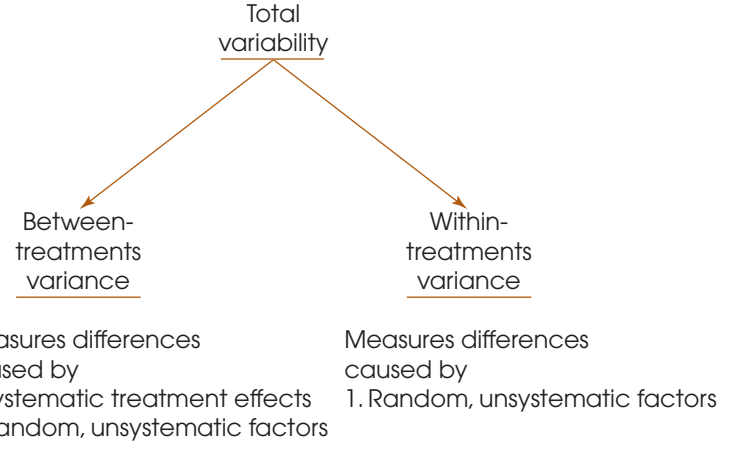
Figure 12.4 shows the overall ANOVA and identifies the sources of variability that are measured by each of the two basic components.

THE *F*-RATIO: THE TEST STATISTIC FOR ANOVA

Once we have analyzed the total variability into two basic components (between treatments and within treatments), we simply compare them. The comparison is made by

FIGURE 12.4

The independent-measures ANOVA partitions, or analyzes, the total variability into two components: variance between treatments and variance within treatments.



computing an F-*ratio*. For the independent-measures ANOVA, the *F*-ratio has the following structure:

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}} = \frac{\text{differences including any treatment effects}}{\text{differences with no treatment effects}} \tag{12.1}$$

When we express each component of variability in terms of its sources (see Figure 12.4), the structure of the *F*-ratio is

$$F = \frac{\text{systematic treatment effects} + \text{random, unsystematic differences}}{\text{random, unsystematic differences}} \tag{12.2}$$

The value obtained for the *F*-ratio helps determine whether any treatment effects exist. Consider the following two possibilities:

1. When there are no systematic treatment effects, the differences between treatments (numerator) are entirely caused by random, unsystematic factors. In this case, the numerator and the denominator of the *F*-ratio are both measuring random differences and should be roughly the same size. With the numerator and denominator roughly equal, the *F*-ratio should have a value around 1.00. In terms of the formula, when the treatment effect is zero, we obtain

$$F = \frac{0 + \text{random, unsystematic differences}}{\text{random, unsystematic differences}}$$

   Thus, an *F*-ratio near 1.00 indicates that the differences between treatments (numerator) are random and unsystematic, just like the differences in the denominator. With an *F*-ratio near 1.00, we conclude that there is no evidence to suggest that the treatment has any effect.

2. When the treatment does have an effect, causing systematic differences between samples, then the combination of systematic and random differences in the numerator should be larger than the random differences alone in the denominator. In this case, the numerator of the *F*-ratio should be noticeably larger than the denominator, and we should obtain an *F*-ratio that is substantially larger than 1.00. Thus, a large *F*-ratio is evidence for the existence of systematic treatment effects; that is, there are consistent differences between treatments.

Because the denominator of the *F*-ratio measures only random and unsystematic variability, it is called the *error term*. The numerator of the *F*-ratio always includes the same unsystematic variability as in the error term, but it also includes any systematic differences caused by the treatment effect. The goal of ANOVA is to find out whether a treatment effect exists.

DEFINITION

For ANOVA, the denominator of the *F*-ratio is called the **error term**. The error term provides a measure of the variance caused by random, unsystematic differences. When the treatment effect is zero (H_0 is true), the error term measures the same sources of variance as the numerator of the *F*-ratio, so the value of the *F*-ratio is expected to be nearly equal to 1.00.

LEARNING CHECK

1. ANOVA is a statistical procedure that compares two or more treatment conditions for differences in variance. (True or false?)
2. In ANOVA, what value is expected, on the average, for the F-ratio when the null hypothesis is true?
3. What happens to the value of the F-ratio if differences between treatments are increased? What happens to the F-ratio if variability inside the treatments is increased?
4. In ANOVA, the total variability is partitioned into two parts. What are these two variability components called, and how are they used in the F-ratio?

ANSWERS

1. False. Although ANOVA uses variance in the computations, the purpose of the test is to evaluate differences in *means* between treatments.
2. When H_0 is true, the expected value for the F-ratio is 1.00 because the top and bottom of the ratio are both measuring the same variance.
3. As differences between treatments increase, the F-ratio increases. As variability within treatments increases, the F-ratio decreases.
4. The two components are between-treatments variability and within-treatments variability. Between-treatments variance is the numerator of the F-ratio, and within-treatments variance is the denominator.

12.3 ANOVA NOTATION AND FORMULAS

Because ANOVA typically is used to examine data from more than two treatment conditions (and more than two samples), we need a notational system to keep track of all the individual scores and totals. To help introduce this notational system, we use the hypothetical data from Table 12.1 again. The data are reproduced in Table 12.2 along with some of the notation and statistics that are described in the following list.

1. The letter k is used to identify the number of treatment conditions—that is, the number of levels of the factor. For an independent-measures study, k also specifies the number of separate samples. For the data in Table 12.2, there are three treatments, so $k = 3$.
2. The number of scores in each treatment is identified by a lowercase letter n. For the example in Table 12.2, $n = 5$ for all the treatments. If the samples are of different sizes, you can identify a specific sample by using a subscript. For example, n_2 is the number of scores in treatment 2.
3. The total number of scores in the entire study is specified by a capital letter N. When all of the samples are the same size (n is constant), $N = kn$. For the data in Table 12.2, there are $n = 5$ scores in each of the $k = 3$ treatments, so we have a total of $N = 3(5) = 15$ scores in the entire study.
4. The sum of the scores (ΣX) for each treatment condition is identified by the capital letter T (for treatment total). The total for a specific treatment can be identified by adding a numerical subscript to the T. For example, the total for the second treatment in Table 12.2 is $T_2 = 5$.
5. The sum of all of the scores in the research study (the grand total) is identified by G. You can compute G by adding up all N scores or by adding up the treatment totals: $G = \Sigma T$.
6. Although there is no new notation involved, we also have computed SS and M for each sample, and we have calculated ΣX^2 for the entire set of $N = 15$ scores

TABLE 12.2

The same data that appeared in Table 12.1 with summary values and notation appropriate for an ANOVA.

Telephone Conditions			
Treatment 1 No Phone (Sample 1)	Treatment 2 Hand-Held Phone (Sample 2)	Treatment 3 Hands-Free Phone (Sample 3)	
4	0	1	$\Sigma X^2 = 106$
3	1	2	$G = 30$
6	3	2	$N = 15$
3	1	0	$k = 3$
4	0	0	
$T_1 = 20$	$T_2 = 5$	$T_3 = 5$	
$SS_1 = 6$	$SS_2 = 6$	$SS_3 = 4$	
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	
$M_1 = 4$	$M_2 = 1$	$M_3 = 1$	

in the study. These values are given in Table 12.2 and are important in the formulas and calculations for ANOVA.

Finally, we should note that there is no universally accepted notation for ANOVA. Although we are using *G*s and *T*s, for example, you may find that other sources use other symbols.

ANOVA FORMULAS

Because ANOVA formulas require ΣX for each treatment and ΣX for the entire set of scores, we have introduced new notation (*T* and *G*) to help identify which ΣX is being used. Remember: *T* stands for *treatment total*, and *G* stands for *grand total*.

Because ANOVA requires extensive calculations and many formulas, one common problem for students is simply keeping track of the different formulas and numbers. Therefore, we examine the general structure of the procedure and look at the organization of the calculations before we introduce the individual formulas.

1. The final calculation for ANOVA is the *F*-ratio, which is composed of two variances:

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

2. Each of the two variances in the *F*-ratio is calculated using the basic formula for sample variance.

$$\text{sample variance} = s^2 = \frac{SS}{df}$$

Therefore, we need to compute an *SS* and a *df* for the variance between treatments (numerator of *F*), and we need another *SS* and *df* for the variance within treatments (denominator of *F*). To obtain these *SS* and *df* values, we must go through two separate analyses: First, compute *SS* for the total study, and analyze it in two components (between and within). Then compute *df* for the total study, and analyze it in two components (between and within).

Thus, the entire process of ANOVA requires nine calculations: three values for *SS*, three values for *df*, two variances (between and within), and a final *F*-ratio. However, these nine calculations are all logically related and are all directed toward finding the final *F*-ratio. Figure 12.5 shows the logical structure of ANOVA calculations.

FIGURE 12.5
The structure and sequence of calculations for the ANOVA.

The final goal for the ANOVA is an *F*-ratio	$F = \dfrac{\text{Variance between treatments}}{\text{Variance within treatments}}$
Each variance in the *F*-ratio is computed as *SS*/*df*	$\text{Variance between treatments} = \dfrac{SS \text{ between}}{df \text{ between}}$ $\quad \text{Variance within treatments} = \dfrac{SS \text{ within}}{df \text{ within}}$
To obtain each of the *SS* and *df* values, the total variability is analyzed into the two components	*SS* total → *SS* between, *SS* within; *df* total → *df* between, *df* within

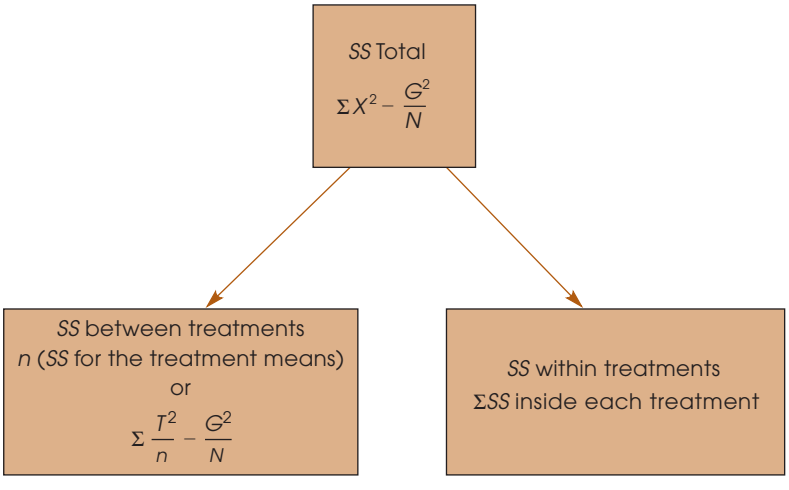
ANALYSIS OF THE SUM OF SQUARES (*SS*)

The ANOVA requires that we first compute a total sum of squares and then partition this value into two components: between treatments and within treatments. This analysis is outlined in Figure 12.6. We will examine each of the three components separately.

1. **Total Sum of Squares, SS_{total}.** As the name implies, SS_{total} is the sum of squares for the entire set of N scores. As described in Chapter 4 (pp. 111–112), this *SS* value can be computed using either a definitional or a computational formula. However, ANOVA typically involves a large number of scores and the mean is often not a whole number. Therefore, it is usually much easier to calculate SS_{total} using the computational formula:

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

FIGURE 12.6
Partitioning the sum of squares (*SS*) for the independent-measures ANOVA.



To make this formula consistent with the ANOVA notation, we substitute the letter G in place of ΣX and obtain

$$SS_{\text{total}} = \Sigma X^2 - \frac{G^2}{N} \tag{12.3}$$

Applying this formula to the set of data in Table 12.2, we obtain

$$SS_{\text{total}} = 106 - \frac{30^2}{15}$$

$$= 106 - 60$$

$$= 46$$

2. **Within-Treatments Sum of Squares, $SS_{\text{within treatments}}$.** Now we are looking at the variability inside each of the treatment conditions. We already have computed the SS within each of the three treatment conditions (Table 12.2): $SS_1 = 6$, $SS_2 = 6$, and $SS_3 = 4$. To find the overall within-treatment sum of squares, we simply add these values together:

$$SS_{\text{within treatments}} = \Sigma SS_{\text{inside each treatment}} \tag{12.4}$$

For the data in Table 12.2, this formula gives

$$SS_{\text{within treatments}} = 6 + 6 + 4$$

$$= 16$$

3. **Between-Treatments Sum of Squares, $SS_{\text{between treatments}}$.** Before we introduce any equations for $SS_{\text{between treatments}}$, consider what we have found so far. The total variability for the data in Table 12.2 is $SS_{\text{total}} = 46$. We intend to partition this total into two parts (see Figure 12.5). One part, $SS_{\text{within treatments}}$, has been found to be equal to 16. This means that $SS_{\text{between treatments}}$ must be equal to 30 so that the two parts (16 and 30) add up to the total (46). Thus, the value for $SS_{\text{between treatments}}$ can be found simply by subtraction:

$$SS_{\text{between}} = SS_{\text{total}} - SS_{\text{within}} \tag{12.5}$$

To simplify the notation, we use the subscripts *between* and *within* in place of *between treatments* and *within treatments.*

However, it is also possible to compute SS_{between} independently, then check your calculations by ensuring that the two components, between and within, add up to the total. Box 12.1 presents two different formulas for calculating SS_{between} directly from the data.

Computing SS_{between} Including the two formulas in Box 12.1, we have presented three different equations for computing SS_{between}. Rather than memorizing all three, however, we suggest that you pick one formula and use it consistently. There are two reasonable alternatives to use. The simplest is Equation 12.5, which finds SS_{between} simply by subtraction: First you compute SS_{total} and SS_{within}, then subtract:

$$SS_{\text{between}} = SS_{\text{total}} - SS_{\text{within}}$$

The second alternative is to use Equation 12.7, which computes SS_{between} using the treatment totals (the T values). The advantage of this alternative is that it provides a way to check your arithmetic: Calculate SS_{total}, SS_{between}, and SS_{within} separately, and then check to be sure that the two components add up to equal SS_{total}.

BOX 12.1 ALTERNATIVE FORMULAS FOR $SS_{between}$

Recall that the variability between treatments is measuring the differences between treatment means. Conceptually, the most direct way of measuring the amount of variability among the treatment means is to compute the sum of squares for the set of sample means, SS_{means}. For the data in Table 12.2, the samples means are 4, 1, and 1. These three values produce $SS_{means} = 6$. However, each of the three means represents a group of $n = 5$ scores. Therefore, the final value for $SS_{between}$ is obtained by multiplying SS_{means} by n.

$$SS_{between} = n(SS_{means}) \tag{12.6}$$

For the data in Table 12.2, we obtain

$$SS_{between} = n(SS_{means}) = 5(6) = 30$$

Unfortunately, Equation 12.6 can only be used when all of the samples are exactly the same size (equal *n*s), and the equation can be very awkward, especially when the treatment means are not whole numbers. Therefore, we also present a computational formula for $SS_{between}$ that uses the treatment totals (T) instead of the treatment means.

$$SS_{between} = \Sigma \frac{T^2}{n} - \frac{G^2}{N} \tag{12.7}$$

For the data in Table 12.2, this formula produces:

$$SS_{between} = \frac{20^2}{5} + \frac{5^2}{5} + \frac{5^2}{5} - \frac{30^2}{15}$$
$$= 80 + 5 + 5 - 60$$
$$= 90 - 60$$
$$= 30$$

Note that all three techniques (Equations 12.5, 12.6, and 12.7) produce the same result, $SS_{between} = 30$.

Using Equation 12.6, which computes *SS* for the set of sample means, is usually not a good choice. Unless the sample means are all whole numbers, this equation can produce very tedious calculations. In most situations, one of the other two equations is a better alternative.

THE ANALYSIS OF DEGREES OF FREEDOM (*DF*)

The analysis of degrees of freedom (*df*) follows the same pattern as the analysis of *SS*. First, we find *df* for the total set of N scores, and then we partition this value into two components: degrees of freedom between treatments and degrees of freedom within treatments. In computing degrees of freedom, there are two important considerations to keep in mind:

1. Each *df* value is associated with a specific *SS* value.
2. Normally, the value of *df* is obtained by counting the number of items that were used to calculate *SS* and then subtracting 1. For example, if you compute *SS* for a set of n scores, then $df = n - 1$.

With this in mind, we examine the degrees of freedom for each part of the analysis.

1. **Total Degrees of Freedom, df_{total}.** To find the *df* associated with SS_{total}, you must first recall that this *SS* value measures variability for the entire set of N scores. Therefore, the *df* value is

$$df_{total} = N - 1 \tag{12.8}$$

For the data in Table 12.2, the total number of scores is $N = 15$, so the total degrees of freedom are

$$df_{total} = 15 - 1$$
$$= 14$$

2. **Within-Treatments Degrees of Freedom, df_{within}.** To find the *df* associated with SS_{within}, we must look at how this *SS* value is computed. Remember, we first find *SS* inside of each of the treatments and then add these values together. Each of the treatment *SS* values measures variability for the *n* scores in the treatment, so each *SS* has $df = n - 1$. When all of these individual treatment values are added together, we obtain

$$df_{within} = \Sigma(n - 1) = \Sigma df_{\text{in each treatment}} \tag{12.9}$$

For the experiment we have been considering, each treatment has $n = 5$ scores. This means there are $n - 1 = 4$ degrees of freedom inside each treatment. Because there are three different treatment conditions, this gives a total of 12 for the within-treatments degrees of freedom. Notice that this formula for *df* simply adds up the number of scores in each treatment (the *n* values) and subtracts 1 for each treatment. If these two stages are done separately, you obtain

$$df_{within} = N - k \tag{12.10}$$

(Adding up all the *n* values gives *N*. If you subtract 1 for each treatment, then altogether you have subtracted *k* because there are *k* treatments.) For the data in Table 12.2, $N = 15$ and $k = 3$, so

$$df_{within} = 15 - 3$$
$$= 12$$

3. **Between-Treatments Degrees of Freedom, $df_{between}$.** The *df* associated with $SS_{between}$ can be found by considering how the *SS* value is obtained. This *SS* formulas measure the variability for the set of treatments (totals or means). To find $df_{between}$, simply count the number of treatments and subtract 1. Because the number of treatments is specified by the letter *k*, the formula for *df* is

$$df_{between} = k - 1 \tag{12.11}$$

For the data in Table 12.2, there are three different treatment conditions (three *T* values or three sample means), so the between-treatments degrees of freedom are computed as follows:

$$df_{between} = 3 - 1$$
$$= 2$$

Notice that the two parts we obtained from this analysis of degrees of freedom add up to equal the total degrees of freedom:

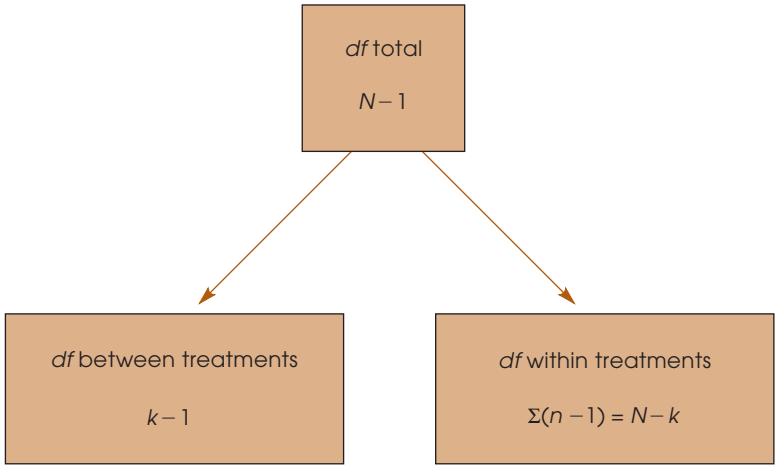
$$df_{total} = df_{within} + df_{between}$$
$$14 = 12 + 2$$

The complete analysis of degrees of freedom is shown in Figure 12.7.

As you are computing the *SS* and *df* values for ANOVA, keep in mind that the labels that are used for each value can help you understand the formulas. Specifically,

1. The term **total** refers to the entire set of scores. We compute *SS* for the whole set of *N* scores, and the *df* value is simply $N - 1$.

FIGURE 12.7
Partitioning degrees of freedom (*df*) for the independent-measures ANOVA.



2. The term **within treatments** refers to differences that exist inside the individual treatment conditions. Thus, we compute *SS* and *df* inside each of the separate treatments.
3. The term **between treatments** refers to differences from one treatment to another. With three treatments, for example, we are comparing three different means (or totals) and have $df = 3 - 1 = 2$.

CALCULATION OF VARIANCES (*MS*) AND THE *F*-RATIO

The next step in the ANOVA procedure is to compute the variance between treatments and the variance within treatments, which are used to calculate the *F*-ratio (see Figure 12.5).

In ANOVA, it is customary to use the term *mean square,* or simply *MS,* in place of the term *variance*. Recall (from Chapter 4) that variance is defined as the mean of the squared deviations. In the same way that we use *SS* to stand for the sum of the squared deviations, we now use *MS* to stand for the mean of the squared deviations. For the final *F*-ratio we need an *MS* (variance) between treatments for the numerator and an *MS* (variance) within treatments for the denominator. In each case

$$MS \text{ (variance)} = s^2 = \frac{SS}{df} \tag{12.12}$$

For the data we have been considering,

$$MS_{\text{between}} = s^2_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{30}{2} = 15$$

and

$$MS_{\text{within}} = s^2_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{16}{12} = 1.33$$

We now have a measure of the variance (or differences) between the treatments and a measure of the variance within the treatments. The *F*-ratio simply compares these two variances:

$$F = \frac{s^2_{\text{between}}}{s^2_{\text{within}}} = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{12.13}$$

For the experiment we have been examining, the data give an *F*-ratio of

$$F = \frac{15}{1.33} = 11.28$$

For this example, the obtained value of $F = 11.28$ indicates that the numerator of the *F*-ratio is substantially bigger than the denominator. If you recall the conceptual structure of the *F*-ratio as presented in Equations 12.1 and 12.2, the *F* value we obtained indicates that the differences between treatments are more than 11 times bigger than what would be expected if there were no treatment effect. Stated in terms of the experimental variables: using a telephone while driving does appear to have an effect on driving performance. However, to properly evaluate the *F*-ratio, we must select an α level and consult the *F*-distribution table that is discussed in the next section.

ANOVA Summary Tables It is useful to organize the results of the analysis in one table called an *ANOVA summary table*. The table shows the source of variability (between treatments, within treatments, and total variability), *SS*, *df*, *MS*, and *F*. For the previous computations, the ANOVA summary table is constructed as follows:

Source	*SS*	*df*	*MS*	
Between treatments	30	2	15	$F = 11.28$
Within treatments	16	12	1.33	
Total	46	14		

Although these tables are no longer used in published reports, they are a common part of computer printouts, and they do provide a concise method for presenting the results of an analysis. (Note that you can conveniently check your work: Adding the first two entries in the *SS* column, 30 + 16, produces SS_{total}. The same applies to the *df* column.) When using ANOVA, you might start with a blank ANOVA summary table and then fill in the values as they are calculated. With this method, you are less likely to "get lost" in the analysis, wondering what to do next.

LEARNING CHECK

1. Calculate SS_{total}, SS_{between}, and SS_{within} for the following set of data:

Treatment 1	Treatment 2	Treatment 3	
$n = 10$	$n = 10$	$n = 10$	$N = 30$
$T = 10$	$T = 20$	$T = 30$	$G = 60$
$SS = 27$	$SS = 16$	$SS = 23$	$\Sigma X^2 = 206$

2. A researcher uses an ANOVA to compare three treatment conditions with a sample of $n = 8$ in each treatment. For this analysis, find df_{total}, df_{between}, and df_{within}.

3. A researcher reports an F-ratio with $df_{between} = 2$ and $df_{within} = 30$ for an independent-measures ANOVA. How many treatment conditions were compared in the experiment? How many subjects participated in the experiment?

4. A researcher conducts an experiment comparing four treatment conditions with a separate sample of $n = 6$ in each treatment. An ANOVA is used to evaluate the data, and the results of the ANOVA are presented in the following table. Complete all missing values in the table. *Hint:* Begin with the values in the *df* column.

Source	*SS*	*df*	*MS*	
Between treatments	__	__	__	$F =$ ____
Within treatments	__	__	2	
Total	58	__		

ANSWERS

1. $SS_{total} = 86$; $SS_{between} = 20$; $SS_{within} = 66$
2. $df_{total} = 23$; $df_{between} = 2$; $df_{within} = 21$
3. There were 3 treatment conditions ($df_{between} = k - 1 = 2$). A total of $N = 33$ individuals participated ($df_{within} = 30 = N - k$).
4.

Source	*SS*	*df*	*MS*	
Between treatments	18	3	6	$F = 3.00$
Within treatments	40	20	2	
Total	58	23		

12.4 THE DISTRIBUTION OF *F*-RATIOS

In ANOVA, the F-ratio is constructed so that the numerator and denominator of the ratio are measuring exactly the same variance when the null hypothesis is true (see Equation 12.2). In this situation, we expect the value of F to be around 1.00.

If the null hypothesis is false, then the F-ratio should be much greater than 1.00. The problem now is to define precisely which values are "around 1.00" and which are "much greater than 1.00." To answer this question, we need to look at all of the possible F values—that is, the *distribution of* F-*ratios*.

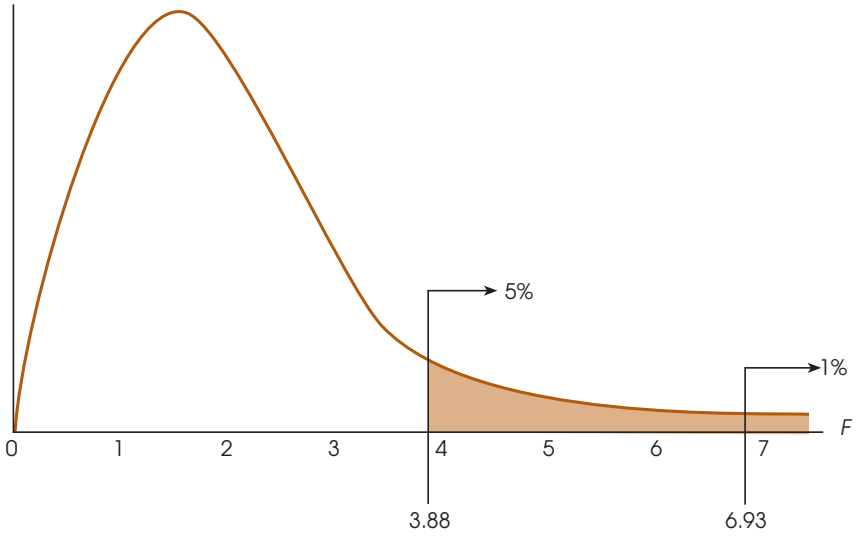
Before we examine this distribution in detail, you should note two obvious characteristics:

1. Because F-ratios are computed from two variances (the numerator and denominator of the ratio), F values always are positive numbers. Remember that variance is always positive.
2. When H_0 is true, the numerator and denominator of the F-ratio are measuring the same variance. In this case, the two sample variances should be about the same size, so the ratio should be near 1. In other words, the distribution of F-ratios should pile up around 1.00.

With these two factors in mind, we can sketch the distribution of F-ratios. The distribution is cut off at zero (all positive values), piles up around 1.00, and then tapers off to the right (Figure 12.8). The exact shape of the F distribution depends on the degrees

FIGURE 12.8

The distribution of F-ratios with $df = 2, 12$. Of all the values in the distribution, only 5% are larger than $F = 3.88$, and only 1% are larger than $F = 6.93$.



of freedom for the two variances in the F-ratio. You should recall that the precision of a sample variance depends on the number of scores or the degrees of freedom. In general, the variance for a large sample (large df) provides a more accurate estimate of the population variance. Because the precision of the MS values depends on df, the shape of the F distribution also depends on the df values for the numerator and denominator of the F-ratio. With very large df values, nearly all of the F-ratios are clustered very near to 1.00. With the smaller df values, the F distribution is more spread out.

THE *F* DISTRIBUTION TABLE

For ANOVA, we expect F near 1.00 if H_0 is true, and we expect a large value for F if H_0 is not true. In the F distribution, we need to separate those values that are reasonably near 1.00 from the values that are significantly greater than 1.00. These critical values are presented in an F distribution table in Appendix B, page 705. A portion of the F distribution table is shown in Table 12.3. To use the table, you must know the df values for the F-ratio (numerator and denominator), and you must know the alpha level for the hypothesis test. It is customary for an F table to have the df values for the numerator of the F-ratio printed across the top of the table. The df values for the denominator of F are printed in a column on the left-hand side. For the experiment we have been considering, the numerator of the F-ratio (between treatments) has $df = 2$, and the denominator of the F-ratio (within treatments) has $df = 12$. This F-ratio is said to have "degrees of freedom equal to 2 and 12." The degrees of freedom would be written as $df = 2, 12$. To use the table, you would first find $df = 2$ across the top of the table and $df = 12$ in the first column. When you line up these two values, they point to a pair of numbers in the middle of the table. These numbers give the critical cutoffs for $\alpha = .05$ and $\alpha = .01$. With $df = 2, 12$, for example, the numbers in the table are 3.88 and 6.93. Thus, only 5% of the distribution ($\alpha = .05$) corresponds to values greater than 3.88, and only1% of the distribution ($\alpha = .01$) corresponds to values greater than 6.93 (see Figure 12.8).

In the experiment comparing driving performance under different telephone conditions, we obtained an F-ratio of 11.28. According to the critical cutoffs in Figure 12.8, this value is extremely unlikely (it is in the most extreme 1%). Therefore, we would reject H_0 with an α level of either .05 or .01, and conclude that the different telephone conditions significantly affect driving performance.

TABLE 12.3

A portion of the *F* distribution table. Entries in roman type are critical values for the .05 level of significance, and bold type values are for the .01 level of significance. The critical values for $df = 2, 12$ have been highlighted (see text).

Degrees of Freedom: Denominator	Degrees of Freedom: Numerator					
	1	2	3	4	5	6
10	4.96	4.10	3.71	3.48	3.33	3.22
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**
11	4.84	3.98	3.59	3.36	3.20	3.09
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**
12	4.75	3.88	3.49	3.26	3.11	3.00
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**
13	4.67	3.80	3.41	3.18	3.02	2.92
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**
14	4.60	3.74	3.34	3.11	2.96	2.85
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**

LEARNING CHECK

1. A researcher obtains $F = 4.18$ with $df = 2, 15$. Is this value sufficient to reject H_0 with $\alpha = .05$? Is it big enough to reject H_0 if $\alpha = .01$?
2. With $\alpha = .05$, what value forms the boundary for the critical region in the distribution of *F*-ratios with $df = 2, 24$?

ANSWERS

1. For $\alpha = .05$, the critical value is 3.68 and you should reject H_0. For $\alpha = .01$, the critical value is 6.36 and you should fail to reject H_0.
2. The critical value is 3.40.

12.5 EXAMPLES OF HYPOTHESIS TESTING AND EFFECT SIZE WITH ANOVA

Although we have now seen all the individual components of ANOVA, the following example demonstrates the complete ANOVA process using the standard four-step procedure for hypothesis testing.

EXAMPLE 12.1

The data in Table 12.4 were obtained from an independent-measures experiment designed to examine people's preferences for viewing distance of a 42-inch, high-definition television. Four viewing distances were evaluated, 9 feet, 12 feet, 15 feet, and 18 feet, with a separate group of participants tested at each distance. Each individual watched a 30-minute television program from a specific distance and then completed a brief questionnaire measuring their satisfaction with the experience. One question asked them to rate the viewing distance on a scale from 1 (Very Bad—definitely need to move closer or farther away) to 7 (Excellent—perfect viewing distance). The purpose of the ANOVA is to determine whether there are any significant differences among the four viewing distances that were tested.

Before we begin the hypothesis test, note that we have already computed several summary statistics for the data in Table 12.4. Specifically, the treatment totals (T) and SS values are shown for each sample, and the grand total (G) as well as N and ΣX^2 are shown for the entire set of data. Having these summary values simplifies the

TABLE 12.4

Satisfaction with different viewing distances of a 42-inch high-definition television.

9 feet	12 feet	15 feet	18 feet	
3	4	7	6	$N = 20$
0	3	6	3	$G = 60$
2	1	5	4	$\Sigma X^2 = 262$
0	1	4	3	
0	1	3	4	
$T = 5$	$T = 10$	$T = 25$	$T = 20$	
$SS = 8$	$SS = 8$	$SS = 10$	$SS = 6$	

computations in the hypothesis test, and we suggest that you always compute these summary statistics before you begin an ANOVA.

STEP 1: State the hypotheses and select an alpha level.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (There is no treatment effect.)

H_1: At least one of the treatment means is different.

We use $\alpha = .05$.

STEP 2: Locate the critical region.

We first must determine degrees of freedom for $MS_{\text{between treatments}}$ and $MS_{\text{within treatments}}$ (the numerator and denominator of the F-ratio), so we begin by analyzing the degrees of freedom. For these data, the total degrees of freedom are

Often it is easier to postpone finding the critical region until after step 3, where you compute the *df* values as part of the calculations for the F-ratio.

$$df_{\text{total}} = N - 1$$
$$= 20 - 1$$
$$= 19$$

Analyzing this total into two components, we obtain

$$df_{\text{between}} = k - 1 = 4 - 1 = 3$$
$$df_{\text{within}} = \Sigma df_{\text{inside each treatment}} = 4 + 4 + 4 + 4 = 16$$

The F-ratio for these data has $df = 3, 16$. The distribution of all the possible F-ratios with $df = 3, 16$ is presented in Figure 12.9. Note that F-ratios larger than 3.24 are extremely rare ($p < .05$) if H_0 is true and, therefore, form the critical region for the test.

STEP 3: Compute the F-ratio.

The series of calculations for computing F is presented in Figure 12.5 and can be summarized as follows:

a. Analyze the *SS* to obtain SS_{between} and SS_{within}.

b. Use the *SS* values and the *df* values (from step 2) to calculate the two variances, MS_{between} and MS_{within}.

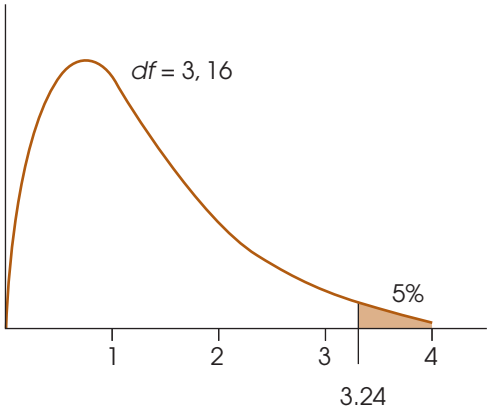
c. Finally, use the two *MS* values (variances) to compute the F-ratio.

Analysis of SS. First, we compute the total *SS* and then the two components, as indicated in Figure 12.6.

SS_{total} is simply the *SS* for the total set of $N = 20$ scores.

FIGURE 12.9

The distribution of F-ratios with $df = 3, 16$. The critical value for $\alpha = .05$ is $F = 3.24$.



$$SS_{\text{total}} = \Sigma X^2 - \frac{G^2}{N}$$
$$= 262 - \frac{60^2}{20}$$
$$= 262 - 180$$
$$= 82$$

SS_{within} combines the SS values from inside each of the treatment conditions.

$$SS_{\text{within}} = \Sigma SS_{\text{inside each treatment}} = 8 + 8 + 10 + 6 = 32$$

SS_{between} measures the differences among the four treatment means (or treatment totals). Because we have already calculated SS_{total} and SS_{within}, the simplest way to obtain SS_{between} is by subtraction (Equation 12.5).

$$SS_{\text{between}} = SS_{\text{total}} - SS_{\text{within}}$$
$$= 82 - 32$$
$$= 50$$

Calculation of mean squares. Because we already found $df_{\text{between}} = 3$ and $df_{\text{within}} = 16$ (Step 2), we now can compute the variance or MS value for each of the two components.

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{50}{3} = 16.67$$
$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{32}{16} = 2.00$$

Calculation of F. We compute the F-ratio:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{16.67}{2.00} = 8.33$$

STEP 4: Make a decision.

The F value we obtained, $F = 8.33$, is in the critical region (see Figure 12.9). It is very unlikely ($p < .05$) that we would obtain a value this large if H_0 is true. Therefore, we reject H_0 and conclude that there is a significant treatment effect.

Example 12.1 demonstrated the complete, step-by-step application of the ANOVA procedure. There are two additional points that can be made using this example.

First, you should look carefully at the statistical decision. We have rejected H_0 and concluded that not all the treatments are the same. But we have not determined which ones are different. Is a 9-foot distance different from 12 feet? Is 12 feet different from 15 feet? Unfortunately, these questions remain unanswered. We do know that at least one difference exists (we rejected H_0), but additional analysis is necessary to find out exactly where this difference is. We address this problem in Section 12.6.

Second, as noted earlier, all of the components of the analysis (the *SS*, *df*, *MS*, and *F*) can be presented together in one summary table. The summary table for the analysis in Example 12.1 is as follows:

Source	*SS*	*df*	*MS*	
Between treatments	50	3	16.67	$F = 8.33$
Within treatments	32	16	2.00	
Total	82	19		

Although these tables are very useful for organizing the components of an ANOVA, they are not commonly used in published reports. The current method for reporting the results from an ANOVA is presented on page 409.

MEASURING EFFECT SIZE FOR ANOVA

As we noted previously, a *significant* mean difference simply indicates that the difference observed in the sample data is very unlikely to have occurred just by chance. Thus, the term significant does not necessarily mean *large,* it simply means larger than expected by chance. To provide an indication of how large the effect actually is, researchers should report a measure of effect size in addition to the measure of significance.

For ANOVA, the simplest and most direct way to measure effect size is to compute the percentage of variance accounted for by the treatment conditions. Like the r^2 value used to measure effect size for the t tests in Chapters 9, 10, and 11, this percentage measures how much of the variability in the scores is accounted for by the differences between treatments. For ANOVA, the calculation and the concept of the percentage of variance is extremely straightforward. Specifically, we determine how much of the total *SS* is accounted for by the $SS_{\text{between treatments}}$.

$$\text{The percentage of variance accounted for} = \frac{SS_{\text{between treatments}}}{SS_{\text{total}}} \tag{12.14}$$

For the data in Example 12.1, the percentage of variance accounted for $= \frac{50}{82} = 0.61$ (or 61%).

In published reports of ANOVA results, the percentage of variance accounted for by the treatment effect is usually called η^2 (the Greek letter *eta squared*) instead of using r^2. Thus, for the study in Example 12.1, $\eta^2 = 0.61$.

IN THE LITERATURE

REPORTING THE RESULTS OF ANOVA

The APA format for reporting the results of ANOVA begins with a presentation of the treatment means and standard deviations in the narrative of the article, a table, or a graph. These descriptive statistics are not part of the calculations for the ANOVA, but you can easily determine the treatment means from n and T ($M = T/n$) and the standard deviations from the SS and $n-1$ values for each treatment. Next, report the results of the ANOVA. For the study described in Example 12.1, the report might state the following:

> The means and standard deviations are presented in Table 1. The analysis of variance indicates that there are significant differences among the four viewing distances, $F(3, 16) = 8.33$, $p < .05$, $\eta^2 = 0.61$.
>
> **TABLE 1**
>
> Ratings of satisfaction with different television viewing distances.

	9 Feet	12 Feet	15 Feet	18 Feet
M	1.00	2.00	5.00	4.00
SD	1.41	1.41	1.58	1.22

Note how the F-ratio is reported. In this example, degrees of freedom for between and within treatments are $df = 3, 16$, respectively. These values are placed in parentheses immediately following the symbol F. Next, the calculated value for F is reported, followed by the probability of committing a Type I error (the alpha level) and the measure of effect size.

When an ANOVA is done using a computer program, the F-ratio is usually accompanied by an exact value for p. The data from Example 12.1 were analyzed using the SPSS program (see Resources at the end of this chapter) and the computer output included a significance level of $p = .001$. Using the exact p value from the computer output, the research report would conclude, "The analysis of variance revealed significant differences among the four viewing distances, $F(3, 16) = 8.33$, $p = .001$, $\eta^2 = 0.61$."

A CONCEPTUAL VIEW OF ANOVA

Because ANOVA requires relatively complex calculations, students encountering this statistical technique for the first time often tend to be overwhelmed by the formulas and arithmetic and lose sight of the general purpose for the analysis. The following two examples are intended to minimize the role of the formulas and shift attention back to the conceptual goal of the ANOVA process.

EXAMPLE 12.2

The following data represent the outcome of an experiment using two separate samples to evaluate the mean difference between two treatment conditions. Take a minute to look at the data and, without doing any calculations, try to predict the outcome of an ANOVA for these values. Specifically, predict what values should be obtained for the between-treatments variance (*MS*) and the *F*-ratio. If you do not "see" the answer after 20 or 30 seconds, try reading the hints that follow the data.

Treatment I	Treatment II	
4	2	$N = 8$
0	1	$G = 16$
1	0	$\Sigma X^2 = 56$
3	5	
$T = 8$	$T = 8$	
$SS = 10$	$SS = 14$	

If you are having trouble predicting the outcome of the ANOVA, read the following hints, and then go back and look at the data.

Hint 1: Remember: $SS_{between}$ and $MS_{between}$ provide a measure of how much difference there is between treatment conditions.

Hint 2: Find the mean or total (T) for each treatment, and determine how much difference there is between the two treatments.

You should realize by now that the data have been constructed so that there is zero difference between treatments. The two sample means (and totals) are identical, so $SS_{between} = 0$, $MS_{between} = 0$, and the *F*-ratio is zero.

Conceptually, the numerator of the *F*-ratio always measures how much difference exists between treatments. In Example 12.2, we constructed an extreme set of scores with zero difference. However, you should be able to look at any set of data and quickly compare the means (or totals) to determine whether there are big differences or small differences between treatments.

Being able to estimate the magnitude of between-treatment differences is a good first step in understanding ANOVA and should help you to predict the outcome of an ANOVA. However, the *between-treatment* differences are only one part of the analysis. You must also understand the *within-treatment* differences that form the denominator of the *F*-ratio. The following example is intended to demonstrate the concepts underlying SS_{within} and MS_{within}. In addition, the example should give you a better understanding of how the between-treatment differences and the within-treatment differences act together within the ANOVA.

EXAMPLE 12.3

The purpose of this example is to present a visual image for the concepts of between-treatments variability and within-treatments variability. In this example, we compare two hypothetical outcomes for the same experiment. In each case, the experiment uses two separate samples to evaluate the mean difference between two treatments. The following data represent the two outcomes, which we call experiment A and experiment B.

Experiment A		Experiment B	
Treatment		Treatment	
I	II	I	II
8	12	4	12
8	13	11	9
7	12	2	20
9	11	17	6
8	13	0	16
9	12	8	18
7	11	14	3
$M = 8$	$M = 12$	$M = 8$	$M = 12$
$s = 0.82$	$s = 0.82$	$s = 6.35$	$s = 6.35$

The data from experiment A are displayed in a frequency distribution graph in Figure 12.10(a). Notice that there is a 4-point difference between the treatment means ($M_1 = 8$ and $M_2 = 12$). This is the *between-treatments* difference that contributes to the numerator of the *F*-ratio. Also notice that the scores in each treatment are clustered closely around the mean, indicating that the variance inside each treatment is relatively small. This is the *within-treatments* variance that contributes to the denominator of the *F*-ratio. Finally, you should realize that it is easy to see the mean difference between the two samples. The fact that there is a clear mean difference between the two treatments is confirmed by computing the *F*-ratio for experiment A.

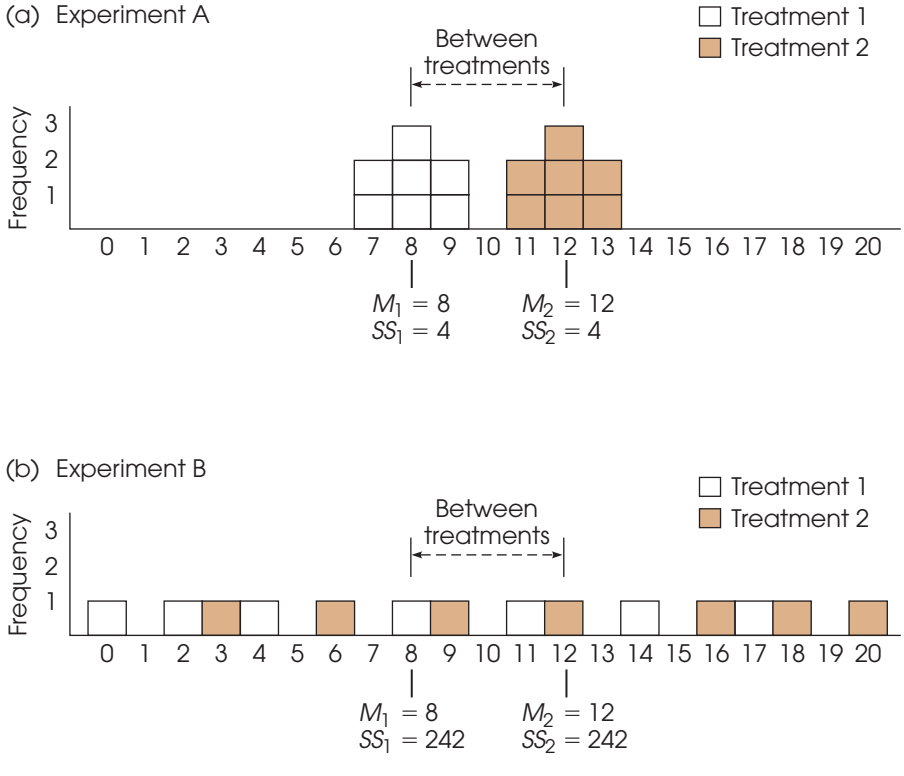


FIGURE 12.10

A visual representation of the between-treatments variability and the within-treatments variability that form the numerator and denominator, respectively, of the *F*-ratio. In (a), the difference between treatments is relatively large and easy to see. In (b), the same 4-point difference between treatments is relatively small and is overwhelmed by the within-treatments variability.

$$F = \frac{\text{between-treatments difference}}{\text{within-treatments differences}} = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{56}{0.667} = 83.96$$

An F-ratio of $F = 83.96$ is sufficient to reject the null hypothesis, so we conclude that there is a significant difference between the two treatments.

Now consider the data from experiment B, which are shown in Figure 12.10(b) and present a very different picture. This experiment has the same 4-point difference between treatment means that we found in experiment A ($M_1 = 8$ and $M_2 = 12$). However, for these data the scores in each treatment are scattered across the entire scale, indicating relatively large variance inside each treatment. In this case, the large variance within treatments overwhelms the relatively small mean difference between treatments. In the figure it is almost impossible to see the mean difference between treatments. For these data, the F-ratio confirms that there is no clear mean difference between treatments.

$$F = \frac{\text{between-treatments difference}}{\text{within-treatments differences}} = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{56}{40.33} = 1.39$$

For experiment B, the F-ratio is not large enough to reject the null hypothesis, so we conclude that there is no significant difference between the two treatments. Once again, the statistical conclusion is consistent with the appearance of the data in Figure 12.10(b). Looking at the figure, we see that the scores from the two samples appear to be intermixed randomly with no clear distinction between treatments.

As a final point, note that the denominator of the F-ratio, MS_{within}, is a measure of the variability (or variance) within each of the separate samples. As we have noted in previous chapters, high variability makes it difficult to see any patterns in the data. In Figure 12.10(a), the 4-point mean difference between treatments is easy to see because the sample variability is small. In Figure 12.10(b), the 4-point difference gets lost because the sample variability is large. In general, you can think of variance as measuring the amount of "noise" or "confusion" in the data. With large variance, there is a lot of noise and confusion and it is difficult to see any clear patterns.

Although Examples 12.2 and 12.3 present somewhat simplified demonstrations with exaggerated data, the general point of the examples is to help you *see* what happens when you perform an ANOVA. Specifically:

1. The numerator of the F-ratio (MS_{between}) measures how much difference exists between the treatment means. The bigger the mean differences, the bigger the F-ratio.
2. The denominator of the F-ratio (MS_{within}) measures the variance of the scores inside each treatment; that is, the variance for each of the separate samples. In general, larger sample variance produces a smaller F-ratio.

We should note that the number of scores in the samples also can influence the outcome of an ANOVA. As with most other hypothesis tests, if other factors are held constant, increasing the sample size tends to increase the likelihood of rejecting the null hypothesis. However, changes in sample size have little or no effect on measures of effect size such as η^2.

Finally, we should note that the problems associated with high variance often can be minimized by transforming the original scores to ranks and then conducting an

alternative statistical analysis known as the *Kruskal-Wallis test,* which is designed specifically for ordinal data. The Kruskal-Wallis test is presented in Appendix E, which also discusses the general purpose and process of converting numerical scores into ranks. The Kruskal-Wallis test also can be used if the data violate one of the assumptions for the independent-measures *ANOVA*, which are outlined at the end of section 12.7.

MS_{within} AND POOLED VARIANCE

You may have recognized that the two research outcomes presented in Example 12.3 are similar to those presented earlier in Example 10.5 in Chapter 10. Both examples are intended to demonstrate the role of variance in a hypothesis test. Both examples show that large values for sample variance can obscure any patterns in the data and reduce the potential for finding significant differences between means.

For the independent-measures t statistic in Chapter 10, the sample variance contributed directly to the standard error in the bottom of the t formula. Now, the sample variance contributes directly to the value of MS_{within} in the bottom of the F-ratio. In the t-statistic and in the F-ratio the variances from the separate samples are pooled together to create one average value for sample variance. For the independent-measures t statistic, we pooled two samples together to compute

$$\text{pooled variance} = s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Now, in ANOVA, we are combining two or more samples to calculate

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{\Sigma SS}{\Sigma df} = \frac{SS_1 + SS_2 + SS_3 + \cdots}{df_1 + df_2 + df_3 + \cdots}$$

Notice that the concept of pooled variance is the same whether you have exactly two samples or more than two samples. In either case, you simply add the SS values and divide by the sum of the df values. The result is an average of all of the different sample variances.

AN EXAMPLE WITH UNEQUAL SAMPLE SIZES

In the previous examples, all of the samples were exactly the same size (equal ns). However, the formulas for ANOVA can be used when the sample size varies within an experiment. You also should note, however, that the general ANOVA procedure is most accurate when used to examine experimental data with equal sample sizes. Therefore, researchers generally try to plan experiments with equal ns. However, there are circumstances in which it is impossible or impractical to have an equal number of subjects in every treatment condition. In these situations, ANOVA still provides a valid test, especially when the samples are relatively large and when the discrepancy between sample sizes is not extreme.

The following example demonstrates an ANOVA with samples of different sizes.

EXAMPLE 12.4

A researcher is interested in the amount of homework required by different academic majors. Students are recruited from Biology, English, and Psychology to participant in the study. The researcher randomly selects one course that each

student is currently taking and asks the student to record the amount of out-of-class work required each week for the course. The researcher used all of the volunteer participants, which resulted in unequal sample sizes. The data are summarized in Table 12.5.

STEP 1: State the hypotheses, and select the alpha level.

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: At least one population is different.

$\alpha = .05$

STEP 2: Locate the critical region.

To find the critical region, we first must determine the *df* values for the *F*-ratio:

$$df_{total} = N - 1 = 20 - 1 = 19$$

$$df_{between} = k - 1 = 3 - 1 = 2$$

$$df_{within} = N - k = 20 - 3 = 17$$

The *F*-ratio for these data has $df = 2, 17$. With $\alpha = .05$, the critical value for the *F*-ratio is 3.59.

STEP 3: Compute the *F*-ratio.

First, compute the three *SS* values. As usual, SS_{total} is the *SS* for the total set of $N = 20$ scores, and SS_{within} combines the *SS* values from inside each of the treatment conditions.

$$SS_{total} = \Sigma X^2 - \frac{G^2}{N}$$

$$= 3377 - \frac{250^2}{20}$$

$$= 3377 - 3125$$

$$= 252$$

$$SS_{within} = \Sigma SS_{\text{inside each treatment}}$$

$$= 37 + 90 + 60$$

$$= 187$$

$SS_{between}$ can be found by subtraction (Equation 12.5).

$$SS_{between} = SS_{total} - SS_{within}$$

$$= 252 - 187$$

$$= 65$$

TABLE 12.5

Average hours of homework per week for one course for students in three academic majors.

Biology	English	Psychology	
$n = 4$	$n = 10$	$n = 6$	$N = 20$
$M = 9$	$M = 13$	$M = 14$	$G = 250$
$T = 36$	$T = 130$	$T = 84$	$\Sigma X^2 = 3377$
$SS = 37$	$SS = 90$	$SS = 60$	

Or, $SS_{between}$ can be calculated using the computation formula (Equation 12.7). If you use the computational formula, be careful to match each treatment total (T) with the appropriate sample size (n) as follows:

$$\begin{aligned} SS_{between} &= \Sigma\frac{T^2}{n} - \frac{G^2}{N} \\ &= \frac{36^2}{4} + \frac{130^2}{10} + \frac{84^2}{6} - \frac{250^2}{20} \\ &= 324 + 1690 + 1176 - 3125 \\ &= 65 \end{aligned}$$

Finally, compute the MS values and the F-ratio:

$$MS_{between} = \frac{SS}{df} = \frac{65}{2} = 32.5$$

$$MS_{within} = \frac{SS}{df} = \frac{187}{17} = 11$$

$$F = \frac{MS_{between}}{MS_{within}} = \frac{32.5}{11} = 2.95$$

STEP 4: Make a decision.

Because the obtained F-ratio is not in the critical region, we fail to reject the null hypothesis and conclude that there are no significant differences among the three populations of students in terms of the average amount of homework each week.

LEARNING CHECK

1. A researcher used ANOVA and computed $F = 4.25$ for the following data.

Treatments		
I	II	III
$n = 10$	$n = 10$	$n = 10$
$M = 20$	$M = 28$	$M = 35$
$SS = 1005$	$SS = 1391$	$SS = 1180$

a. If the mean for treatment III were changed to $M = 25$, what would happen to the size of the F-ratio (increase or decrease)? Explain your answer.

b. If the SS for treatment I were changed to $SS = 1400$, what would happen to the size of the F-ratio (increase or decrease)? Explain your answer.

2. A research study comparing three treatment conditions produces $T = 20$ with $n = 4$ for the first treatment, $T = 10$ with $n = 5$ for the second treatment, and $T = 30$ with $n = 6$ for the third treatment. Calculate $SS_{between\ treatments}$ for these data.

ANSWERS

1. a. If the mean for treatment III were changed to $M = 25$, it would reduce the size of the mean differences (the three means would be closer together). This would reduce the size of $MS_{between}$ and would reduce the size of the F-ratio.

b. If the SS in treatment I were increased to $SS = 1400$, it would increase the size of the variability within treatments. This would increase MS_{within} and would reduce the size of the F-ratio.

2. With $G = 60$ and $N = 15$, $SS_{between} = 30$.

12.6 POST HOC TESTS

As noted earlier, the primary advantage of ANOVA (compared to *t* tests) is that it allows researchers to test for significant mean differences when there are *more than two* treatment conditions. ANOVA accomplishes this feat by comparing all the individual mean differences simultaneously within a single test. Unfortunately, the process of combining several mean differences into a single test statistic creates some difficulty when it is time to interpret the outcome of the test. Specifically, when you obtain a significant *F*-ratio (reject H_0), it simply indicates that somewhere among the entire set of mean differences there is at least one that is statistically significant. In other words, the overall *F*-ratio only tells you that a significant difference exists; it does not tell exactly which means are significantly different and which are not.

Consider, for example, a research study that uses three samples to compare three treatment conditions. Suppose that the three sample means are $M_1 = 3$, $M_2 = 5$, and $M_3 = 10$. In this hypothetical study there are three mean differences:

1. There is a 2-point difference between M_1 and M_2.
2. There is a 5-point difference between M_2 and M_3.
3. There is a 7-point difference between M_1 and M_3.

If an ANOVA were used to evaluate these data, a significant *F*-ratio would indicate that at least one of the sample mean differences is large enough to satisfy the criterion of statistical significance. In this example, the 7-point difference is the biggest of the three and, therefore, it must indicate a significant difference between the first treatment and the third treatment ($\mu_1 \neq \mu_3$). But what about the 5-point difference? Is it also large enough to be significant? And what about the 2-point difference between M_1 and M_2? Is it also significant? The purpose of *post hoc tests* is to answer these questions.

DEFINITION

Post hoc tests (or **posttests**) are additional hypothesis tests that are done after an ANOVA to determine exactly which mean differences are significant and which are not.

As the name implies, post hoc tests are done after an ANOVA. More specifically, these tests are done after ANOVA when

1. You reject H_0 and
2. There are three or more treatments ($k \geq 3$).

Rejecting H_0 indicates that at least one difference exists among the treatments. If there are only two treatments, then there is no question about which means are different and, therefore, no need for posttests. However, with three or more treatments ($k \geq 3$), the problem is to determine exactly which means are significantly different.

POSTTESTS AND TYPE I ERRORS

In general, a post hoc test enables you to go back through the data and compare the individual treatments two at a time. In statistical terms, this is called making *pairwise comparisons*. For example, with $k = 3$, we would compare μ_1 versus μ_2, then μ_2 versus μ_3, and then μ_1 versus μ_3. In each case, we are looking for a significant mean difference. The process of conducting pairwise comparisons involves performing a series of separate hypothesis tests, and each of these tests includes the risk of a Type I error. As you do more and more separate tests, the risk of a Type I error accumulates and is called the *experimentwise alpha level* (see p. 391).

We have seen, for example, that a research study with three treatment conditions produces three separate mean differences, each of which could be evaluated using a post hoc test. If each test uses $\alpha = .05$, then there is a 5% risk of a Type I error for the first posttest, another 5% risk for the second test, and one more 5% risk for the third test. Although the probability of error is not simply the sum across the three tests, it should be clear that increasing the number of separate tests definitely increases the total, experimentwise probability of a Type I error.

Whenever you are conducting posttests, you must be concerned about the experimentwise alpha level. Statisticians have worked with this problem and have developed several methods for trying to control Type I errors in the context of post hoc tests. We consider two alternatives.

TUKEY'S HONESTLY SIGNIFICANT DIFFERENCE (HSD) TEST

The first post hoc test we consider is *Tukey's HSD test.* We selected Tukey's HSD test because it is a commonly used test in psychological research. Tukey's test allows you to compute a single value that determines the minimum difference between treatment means that is necessary for significance. This value, called the *honestly significant difference,* or HSD, is then used to compare any two treatment conditions. If the mean difference exceeds Tukey's HSD, then you conclude that there is a significant difference between the treatments. Otherwise, you cannot conclude that the treatments are significantly different. The formula for Tukey's HSD is

$$HSD = q\sqrt{\frac{MS_{\text{within}}}{n}} \tag{12.15}$$

The q value used in Tukey's HSD test is called a Studentized range statistic.

where the value of q is found in Table B.5 (Appendix B, p. 708), MS_{within} is the within-treatments variance from the ANOVA, and n is the number of scores in each treatment. Tukey's test requires that the sample size, n, be the same for all treatments. To locate the appropriate value of q, you must know the number of treatments in the overall experiment (k), the degrees of freedom for MS_{within} (the error term in the F-ratio), and you must select an alpha level (generally the same α used for the ANOVA).

EXAMPLE 12.5

To demonstrate the procedure for conducting post hoc tests with Tukey's HSD, we use the hypothetical data shown in Table 12.6. The data represent the results of a study comparing scores in three different treatment conditions. Note that the table displays summary statistics for each sample and the results from the overall ANOVA. With $k = 3$ treatments, $df_{\text{within}} = 24$, and $\alpha = .05$, you should find that the value of q for the test is $q = 3.53$ (see Table B.5). Therefore, Tukey's HSD is

$$HSD = q\sqrt{\frac{MS_{\text{within}}}{n}} = 3.53\sqrt{\frac{4.00}{9}} = 2.36$$

TABLE 12.6

Hypothetical results from a research study comparing three treatment conditions. Summary statistics are presented for each treatment along with the outcome from the ANOVA.

Treatment A	Treatment B	Treatment C
$n = 9$	$n = 9$	$n = 9$
$T = 27$	$T = 49$	$T = 63$
$M = 3.00$	$M = 5.44$	$M = 7.00$

Source	*SS*	*df*	*MS*
Between	73.19	2	36.60
Within	96.00	24	4.00
Total	169.19	26	
Overall $F(2, 24) = 9.15$			

Thus, the mean difference between any two samples must be at least 2.36 to be significant. Using this value, we can make the following conclusions:

1. Treatment A is significantly different from treatment B ($M_A - M_B = 2.44$).
2. Treatment A is also significantly different from treatment C ($M_A - M_C = 4.00$).
3. Treatment B is not significantly different from treatment C ($M_B - M_C = 1.56$).

THE SCHEFFÉ TEST

Because it uses an extremely cautious method for reducing the risk of a Type I error, the *Scheffé test* has the distinction of being one of the safest of all possible post hoc tests (smallest risk of a Type I error). The Scheffé test uses an *F*-ratio to evaluate the significance of the difference between any two treatment conditions. The numerator of the *F*-ratio is an $MS_{between}$ that is calculated using *only the two treatments you want to compare*. The denominator is the same MS_{within} that was used for the overall ANOVA. The "safety factor" for the Scheffé test comes from the following two considerations:

1. Although you are comparing only two treatments, the Scheffé test uses the value of k from the original experiment to compute *df* between treatments. Thus, *df* for the numerator of the *F*-ratio is $k - 1$.
2. The critical value for the Scheffé *F*-ratio is the same as was used to evaluate the *F*-ratio from the overall ANOVA. Thus, Scheffé requires that every posttest satisfy the same criterion that was used for the complete ANOVA. The following example uses the data from Table 12.6 to demonstrate the Scheffé posttest procedure.

EXAMPLE 12.6

Remember that the Scheffé procedure requires a separate $SS_{between}$, $MS_{between}$, and *F*-ratio for each comparison being made. Although Scheffé computes $SS_{between}$ using the regular computational formula (Equation 12.7), you must remember that all of the numbers in the formula are entirely determined by the two treatment conditions being compared. We begin by comparing treatment A (with $T = 27$ and $n = 9$) and treatment B (with $T = 49$ and $n = 9$). The first step is to compute $SS_{between}$ for these two groups. In the formula for *SS,* notice that the grand total for the two groups is $G = 27 + 49 = 76$, and the total number of scores for the two groups is $N = 9 + 9 = 18$.

$$SS_{between} = \Sigma\frac{T^2}{n} - \frac{G^2}{N}$$
$$= \frac{27^2}{9} + \frac{49^2}{9} - \frac{76^2}{18}$$
$$= 81 + 266.78 - 320.89$$
$$= 26.89$$

Although we are comparing only two groups, these two were selected from a study consisting of $k = 3$ samples. The Scheffé test uses the overall study to determine the degrees of freedom between treatments. Therefore, $df_{between} = 3 - 1 = 2$, and the $MS_{between}$ is

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{26.89}{2} = 13.45$$

Finally, the Scheffé procedure uses the error term from the overall ANOVA to compute the F-ratio. In this case, $MS_{within} = 4.00$ with $df_{within} = 24$. Thus, the Scheffé test produces an F-ratio of

$$F_{\text{A verus B}} = \frac{MS_{between}}{MS_{within}} = \frac{13.45}{4.00} = 3.36$$

With $df = 2, 24$ and $\alpha = .05$, the critical value for F is 3.40 (see Table B.4). Therefore, our obtained F-ratio is not in the critical region, and we must conclude that these data show no significant difference between treatment A and treatment B.

The second comparison involves treatment B ($T = 49$) and treatment C ($T = 63$). This time the data produce $SS_{between} = 10.89$, $MS_{between} = 5.45$, and $F(2, 24) = 1.36$ (check the calculations for yourself). Once again the critical value for F is 3.40, so we must conclude that the data show no significant difference between treatment B and treatment C.

The final comparison is treatment A ($T = 27$) and treatment C ($T = 63$). This time the data produce $SS_{between} = 72$, $MS_{between} = 36$, and $F(2, 24) = 9.00$ (check the calculations for yourself). Once again the critical value for F is 3.40, and this time we conclude that the data show a significant difference.

Thus, the Scheffé posttest indicates that the only significant difference is between treatment A and treatment C.

There are two interesting points to be made from the posttest outcomes presented in the preceding two examples. First, the Scheffé test was introduced as being one of the safest of the posttest techniques because it provides the greatest protection from Type I errors. To provide this protection, the Scheffé test simply requires a larger difference between sample means before you may conclude that the difference is significant. For example, using Tukey's test in Example 12.5, we found that the difference between treatment A and treatment B was large enough to be significant. However, this same difference failed to reach significance according to the Scheffé test (Example 12.6). The discrepancy between the results is an example of the Scheffé test's extra demands: The Scheffé test simply requires more evidence and, therefore, it is less likely to lead to a Type I error.

The second point concerns the pattern of results from the three Scheffé tests in Example 12.6. You may have noticed that the posttests produce what are apparently contradictory results. Specifically, the tests show no significant difference between A and B and they show no significant difference between B and C. This combination of outcomes might lead you to suspect that there is no significant difference between A and C. However, the test did show a significant difference. The answer to this apparent contradiction lies in the criterion of statistical significance. The differences between A and B and between B and C are too small to satisfy the criterion of significance. However, when these differences are combined, the total difference between A and C is large enough to meet the criterion for significance.

LEARNING CHECK

1. With $k = 2$ treatments, are post hoc tests necessary when the null hypothesis is rejected? Explain why or why not.
2. An ANOVA comparing three treatments produces an overall F-ratio with $df = 2, 27$. If the Scheffé test was used to compare two of the three treatments, then the Scheffé F-ratio would also have $df = 2, 27$. (True or false?)

3. Using the data and the results from Example 12.1,

a. Use Tukey's HSD test to determine whether there is a significant mean difference between a 12-foot and a 15-foot distance. Use $\alpha = .05$.

b. Use the Scheffé test to determine whether there is a significant mean difference between 12 feet and 15 feet. Use $\alpha = .05$.

ANSWERS

1. No. Post hoc tests are used to determine which treatments are different. With only two treatment conditions, there is no uncertainty as to which two treatments are different.

2. True

3. a. For this test, $q = 4.05$ and HSD = 2.55. There is a 3-point mean difference between 12 feet and 15 feet, which is large enough to be significant.

b. The Scheffé $F = 3.75$, which is greater than the critical value of 3.24. Conclude that the mean difference between 12 feet and 15 feet is significant.

12.7 THE RELATIONSHIP BETWEEN ANOVA AND *t* TESTS

When you are evaluating the mean difference from an independent-measures study comparing only two treatments (two separate samples), you can use either an independent-measures *t* test (Chapter 10) or the ANOVA presented in this chapter. In practical terms, it makes no difference which you choose. These two statistical techniques always result in the same statistical decision. In fact the two methods use many of the same calculations and are very closely related in several other respects. The basic relationship between *t* statistics and *F*-ratios can be stated in an equation:

$$F = t^2$$

This relationship can be explained by first looking at the structure of the formulas for *F* and *t*. The *t* statistic compares *distances*: the distance between two sample means (numerator) and the distance computed for the standard error (denominator). The *F*-ratio, on the other hand, compares *variances*. You should recall that variance is a measure of squared distance. Hence, the relationship: $F = t^2$.

There are several other points to consider in comparing the *t* statistic to the *F*-ratio.

1. It should be obvious that you are testing the same hypotheses whether you choose a *t* test or an ANOVA. With only two treatments, the hypotheses for either test are

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

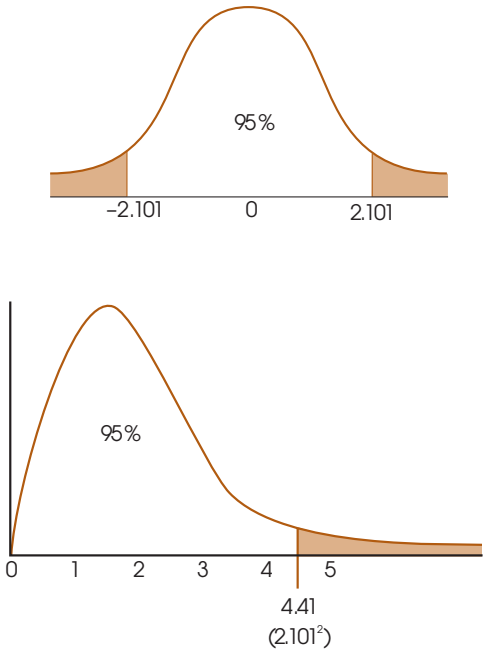
2. The degrees of freedom for the *t* statistic and the *df* for the denominator of the *F*-ratio (df_{within}) are identical. For example, if you have two samples, each with six scores, the independent-measures *t* statistic has $df = 10$, and the *F*-ratio has $df = 1, 10$. In each case, you are adding the *df* from the first sample $(n - 1)$ and the *df* from the second sample $(n - 1)$.

3. The distribution of *t* and the distribution of *F*-ratios match perfectly if you take into consideration the relationship $F = t^2$. Consider the *t* distribution with $df = 18$ and the corresponding *F* distribution with $df = 1, 18$ that are presented in Figure 12.11. Notice the following relationships:

a. If each of the *t* values is squared, then all of the negative values become positive. As a result, the whole left-hand side of the *t* distribution (below

FIGURE 12.11

The distribution of t statistics with $df = 18$ and the corresponding distribution of F-ratios with $df = 1, 18$. Notice that the critical values for $\alpha = .05$ are $t = \pm 2.101$ and that $F = 2.101^2 = 4.41$.



zero) is flipped over to the positive side. This creates an asymmetrical, positively skewed distribution—that is, the F distribution.

b. For $\alpha = .05$, the critical region for t is determined by values greater than $+2.101$ or less than -2.101. When these boundaries are squared, you get $\pm 2.101^2 = 4.41$

Notice that 4.41 is the critical value for $\alpha = .05$ in the F distribution. Any value that is in the critical region for t ends up in the critical region for F-ratios after it is squared.

ASSUMPTIONS FOR THE INDEPENDENT-MEASURES ANOVA

The independent-measures ANOVA requires the same three assumptions that were necessary for the independent-measures t hypothesis test:

1. The observations within each sample must be independent (see p. 254).
2. The populations from which the samples are selected must be normal.
3. The populations from which the samples are selected must have equal variances (homogeneity of variance).

Ordinarily, researchers are not overly concerned with the assumption of normality, especially when large samples are used, unless there are strong reasons to suspect that the assumption has not been satisfied. The assumption of homogeneity of variance is an important one. If a researcher suspects that it has been violated, it can be tested by Hartley's F-max test for homogeneity of variance (Chapter 10, p. 338).

Finally, if you suspect that one of the assumptions for the independent-measures ANOVA has been violated, you can still proceed by transforming the original scores into ranks and then using an alternative statistical analysis known as the Kruskal-Wallis test, which is designed specifically for ordinal data. The Kruskal-Wallis test is

presented in Appendix E. As noted earlier, the Kruskal-Wallis test also can be useful if large sample variance prevents the independent-measures ANOVA from producing a significant result.

LEARNING CHECK

1. A researcher uses an independent-measures t test to evaluate the mean difference obtained in a research study, and obtains a t statistic of $t = 3.00$. If the researcher had used an ANOVA to evaluate the results, the F-ratio would be $F = 9.00$. (True or false?)
2. An ANOVA produces an F-ratio with $df = 1, 34$. Could the data have been analyzed with a t test? What would be the degrees of freedom for the t statistic?

ANSWERS

1. True. $F = t^2$
2. If the F-ratio has $df = 1, 34$, then the experiment compared only two treatments, and you could use a t statistic to evaluate the data. The t statistic would have $df = 34$.

SUMMARY

1. Analysis of variance (ANOVA) is a statistical technique that is used to test the significance of mean differences among two or more treatment conditions. The null hypothesis for this test states that, in the general population, there are no mean differences among the treatments. The alternative states that at least one mean is different from another.
2. The test statistic for ANOVA is a ratio of two variances called an F-ratio. The variances in the F-ratio are called mean squares, or MS values. Each MS is computed by

$$MS = \frac{SS}{df}$$

3. For the independent-measures ANOVA, the F-ratio is

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

The MS_{between} measures differences between the treatments by computing the variability of the treatment means or totals. These differences are assumed to be produced by

   a. Treatment effects (if they exist)
   b. Random, unsystematic differences (chance)

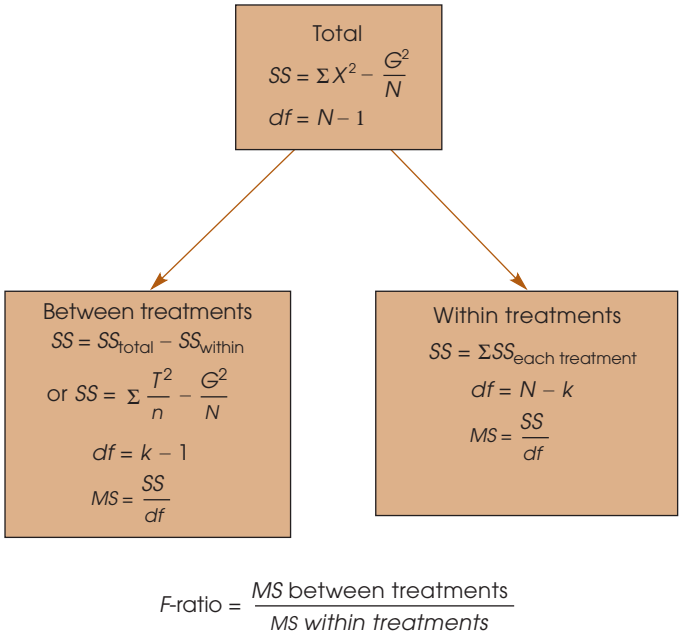
The MS_{within} measures variability inside each of the treatment conditions. Because individuals inside a treatment condition are all treated exactly the same, any differences within treatments cannot be caused by treatment effects. Thus, the within-treatments MS is produced only by random, unsystematic differences. With these factors in mind, the F-ratio has the following structure:

$$F = \frac{\text{treatment effect} + \text{differences due to chance}}{\text{differences due to chance}}$$

When there is no treatment effect (H_0 is true), the numerator and the denominator of the F-ratio are measuring the same variance, and the obtained ratio should be near 1.00. If there is a significant treatment effect, then the numerator of the ratio should be larger than the denominator, and the obtained F value should be much greater than 1.00.

4. The formulas for computing each SS, df, and MS value are presented in Figure 12.12, which also shows the general structure for the ANOVA.
5. The F-ratio has two values for degrees of freedom, one associated with the MS in the numerator and one associated with the MS in the denominator. These df values are used to find the critical value for the F-ratio in the F distribution table.
6. Effect size for the independent-measures ANOVA is measured by computing eta squared, the percentage of variance accounted for by the treatment effect.

FIGURE 12.12
Formulas for ANOVA.



$$\eta^2 = \frac{SS_{between}}{SS_{between} + SS_{within}} = \frac{SS_{between}}{SS_{total}}$$

7. When the decision from an ANOVA is to reject the null hypothesis and when the experiment contains more than two treatment conditions, it is necessary to continue the analysis with a post hoc test, such as Tukey's HSD test or the Scheffé test. The purpose of these tests is to determine exactly which treatments are significantly different and which are not.

KEY TERMS

analysis of variance (ANOVA) (386)
factor (388)
levels (388)
testwise alpha level (391)
experimentwise alpha level (391)
between-treatments variance (392)
treatment effect (393)
within-treatments variance (393)
F-ratio (394)
error term (394)
mean square (*MS*) (401)
ANOVA summary table (402)
distribution of *F*-ratios (403)
eta squared (η^2) (408)
Kruskal-Wallis test (413)
post hoc tests (416)
pairwise comparisons (416)
Tukey's HSD test (417)
Scheffé test (418)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter.
You can find a tutorial quiz and other learning exercises for Chapter 12 on the book companion website.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Single-Factor, Independent-Measures Analysis of Variance (ANOVA)** presented in this chapter.

Data Entry

1. The scores are entered in a *stacked format* in the data editor, which means that all of the scores from all of the different treatments are entered in a single column (VAR00001). Enter the scores for treatment #2 directly beneath the scores from treatment #1 with no gaps or extra spaces. Continue in the same column with the scores from treatment #3, and so on.
2. In the second column (VAR00002), enter a number to identify the treatment condition for each score. For example, enter a 1 beside each score from the first treatment, enter a 2 beside each score from the second treatment, and so on.

Data Analysis

1. Click **Analyze** on the tool bar, select **Compare Means,** and click on **One-Way ANOVA.**
2. Highlight the column label for the set of scores (VAR0001) in the left box and click the arrow to move it into the **Dependent List** box.
3. Highlight the label for the column containing the treatment numbers (VAR0002) in the left box and click the arrow to move it into the **Factor** box.
4. If you want descriptive statistics for each treatment, click on the **Options** box, select **Descriptives,** and click **Continue.**
5. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the television viewing study in Example 12.1 and the program output is shown in Figure 12.13. The output begins with a table showing descriptive statistics (number of scores, mean, standard deviation, standard error for the mean, a 95% confidence interval for the mean, maximum and minimum scores) for each sample. The second part of the output presents a summary table showing the results from the ANOVA.

FOCUS ON PROBLEM SOLVING

1. It can be helpful to compute all three *SS* values separately, then check to verify that the two components (between and within) add up to the total. However, you can

Descriptives

VAR00001

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1.00	5	1.0000	1.41421	.63246	−.7560	2.7560	.00	3.00
2.00	5	2.0000	1.41421	.63246	.2440	3.7560	1.00	4.00
3.00	5	5.0000	1.58114	.70711	3.0368	6.9632	3.00	7.00
4.00	5	4.0000	1.22474	.54772	2.4793	5.5207	3.00	6.00
Total	20	3.0000	2.07745	.46453	2.0277	3.9723	.00	7.00

ANOVA

VAR00001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	50.000	3	16.667	8.333	.001
Within Groups	32.000	16	2.000		
Total	82.000	19			

FIGURE 12.13

SPSS output of the ANOVA for the television-viewing distance study in Example 12.1.

greatly simplify the calculations if you simply find SS_{total} and SS_{within}, then obtain $SS_{between}$ by subtraction.

2. Remember that an F-ratio has two separate values for df: a value for the numerator and one for the denominator. Properly reported, the $df_{between}$ value is stated first. You will need both df values when consulting the F distribution table for the critical F value. You should recognize immediately that an error has been made if you see an F-ratio reported with a single value for df.

3. When you encounter an F-ratio and its df values reported in the literature, you should be able to reconstruct much of the original experiment. For example, if you see "$F(2, 36) = 4.80$," you should realize that the experiment compared $k = 3$ treatment groups (because $df_{between} = k - 1 = 2$), with a total of $N = 39$ subjects participating in the experiment (because $df_{within} = N - k = 36$).

DEMONSTRATION 12.1

ANALYSIS OF VARIANCE

A human-factors psychologist studied three computer keyboard designs. Three samples of individuals were given material to type on a particular keyboard, and the number of errors committed by each participant was recorded. The data are as follows:

Keyboard A	Keyboard B	Keyboard C	
0	6	6	$N = 15$
4	8	5	$G = 60$
0	5	9	$\Sigma X^2 = 356$
1	4	4	
0	2	6	
$T = 5$	$T = 25$	$T = 30$	
$SS = 12$	$SS = 20$	$SS = 14$	

Are these data sufficient to conclude that there are significant differences in typing performance among the three keyboard designs?

STEP 1 **State the hypotheses, and specify the alpha level.** The null hypothesis states that there is no difference among the keyboards in terms of number of errors committed. In symbols, we would state

$$H_0: \mu_1 = \mu_2 = \mu_3 \qquad \text{(Type of keyboard used has no effect.)}$$

As noted previously in this chapter, there are a number of possible statements for the alternative hypothesis. Here we state the general alternative hypothesis:

$$H_1: \quad \text{At least one of the treatment means is different.}$$

We set alpha at $\alpha = .05$.

STEP 2 **Locate the critical region.** To locate the critical region, we must obtain the values for $df_{between}$ and df_{within}.

$$df_{\text{between}} = k - 1 = 3 - 1 = 2$$

$$df_{\text{within}} = N - k = 15 - 3 = 12$$

The F-ratio for this problem has $df = 2, 12$, and the critical F value for $\alpha = .05$ is $F = 3.88$.

STEP 3 Perform the analysis. The analysis involves the following steps:

1. Perform the analysis of *SS*.
2. Perform the analysis of *df*.
3. Calculate mean squares.
4. Calculate the F-ratio.

Perform the analysis of SS. We compute SS_{total} followed by its two components.

$$SS_{\text{total}} = \Sigma X^2 - \frac{G^2}{N} = 356 - \frac{60^2}{15} = 356 - \frac{3600}{15}$$
$$= 356 - 240 = 116$$

$$SS_{\text{within}} = \Sigma SS_{\text{inside each treatment}}$$
$$= 12 + 20 + 14$$
$$= 46$$

$$SS_{\text{between}} = \Sigma \frac{T^2}{n} - \frac{G^2}{N}$$
$$= \frac{5^2}{5} + \frac{25^2}{5} + \frac{30^2}{5} - \frac{60^2}{15}$$
$$= \frac{25}{5} + \frac{625}{5} + \frac{900}{5} - \frac{3600}{15}$$
$$= 5 + 125 + 180 - 240$$
$$= 70$$

Analyze degrees of freedom. We compute df_{total}. Its components, df_{between} and df_{within}, were previously calculated (see step 2).

$$df_{\text{total}} = N - 1 = 15 - 1 = 14$$

$$df_{\text{between}} = 2$$

$$df_{\text{within}} = 12$$

Calculate the MS values. We determine the values for MS_{between} and MS_{within}.

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{70}{2} = 35$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{46}{12} = 3.83$$

Compute the F-ratio. Finally, we can compute F.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{35}{3.83} = 9.14$$

STEP 4 Make a decision about H_0, and state a conclusion. The obtained F of 9.14 exceeds the critical value of 3.88. Therefore, we can reject the null hypothesis. The type of keyboard used has a significant effect on the number of errors committed, $F(2, 12) = 9.14$, $p < .05$. The following table summarizes the results of the analysis:

Source	*SS*	*df*	*MS*	
Between treatments	70	2	35	$F = 9.14$
Within treatments	46	12	3.83	
Total	116	14		

DEMONSTRATION 12.2

COMPUTING EFFECT SIZE FOR ANOVA

We compute eta squared (η^2), the percentage of variance explained, for the data that were analyzed in Demonstration 12.1. The data produced a between-treatments *SS* of 70 and a total *SS* of 116. Thus,

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} = \frac{70}{116} = 0.60 \text{ (or } 60\%)$$

PROBLEMS

1. Explain why the F-ratio is expected to be near 1.00 when the null hypothesis is true.

2. Describe the similarities between an F-ratio and a t statistic.

3. Several factors influence the size of the F-ratio. For each of the following, indicate whether it would influence the numerator or the denominator of the F-ratio, and indicate whether the size of the F-ratio would increase or decrease.
- **a.** Increase the differences between the sample means.
- **b.** Increase the size of the sample variances.

4. Why should you use ANOVA instead of several t tests to evaluate mean differences when an experiment consists of three or more treatment conditions?

5. Posttests are done after an ANOVA.
- **a.** What is the purpose of posttests?
- **b.** Explain why you do not need posttests if the analysis is comparing only two treatments.
- **c.** Explain why you do not need posttests if the decision from the ANOVA is to fail to reject the null hypothesis.

6. An independent-measures research study compares three treatment conditions with a sample of $n = 10$ in each condition. The sample means are $M_1 = 2$, $M_2 = 3$, and $M_3 = 7$.
- **a.** Compute *SS* for the set of 3 treatment means. (Use the three means as a set of $n = 3$ scores and compute *SS*.)
- **b.** Using the result from part a, compute $n(SS_{\text{means}})$. Note that this value is equal to SS_{between} (see Equation 12.6).
- **c.** Now, compute SS_{between} with the computational formula using the T values (Equation 12.7). You should obtain the same result as in part b.

7. The following data summarize the results from an independent-measures study comparing three treatment conditions.

I	II	III	
$n = 6$	$n = 6$	$n = 6$	
$M = 1$	$M = 5$	$M = 6$	$N = 18$
$T = 6$	$T = 30$	$T = 36$	$G = 72$
$SS = 30$	$SS = 35$	$SS = 40$	$\Sigma X^2 = 477$

a. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant differences among the three treatment means.
b. Calculate η^2 to measure the effect size for this study.
c. Write a sentence demonstrating how a research report would present the results of the hypothesis test and the measure of effect size.

8. For the preceding problem you should find that there are significant differences among the three treatments. The primary reason for the significance is that the mean for treatment I is substantially smaller than the means for the other two treatments. To create the following data, we started with the values from problem 7 and added 3 points to each score in treatment I. Recall that adding a constant causes the mean to change but has no influence on the variability of the sample. In the resulting data, the mean differences are much smaller than those in problem 7.

I	II	III	
$n = 6$	$n = 6$	$n = 6$	
$M = 4$	$M = 5$	$M = 6$	$N = 18$
$T = 24$	$T = 30$	$T = 36$	$G = 90$
$SS = 30$	$SS = 35$	SS = 40	$\Sigma X^2 = 567$

a. Before you begin any calculations, predict how the change in the data should influence the outcome of the analysis. That is, how will the *F*-ratio and the value of η^2 for these data compare with the values obtained in problem 7?
b. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant differences among the three treatment means. (Does your answer agree with your prediction in part a?)
c. Calculate η^2 to measure the effect size for this study. (Does your answer agree with your prediction in part a?)

9. The following data summarize the results from an independent-measures study comparing three treatment conditions.

I	II	III	
$n = 5$	$n = 5$	$n = 5$	
$M = 2$	$M = 5$	$M = 8$	$N = 15$
$T = 10$	$T = 25$	$T = 40$	$G = 75$
$SS = 16$	$SS = 20$	$SS = 24$	$\Sigma X^2 = 525$

a. Calculate the sample variance for each of the three samples.
b. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant differences among the three treatment means.

10. For the preceding problem you should find that there are significant differences among the three treatments. One reason for the significance is that the sample variances are relatively small. To create the following data, we started with the values from problem 9 and increased the variability (the *SS* values) within each sample.

I	II	III	
$n = 5$	$n = 5$	$n = 5$	
$M = 2$	$M = 5$	$M = 8$	$N = 15$
$T = 10$	$T = 25$	$T = 40$	$G = 75$
$SS = 64$	$SS = 80$	$SS = 96$	$\Sigma X^2 = 705$

a. Calculate the sample variance for each of the three samples. Describe how these sample variances compare with those from problem 9.
b. Predict how the increase in sample variance should influence the outcome of the analysis. That is, how will the *F*-ratio for these data compare with the value obtained in problem 9?
c. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant differences among the three treatment means. (Does your answer agree with your prediction in part b?)

11. Binge drinking on college campuses has been a hot topic in the popular media and in scholarly research. Flett, Goldstein, Wall, Hewitt, Wekerle, and Azzi (2008) report the results of a study relating perfectionism to binge drinking. In the study, students were classified into three groups based on the number of binge drinking episodes they experienced during the past month (0, 1, 2 or more). The students then completed a perfectionism questionnaire including one scale measuring parental criticism. One sample item is "I never felt that I could meet my parents' standards." Students rated their level of agreement with each item, and the total score was calculated for each student. The following results are similar to those obtained by the researchers.

Binge Drinking Episodes in Past Month			
0	1	2 or more	
8	10	13	$N = 15$
8	12	14	
10	8	12	$G = 165$
9	9	15	
10	11	16	$\Sigma X^2 = 1909$
$M = 9$	$M = 10$	$M = 14$	
$T = 45$	$T = 50$	$T = 70$	
$SS = 4$	$SS = 10$	$SS = 10$	

a. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant differences among the three treatment means.
b. Calculate η^2 to measure the effect size for this study.
c. Write a sentence demonstrating how a research report would present the results of the hypothesis test and the measure of effect size.

12. A researcher reports an F-ratio with $df = 3, 36$ from an independent-measures research study.
a. How many treatment conditions were compared in the study?
b. What was the total number of participants in the study?

13. A research report from an independent-measures study states that there are significant differences between treatments, $F(2, 54) = 3.58$, $p < .05$.
a. How many treatment conditions were compared in the study?
b. What was the total number of participants in the study?

14. There is some evidence that high school students justify cheating in class on the basis of poor teacher skills or low levels of teacher caring (Murdock, Miller, and Kohlhardt, 2004). Students appear to rationalize their illicit behavior based on perceptions of how their teachers view cheating. Poor teachers are thought not to know or care whether students cheat, so cheating in their classes is okay. Good teachers, on the other hand, do care and are alert to cheating, so students tend not to cheat in their classes. Following are hypothetical data similar to the actual research results. The scores represent judgments of the acceptability of cheating for the students in each sample.

Poor Teacher	Average Teacher	Good Teacher	
$n = 6$	$n = 8$	$n = 10$	$N = 24$
$M = 6$	$M = 2$	$M = 2$	$G = 72$
$SS = 30$	$SS = 33$	$SS = 42$	$\Sigma X^2 = 393$

a. Use an ANOVA with $\alpha = .05$ to determine whether there are significant differences in student judgments depending on how they see their teachers.
b. Calculate η^2 to measure the effect size for this study.
c. Write a sentence demonstrating how a research report would present the results of the hypothesis test and the measure of effect size.

15. The following summary table presents the results from an ANOVA comparing three treatment conditions with $n = 8$ participants in each condition. Complete all missing values. (*Hint:* Start with the *df* column.)

Source	*SS*	*df*	*MS*	
Between treatments	____	____	15	$F =$ ____
Within treatments	____	____	____	
Total	93	____		

16. A pharmaceutical company has developed a drug that is expected to reduce hunger. To test the drug, two samples of rats are selected with $n = 20$ in each sample. The rats in the first sample receive the drug every day and those in the second sample are given a placebo. The dependent variable is the amount of food eaten by each rat over a 1-month period. An ANOVA is used to evaluate the difference between the two sample means and the results are reported in the following summary table. Fill in all missing values in the table. (*Hint:* Start with the *df* column.)

Source	*SS*	*df*	*MS*	
Between treatments	____	____	20	$F = 4.00$
Within treatments	____	____	____	
Total	____	____		

17. A developmental psychologist is examining the development of language skills from age 2 to age 4. Three different groups of children are obtained, one for each age, with $n = 16$ children in each group. Each child is given a language-skills assessment test. The resulting data were analyzed with an ANOVA to test for mean differences between age groups. The results of the ANOVA are presented in the following table. Fill in all missing values.

Source	*SS*	*df*	*MS*	
Between treatments	20	____	____	$F =$ ____
Within treatments	____	____	____	
Total	200	____		

18. The following data were obtained from an independent-measures research study comparing three treatment conditions. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant mean differences among the treatments.

Treatment			
I	II	III	
2	5	7	$N = 14$
5	2	3	$G = 42$
0	1	6	$\Sigma X^2 = 182$
1	2	4	
2			
2			
$T = 12$	$T = 10$	$T = 20$	
$SS = 14$	$SS = 9$	$SS = 10$	

19. The following values summarize the results from an independent-measures study comparing two treatment conditions.

a. Use an independent-measures t test with $\alpha = .05$ to determine whether there is a significant mean difference between the two treatments.

b. Use an ANOVA with $\alpha = .05$ to determine whether there is a significant mean difference between the two treatments.

Treatment		
I	II	
$n = 8$	$n = 4$	
$M = 4$	$M = 10$	$N = 12$
$T = 32$	$T = 40$	$G = 72$
$SS = 45$	$SS = 15$	$\Sigma X^2 = 588$

20. The following data represent the results from an independent-measures study comparing two treatment conditions.

a. Use an independent-measures t test with $\alpha = .05$ to determine whether there is a significant mean difference between the two treatments.

b. Use an ANOVA with $\alpha = .05$ to determine whether there is a significant mean difference between the two treatments.

Treatment		
I	II	
8	2	$N = 10$
7	3	$G = 50$
6	3	$\Sigma X^2 = 306$
5	5	
9	2	
$M = 7$	$M = 3$	
$T = 35$	$T = 15$	
$SS = 10$	$SS = 6$	

21. One possible explanation for why some birds migrate and others maintain year round residency in a single location is intelligence. Specifically, birds with small brains, relative to their body size, are simply not smart enough to find food during the winter and must migrate to warmer climates where food is easily available (Sol, Lefebvre, & Rodriguez-Teijeiro, 2005). Birds with bigger brains, on the other hand, are more creative and can find food even when the weather turns harsh. Following are hypothetical data similar to the actual research results. The numbers represent relative brain size for the individual birds in each sample.

Non-Migrating	Short-Distance Migrants	Long-Distance Migrants	
18	6	4	$N = 18$
13	11	9	$G = 180$
19	7	5	$\Sigma X^2 = 2150$
12	9	6	
16	8	5	
12	13	7	
$M = 15$	$M = 9$	$M = 6$	
$T = 90$	$T = 54$	$T = 36$	
$SS = 48$	$SS = 34$	$SS = 16$	

a. Use an ANOVA with $\alpha = .05$ to determine whether there are any significant mean differences among the three groups of birds.

b. Compute η^2, the percentage of variance explained by the group differences, for these data.

c. Write a sentence demonstrating how a research report would present the results of the hypothesis test and the measure of effect size.
d. Use the Tukey HSD posttest to determine which groups are significantly different.

22. There is some research indicating that college students who use Facebook while studying tend to have lower grades than non-users (Kirschner & Karpinski, 2010). A representative study surveys students to determine the amount of Facebook use during the time they are studying or doing homework. Based on the amount of time spent on Facebook, students are classified into three groups and their grade point averages are recorded. The following data show the typical pattern of results.

Facebook Use While Studying		
Non-User	Rarely Use	Regularly Use
3.70	3.51	3.02
3.45	3.42	2.84
2.98	3.81	3.42
3.94	3.15	3.10
3.82	3.64	2.74
3.68	3.20	3.22
3.90	2.95	2.58
4.00	3.55	3.07
3.75	3.92	3.31
3.88	3.45	2.80

a. Use an ANOVA with $\alpha = .05$ to determine whether there are significant mean differences among the three groups.
b. Compute η^2 to measure the size of the effect.
c. Write a sentence demonstrating how the result from the hypothesis test and the measure of effect size would appear in a research report.

23. New research suggests that watching television, especially medical shows such as *Grey's Anatomy* and *House* can result in more concern about personal health (Ye, 2010). Surveys administered to college students measure television viewing habits and health concerns such as fear of developing the diseases and disorders seen on television. For the following data, students are classified into three categories based on their television viewing patterns and health concerns are measured on a 10-point scale with 0 indicating "none."

Television Viewing		
Little or None	Moderate	Substantial
4	5	5
2	7	7
5	3	6
1	4	6
3	8	8
7	6	9
4	2	6
4	7	4
8	3	6
2	5	8

a. Use an ANOVA with $\alpha = .05$ to determine whether there are significant mean differences among the three groups.
b. Compute η^2 to measure the size of the effect.
c. Use Tukey's HSD test with $\alpha = .05$ to determine which groups are significantly different.

CHAPTER

13

Repeated-Measures Analysis of Variance

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Independent-measures analysis of variance (Chapter 12)
- Repeated-measures designs (Chapter 11)
- Individual differences

Preview

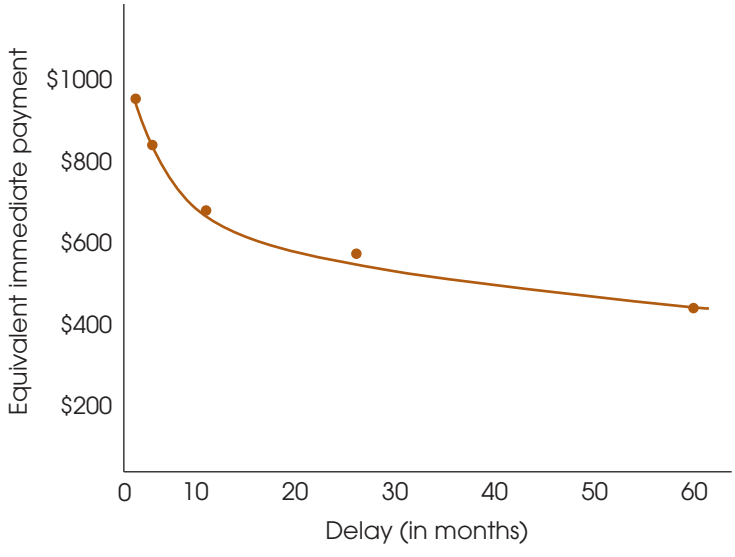
Suppose that you were offered a choice between receiving $1000 in 5 years or a smaller amount today. How much would you be willing to take today to avoid waiting 5 years to get the full $1000 payment?

The general result for this kind of decision is that the longer the $1000 payment is delayed, the smaller the amount that people will accept today. For example, you may be willing to take $300 today rather than waiting 5 years for the $1000. However, you might be willing to settle for $100 today if you have to wait 10 years for the $1000. This phenomenon is known as delayed discounting because people discount the value of a future reward depending on how long it is delayed (Green, Fry, & Myerson, 1994). In a typical study examining delayed discounting, people are asked to place a value on a future reward for several different delay periods. For example, how much would you accept today instead of waiting for a future reward of $1000 if you had to wait 1 month before receiving the payment? How about waiting 6 months? 12 months? 24 months? 60 months?

Typical results for a sample of college students are shown in Figure 13.1. Note that the average value declines regularly as the delay period increases. The statistical question is whether the mean differences from one delay period to another are significant.

FIGURE 13.1

Mean amount of immediate payment selected as equivalent to receiving a $1000 payment at each delay.



The Problem: You should recognize that evaluating mean differences for more than two sample means is a job for analysis of variance (ANOVA). However, the discounting study is a repeated-measures design with five scores for each individual, and the ANOVA introduced in Chapter 12 is intended for independent-measures studies. Once again, a new hypothesis test is needed.

The Solution: In this chapter we introduce the *repeated-measures ANOVA*. As the name implies, this new procedure is used to evaluate the differences between two or more sample means obtained from a repeated-measures research study. As you will see, many of the notational symbols and computations are the same as those used for the independent-measures ANOVA. In fact, your best preparation for this chapter is a good understanding of the basic ANOVA procedure presented in Chapter 12.

13.1 OVERVIEW OF REPEATED-MEASURES DESIGNS

In the preceding chapter, we introduced ANOVA as a hypothesis-testing procedure for evaluating differences among two or more sample means. The specific advantage of ANOVA, especially in contrast to *t* tests, is that ANOVA can be used to evaluate the significance of mean differences in situations in which there are more than two sample means being compared. However, the presentation of ANOVA in Chapter 12 was limited to single-factor, independent-measures research designs. Recall that *single factor* indicates that the research study involves only one independent variable (or only one quasi-independent variable), and the term *independent-measures* indicates that the study uses a separate sample for each of the different treatment conditions being compared.

In this chapter, we extend the ANOVA procedure to single-factor, repeated-measures designs. The defining characteristic of a repeated-measures design is that one group of individuals participates in all of the different treatment conditions. The repeated-measures ANOVA is used to evaluate mean differences in two general research situations:

1. An experimental study in which the researcher manipulates an independent variable to create two or more treatment conditions, with the same group of individuals tested in all of the conditions.
2. A nonexperimental study in which the same group of individuals is simply observed at two or more different times.

Examples of these two research situations are presented in Table 13.1. Table 13.1(a) shows data from a study in which the researcher changes the type of distraction to

TABLE 13.1

Two sets of data representing typical examples of single-factor, repeated-measures research designs.

(a) Data from an experimental study evaluating the effects of different types of distraction on the performance of a visual detection task.

	Visual Detection Scores		
Participant	No Distraction	Visual Distraction	Auditory Distraction
A	47	22	41
B	57	31	52
C	38	18	40
D	45	32	43

(b) Data from a nonexperimental design evaluating the effectiveness of a clinical therapy for treating depression.

	Depression Scores		
Participant	Before Therapy	After Therapy	6-Month Follow-Up
A	71	53	55
B	62	45	44
C	82	56	61
D	77	50	46
E	81	54	55

create three treatment conditions. One group of participants is then tested in all three conditions. In this study, the factor being examined is the type of distraction.

Table 13.1(b) shows a study in which a researcher observes depression scores for the same group of individuals at three different times. In this study, the time of measurement is the factor being examined. Another common example of this type of design is found in developmental psychology when the participants' age is the factor being studied. For example, a researcher could study the development of vocabulary skill by measuring vocabulary for a sample of 3-year-old children, then measuring the same children again at ages 4 and 5.

13.2 THE REPEATED-MEASURES ANOVA

HYPOTHESES FOR THE REPEATED-MEASURES ANOVA

The hypotheses for the repeated-measures ANOVA are exactly the same as those for the independent-measures ANOVA presented in Chapter 12. Specifically, the null hypothesis states that, for the general population, there are no mean differences among the treatment conditions being compared. In symbols,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots$$

The null hypothesis states that, on average, all of the treatments have exactly the same effect. According to the null hypothesis, any differences that may exist among the sample means are not caused by systematic treatment effects but rather are the result of random and unsystematic factors.

The alternative hypothesis states that there are mean differences among the treatment conditions. Rather than specifying exactly which treatments are different, we use a generic version of H_1, which simply states that differences exist:

H_1: At least one treatment mean (μ) is different from another.

Notice that the alternative says that, on average, the treatments do have different effects. Thus, the treatment conditions may be responsible for causing mean differences among the samples. As always, the goal of the ANOVA is to use the sample data to determine which of the two hypotheses is more likely to be correct.

THE *F*-RATIO FOR REPEATED-MEASURES ANOVA

The F-ratio for the repeated-measures ANOVA has the same structure that was used for the independent-measures ANOVA in Chapter 12. In each case, the F-ratio compares the actual mean differences between treatments with the amount of difference that would be expected if there were no treatment effect. The numerator of the F-ratio measures the actual mean differences between treatments. The denominator measures how big the differences should be if there is no treatment effect. As always, the F-ratio uses variance to measure the size of the differences. Thus, the F-ratio for both ANOVAs has the general structure

$$F = \frac{\text{variance (differences) between treatments}}{\text{variance (differences) expected if there is no treatment effect}}$$

A large value for the F-ratio indicates that the differences between treatments are greater than would be expected without any treatment effect. If the F-ratio is larger than

the critical value in the *F* distribution table, then we can conclude that the differences between treatments are *significantly* larger than would be caused by chance.

Individual Differences in the *F*-ratio Although the structure of the *F*-ratio is the same for independent-measures and repeated-measures designs, there is a fundamental difference between the two designs that produces a corresponding difference in the two *F*-ratios. Specifically, individual differences are a part of one ratio but are eliminated from the other.

You should recall that the term *individual differences* refers to participant characteristics such as age, personality, and gender that vary from one person to another and may influence the measurements that you obtain for each person. Suppose, for example, that you are measuring reaction time. The first participant in your study is a 19-year-old female with an IQ of 136 who is on the college varsity volleyball team. The next participant is a 42-year-old male with an IQ of 111 who returned to college after losing his job and comes to the research study with a head cold. Would you expect to obtain the same reaction time score for these two individuals?

Individual differences are a part of the variance in the numerator and in the denominator of the *F*-ratio for the independent-measures ANOVA. However, individual difference are eliminated or removed from the variances in the F-ratio for the repeated measures ANOVA. The idea of removing individual differences was first presented in Chapter 11 when we introduced the repeated-measures design (p. 367), but we review it briefly now.

In a repeated-measures study, exactly the same individuals participate in all of the treatment conditions. Therefore, if there are any mean differences between treatments, they cannot be explained by individual differences. Thus, individual differences are automatically eliminated from the numerator of the repeated-measures *F*-ratio.

A repeated-measures design also allows you to remove individual differences from the variance in the denominator of the *F*-ratio. Because the same individuals are measured in every treatment condition, it is possible to measure the size of the individual differences. In Table 13.1(a), for example, participant A has scores that are consistently 10 points lower than the scores for participant B. Because the individual differences are systematic and predictable, they can be measured and separated from the random, unsystematic differences in the denominator of the *F*-ratio.

Thus, individual differences are automatically eliminated from the numerator of the repeated-measures *F*-ratio. In addition, they can be measured and removed from the denominator. As a result, the structure of the final *F*-ratio is as follows:

$$F = \frac{\begin{array}{c}\text{variance/differences between treatments}\\ \text{(without individual differences)}\end{array}}{\begin{array}{c}\text{variance/differences with no treatment effect}\\ \text{(with individual differences removed)}\end{array}}$$

The process of removing individual differences is an important part of the procedure for a repeated-measures ANOVA.

THE LOGIC OF THE REPEATED-MEASURES ANOVA

The general purpose of the repeated-measures ANOVA is to determine whether the differences that are found between treatment conditions are significantly greater than would be expected if there is no treatment effect. In the numerator of the *F*-ratio, the *between-treatments variance* measures the actual mean differences between the treatment conditions. The variance in the denominator is intended to measure how much

difference is reasonable to expect if there are no systematic treatment effects and no systematic individual differences. In other words, the denominator measures variability caused entirely by random and unsystematic factors. For this reason, the variance in the denominator is called the *error variance*. In this section we examine the elements that make up the two variances in the repeated-measures F-ratio.

The numerator of the *F*-ratio: between-treatments variance Logically, any differences that are found between treatments can be explained by only two factors:

1. **Systematic Differences Caused by the Treatments.** It is possible that the different treatment conditions really do have different effects and, therefore, cause the individuals' scores in one condition to be higher (or lower) than in another. Remember that the purpose for the research study is to determine whether a *treatment effect* exists.
2. **Random, Unsystematic Differences.** Even if there is no treatment effect, it is possible for the scores in one treatment condition to be different from the scores in another. For example, suppose that I measure your IQ score on a Monday morning. A week later I come back and measure your IQ again under exactly the same conditions. Will you get exactly the same IQ score both times? In fact, minor differences between the two measurement situations would probably cause you to end up with two different scores. For example, for one of the IQ tests you might be more tired, or hungry, or worried, or distracted than you were on the other test. These differences can cause your scores to vary. The same thing can happen in a repeated-measures research study. The same individuals are measured at two or more different times and, even though there may be no difference between the two treatment conditions, you can still end up with different scores. However, these differences are random and unsystematic and are classified as error variance.

Thus, it is possible that any differences (or variance) found between treatments could be caused by treatment effects, and it is possible that the differences could simply be the result of chance. On the other hand, it is *impossible* that the differences between treatments are caused by individual differences. Because the repeated-measures design uses exactly the same individuals in every treatment condition, individual differences are *automatically eliminated* from the variance between treatments in the numerator of the F-ratio.

The denominator of the *F*-ratio: error variance The goal of the ANOVA is to determine whether the differences that are observed in the data are greater than would be expected without any systematic treatment effects. To accomplish this goal, the denominator of the F-ratio is intended to measure how much difference (or variance) is reasonable to expect from random and unsystematic factors. This means that we must measure the variance that exists when there are no treatment effects or any other systematic differences.

We begin exactly as we did with the independent-measures F-ratio; specifically, we calculate the variance that exists within treatments. Recall from Chapter 12 that within each treatment all of the individuals are treated in exactly the same way. Therefore, any differences that exist within treatments cannot be caused by treatment effects.

In a repeated-measures design, however, it is also possible that individual differences can cause systematic differences between the scores within treatments. For example, one individual may score consistently higher than another. To eliminate the

individual differences from the denominator of the *F*-ratio, we measure the individual differences and then subtract them from the rest of the variability. The variance that remains is a measure of pure *error* without any systematic differences that can be explained by treatment effects or by individual differences.

In summary, the *F*-ratio for a repeated-measures ANOVA has the same basic structure as the *F*-ratio for independent measures (Chapter 12) except that it includes no variability caused by individual differences. The individual differences are automatically eliminated from the variance between treatments (numerator) because the repeated-measures design uses the same individuals in all treatments. In the denominator, the individual differences are subtracted during the analysis. As a result, the repeated-measures *F*-ratio has the following structure:

$$F = \frac{\text{between-treatments variance}}{\text{error variance}}$$

$$= \frac{\text{treatment effects} + \text{random, unsystematic differences}}{\text{random, unsystematic differences}} \qquad (13.1)$$

Note that this *F*-ratio is structured so that there are no individual differences contributing to either the numerator or the denominator. When there is no treatment effect, the *F*-ratio is balanced because the numerator and denominator are both measuring exactly the same variance. In this case, the *F*-ratio should have a value near 1.00. When research results produce an *F*-ratio near 1.00, we conclude that there is no evidence of a treatment effect and we fail to reject the null hypothesis. On the other hand, when a treatment effect does exist, it contributes only to the numerator and should produce a large value for the *F*-ratio. Thus, a large value for *F* indicates that there is a real treatment effect and, therefore, we should reject the null hypothesis.

LEARNING CHECK

1. Explain why individual differences do not contribute to the between-treatments variability in a repeated-measures study.
2. What sources of variability contribute to the within-treatment variability for a repeated-measures study?
3. Describe the structure of the *F*-ratio for the repeated-measures ANOVA.

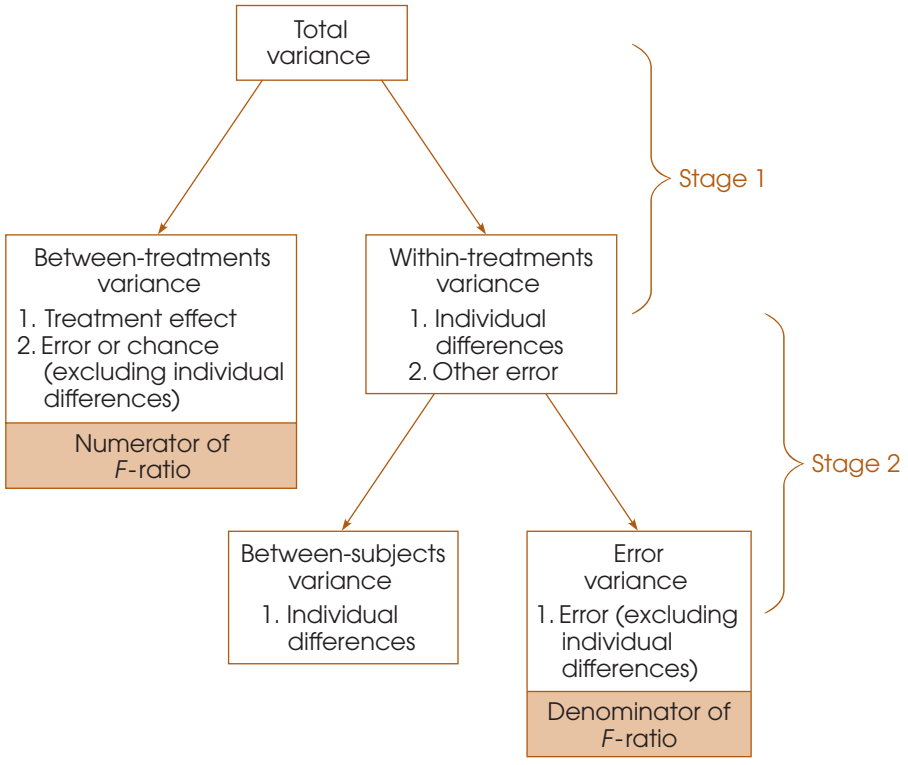
ANSWERS

1. Because the individuals in one treatment are exactly the same as the individuals in every other treatment, there are no individual differences from one treatment to another.
2. Variability (differences) within treatments is caused by individual differences and random, unsystematic differences.
3. The numerator of the *F*-ratio measures between-treatments variability, which consists of treatment effects and random, unsystematic differences. The denominator measures variability that is exclusively caused by random, unsystematic differences.

13.3 HYPOTHESIS TESTING AND EFFECT SIZE WITH THE REPEATED-MEASURES ANOVA

The overall structure of the repeated-measures ANOVA is shown in Figure 13.2. Note that the ANOVA can be viewed as a two-stage process. In the first stage, the total variance is partitioned into two components: *between-treatments variance* and *within-treatments*

FIGURE 13.2
The partitioning of variance for a repeated-measures experiment.



variance. This stage is identical to the analysis that we conducted for an independent-measures design in Chapter 12.

The second stage of the analysis is intended to remove the individual differences from the denominator of the *F*-ratio. In the second stage, we begin with the variance within treatments and then measure and subtract out the *between-subject variance,* which measures the size of the individual differences. The remaining variance, often called the *residual variance,* or *error variance,* provides a measure of how much variance is reasonable to expect after the treatment effects and individual differences have been removed. The second stage of the analysis is what differentiates the repeated-measures ANOVA from the independent-measures ANOVA. Specifically, the repeated-measures design requires that the individual differences be removed.

DEFINITION

In a repeated-measures ANOVA, the denominator of the *F*-ratio is called the **residual variance,** or the **error variance,** and measures how much variance is expected if there are no systematic treatment effects and no individual differences contributing to the variability of the scores.

NOTATION FOR THE REPEATED-MEASURES ANOVA

We use the data in Table 13.2 to introduce the notation for the repeated-measures ANOVA. The data represent the results of a study comparing different viewing distances for a 42-inch high-definition television. Four viewing distances were evaluated,

TABLE 13.2

Satisfaction with different viewing distances of a 42-inch, high-definition television.
Note: For comparison, the scores are identical to the values in Example 12.1 on page 405.

	Viewing Distance					
Person	9 Feet	12 Feet	15 Feet	18 Feet	Person Totals	
A	3	4	7	6	$P = 20$	$n = 5$
B	0	3	6	3	$P = 12$	$k = 4$
C	2	1	5	4	$P = 12$	$N = 20$
D	0	1	4	3	$P = 8$	$G = 60$
E	0	1	3	4	$P = 8$	$\Sigma X^2 = 262$
	$T = 5$	$T = 10$	$T = 25$	$T = 20$		
	$SS = 8$	$SS = 8$	$SS = 10$	$SS = 6$		

9 feet, 12 feet, 15 feet, and 18 feet. Each participant was free to move back and forth among the four distances while watching a 30-minute video on the television. The only restriction was that each person had to spend at least 2 minutes watching from each of the four distances. At the end of the video, each participant rated the all of the viewing distances on a scale from 1 (Very Bad, definitely need to move closer or farther away) to 7 (excellent, perfect viewing distance). You may notice that this research study and the numerical values in the table are identical to those used to demonstrate the independent-measures ANOVA in the previous chapter (Example 12.1, page 405). In this case, however, the data represent a repeated-measures study in which the same group of $n = 5$ individuals is tested in all four treatment conditions.

You should recognize that most of the notation in Table 13.2 is identical to the notation used in an independent-measures analysis (Chapter 12). For example, there are $n = 5$ participants who are tested in $k = 4$ treatment conditions, producing a total of $N = 20$ scores that add up to a grand total of $G = 60$. Note, however, that $N = 20$ now refers to the total number of scores in the study, not the number of participants.

The repeated-measures ANOVA introduces only one new notational symbol. The letter P is used to represent the total of all of the scores for each individual in the study. You can think of the P values as "Person totals" or "Participant totals." In Table 13.2, for example, participant A had scores of 3, 4, 6, and 7 for a total of $P = 20$. The P values are used to define and measure the magnitude of the individual differences in the second stage of the analysis.

EXAMPLE 13.1

We use the data in Table 13.2 to demonstrate the repeated-measures ANOVA. Again, the goal of the test is to determine whether there are any significant differences among the four distances being compared. Specifically, are any of the mean differences in the data greater than would be expected if there are no systematic differences among the four viewing distances?

STAGE 1 OF THE REPEATED-MEASURES ANOVA

The first stage of the repeated-measures analysis is identical to the independent-measures ANOVA that was presented in Chapter 12. Specially, the *SS* and *df* for the total set of scores are analyzed into within-treatments and between-treatments components.

Because the numerical values in Table 13.2 are the same as the values used in Example 12.1 (p. 405), the computations for the first stage of the repeated-measures analysis are identical to those in Example 12.1. Rather than repeating the same

arithmetic, the results of the first stage of the repeated-measures analysis can be summarized as follows:

Total:

$$SS_{\text{total}} = \Sigma X^2 - \frac{G^2}{N} = 262 - \frac{(60)^2}{20} = 262 - 180 = 82$$

$$df_{\text{total}} = N - 1 = 19$$

Within treatments:

$$SS_{\text{within treatments}} = \Sigma SS_{\text{inside each treatment}} = 8 + 8 + 10 + 6 = 32$$

$$df_{\text{within treatments}} = \Sigma df_{\text{inside each treatment}} = 4 + 4 + 4 + 4 = 16$$

Between treatments: For this example we use the computational formula for $SS_{\text{between treatments}}$.

$$SS_{\text{between treatments}} = \Sigma \frac{T^2}{n} - \frac{G^2}{N} = \frac{5^2}{5} + \frac{10^2}{5} + \frac{25^2}{5} + \frac{20^2}{5} - \frac{60^2}{20} = 50$$

$$df_{\text{between treatments}} = k - 1 = 3$$

For more details on the formulas and calculations, see Example 12.1, pages 405–407.

This completes the first stage of the repeated-measures ANOVA. Note that the two components, between and within, add up to the total for the *SS* values and for the *df* values. Also note that the between-treatments *SS* and *df* values provide a measure of the mean differences between treatments and are used to compute the variance in the numerator of the final *F*-ratio.

STAGE 2 OF THE REPEATED-MEASURES ANOVA

The second stage of the analysis involves removing the individual differences from the denominator of the *F*-ratio. Because the same individuals are used in every treatment, it is possible to measure the size of the individual differences. For the data in Table 13.2, for example, participant A tends to have the highest scores and participants D and E tend to have the lowest scores. These individual differences are reflected in the *P* values, or person totals, in the right-hand column. We use these *P* values to create a computational formula for $SS_{\text{between subjects}}$ in much the same way that we used the treatment totals, the *T* values, in the computational formula for $SS_{\text{between treatments}}$. Specifically, the formula for the between-subjects *SS* is

$$SS_{\text{between subjects}} = \Sigma \frac{P^2}{k} - \frac{G^2}{N} \tag{13.2}$$

Notice that the formula for the between-subjects *SS* has exactly the same structure as the computational formula for the between-treatments *SS* (see the calculation above). In this case we use the person totals (*P* values) instead of the treatment totals (*T* values). Each *P* value is squared and divided by the number of scores that were added to obtain the total. In this case, each person has *k* scores, one for each treatment. Box 13.1 presents another demonstration of the similarity

of the formulas for *SS* between subjects and *SS* between treatments. For the data in Table 13.2,

$$SS_{\text{between subjects}} = \frac{20^2}{4} + \frac{12^2}{4} + \frac{12^2}{4} + \frac{8^2}{4} + \frac{8^2}{4} - \frac{60^2}{20}$$
$$= 100 + 36 + 36 + 16 + 16 - 180$$
$$= 24$$

The value of $SS_{\text{between subjects}}$ provides a measure of the size of the individual differences—that is, the differences between subjects. In the second stage of the analysis, we simply subtract the individual differences to obtain the measure of error that forms the denominator of the *F*-ratio. Thus, the final step in the analysis of *SS* is

$$SS_{\text{error}} = SS_{\text{within treatments}} - SS_{\text{between subjects}} \tag{13.3}$$

We have already computed $SS_{\text{within treatments}} = 32$ and $SS_{\text{between subjects}} = 24$, therefore

$$SS_{\text{error}} = 32 - 24 = 8$$

The analysis of degrees of freedom follows exactly the same pattern that was used to analyze *SS*. Remember that we are using the *P* values to measure the magnitude of the individual differences. The number of *P* values corresponds to the number of subjects, *n,* so the corresponding *df* is

$$df_{\text{between subjects}} = n - 1 \tag{13.4}$$

For the data in Table 13.2, there are $n = 5$ subjects and

$$df_{\text{between subjects}} = 5 - 1 = 4$$

BOX 13.1 $SS_{\text{between subjects}}$ AND $SS_{\text{between treatments}}$

The data for a repeated-measures study are normally presented in a matrix, with the treatment conditions determining the columns and the participants defining the rows. The data in Table 13.2 demonstrate this normal presentation. The calculation of $SS_{\text{between treatments}}$ provides a measure of the differences between treatment conditions—that is, a measure of the mean differences between the *columns* in the data matrix. For the data in Table 13.2, the column totals are 5, 10, 20, and 25. These values are variable, and $SS_{\text{between treatments}}$ measures the amount of variability.

The following table reproduces the data from Table 13.2, but now we have turned the data matrix on its side so that the participants define the columns and the treatment conditions define the rows.

In this new format, the differences between the columns represent the between-subjects variability. The column totals are now *P* values (instead of *T* values) and the number of scores in each column is now identified by *k* (instead of *n*). With these changes in notation, the formula for $SS_{\text{between subjects}}$ has exactly the same structure as the formula for $SS_{\text{between treatments}}$. If you examine the two equations, the similarity should be clear.

	Participant A	B	C	D	E	
9 feet	3	0	2	0	0	$T = 5$
12 feet	4	3	1	1	1	$T = 10$
15 feet	7	6	5	4	3	$T = 25$
18 feet	6	3	4	3	4	$T = 20$
	$P = 20$	$P = 12$	$P = 12$	$P = 8$	$P = 8$	

Next, we subtract the individual differences from the within-subjects component to obtain a measure of error. In terms of degrees of freedom,

$$df_{\text{error}} = df_{\text{within treatments}} - df_{\text{between subjects}} \tag{13.5}$$

For the data in Table 13.2,

$$df_{\text{error}} = 16 - 4 = 12$$

An algebraically equivalent formula for df_{error} uses only the number of treatment conditions (k) and the number of participants (n):

$$df_{\text{error}} = (k - 1)(n - 1) \tag{13.6}$$

The usefulness of equation 13.6 is discussed in Box 13.2.

Remember: The purpose for the second stage of the analysis is to measure the individual differences and then remove the individual differences from the denominator of the F-ratio. This goal is accomplished by computing SS and df between subjects (the individual differences) and then subtracting these values from the within-treatments values. The result is a measure of variability resulting from error with the individual differences removed. This error variance (SS and df) is used in the denominator of the F-ratio.

CALCULATION OF THE VARIANCES (*MS* VALUES) AND THE *F*-RATIO

The final calculation in the analysis is the F-ratio, which is a ratio of two variances. Each variance is called a *mean square,* or MS, and is obtained by dividing the appropriate SS by its corresponding df value. The MS in the numerator of the F-ratio measures the size of the differences between treatments and is calculated as

$$MS_{\text{between treatments}} = \frac{SS_{\text{between treatments}}}{df_{\text{between treatments}}} \tag{13.7}$$

For the data in Table 13.2,

$$MS_{\text{between treatments}} = \frac{50}{3} = 16.67$$

BOX 13.2 USING THE ALTERNATIVE FORMULA FOR df_{error}

The statistics presented in a research report not only describe the significance of the results but typically provide enough information to reconstruct the research design. The alternative formula for df_{error} is particularly useful for this purpose. Suppose, for example, that a research report for a repeated-measures study includes an F-ratio with $df = 2, 10$. How many treatment conditions were compared in the study, and how many individuals participated?

To answer these questions, begin with the first df value, which is $df_{\text{between treatments}} = 2 = k - 1$. From this value, it is clear that $k = 3$ treatments. Next, use the second df value, which is $df_{\text{error}} = 10$. Using this value and the fact that $k - 1 = 2$, use equation 13.6 to find the number of participants.

$$df_{\text{error}} = 10 = (k - 1)(n - 1) = 2(n - 1)$$

If $2(n - 1) = 10$, then $n - 1$ must equal 5. Therefore, $n = 6$.

Therefore, we conclude that a repeated-measures study producing an F-ratio with $df = 2, 10$ must have compared 3 treatment conditions using a sample of 6 participants.

The denominator of the *F*-ratio measures how much difference is reasonable to expect if there are no systematic treatment effects and the individual differences have been removed. This is the error variance, or the residual variance, obtained in stage 2 of the analysis.

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{df_{\text{error}}} \tag{13.8}$$

For the data in Table 13.2,

$$MS_{\text{error}} = \frac{8}{12} = 0.67$$

Finally, the *F*-ratio is computed as

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{error}}} \tag{13.9}$$

For the data in Table 13.2,

$$F = \frac{16.67}{0.67} = 24.88$$

Once again, notice that the repeated-measures ANOVA uses MS_{error} in the denominator of the *F*-ratio. This *MS* value is obtained in the second stage of the analysis, after the individual differences have been removed. As a result, individual differences are completely eliminated from the repeated-measures *F*-ratio, so that the general structure is

$$F = \frac{\text{treatment effects} + \text{unsystematic differences (without individual diffs)}}{\text{unsystematic differences (without individual diffs)}}$$

For the data we have been examining, the *F*-ratio is $F = 24.88$, indicating that the differences between treatments (numerator) are almost 25 times bigger than you would expect without any treatment effects (denominator). A ratio this large provides clear evidence that there is a real treatment effect. To verify this conclusion you must consult the *F* distribution table to determine the appropriate critical value for the test. The degrees of freedom for the *F*-ratio are determined by the two variances that form the numerator and the denominator. For a repeated-measures ANOVA, the *df* values for the *F*-ratio are reported as

$$df = df_{\text{between treatments}}, df_{\text{error}}$$

For the example we are considering, the *F*-ratio has $df = 2, 12$ ("degrees of freedom equal two and twelve"). Using the *F* distribution table (p. 705) with $\alpha = .05$, the critical value is $F = 3.88$, and with $\alpha = .01$ the critical value is $F = 6.93$. Our obtained *F*-ratio, $F = 24.88$, is well beyond either of the critical values, so we can conclude that the differences between treatments are *significantly* greater than expected by chance using either $\alpha = .05$ or $\alpha = .01$.

The summary table for the repeated-measures ANOVA from Example 13.1 is presented in Table 13.3. Although these tables are no longer commonly used in research reports, they provide a concise format for displaying all of the elements of the analysis.

MEASURING EFFECT SIZE FOR THE REPEATED-MEASURES ANOVA

The most common method for measuring effect size with ANOVA is to compute the percentage of variance that is explained by the treatment differences. In the context of ANOVA, the percentage of variance is commonly identified as η^2 (eta squared). In Chapter 12, for the independent-measures analysis, we computed η^2 as

$$\eta^2 = \frac{SS_{\text{between treatments}}}{SS_{\text{between treatments}} + SS_{\text{within treatments}}} = \frac{SS_{\text{between treatments}}}{SS_{\text{total}}}$$

The intent is to measure how much of the total variability is explained by the differences between treatments. With a repeated-measures design, however, there is another component that can explain some of the variability in the data. Specifically, part of the variability is caused by differences between individuals. In Table 13.2, for example, person A consistently scored higher than person B. This consistent difference explains some of the variability in the data. When computing the size of the treatment effect, it is customary to remove any variability that can be explained by other factors, and then compute the percentage of the remaining variability that can be explained by the treatment effects. Thus, for a repeated-measures ANOVA, the variability from the individual differences is removed before computing η^2. As a result, η^2 is computed as

$$\eta^2 = \frac{SS_{\text{between treatments}}}{SS_{\text{total}} - SS_{\text{between subjects}}} \tag{13.10}$$

Because Equation 13.10 computes a percentage that is not based on the total variability of the scores (one part, $SS_{\text{between subjects}}$, is removed), the result is often called a *partial* eta squared.

The general goal of Equation 13.10 is to calculate a percentage of the variability that has not already been explained by other factors. Thus, the denominator of Equation 13.10 is limited to variability from the treatment differences and variability that is exclusively from random, unsystematic factors. With this in mind, an equivalent version of the η^2 formula is

$$\eta^2 = \frac{SS_{\text{between treatments}}}{SS_{\text{between treatments}} + SS_{\text{error}}} \tag{13.11}$$

TABLE 13.3

A summary table for the repeated-measures ANOVA for the data from Example 13.1.

Source	*SS*	*df*	*MS*	*F*
Between treatments	50	3	16.67	$F(3,12) = 24.88$
Within treatments	32	16		
Between subjects	24	4		
Error	8	12	0.67	
Total	82	19		

In this new version of the eta-squared formula, the denominator consists of the variability that is explained by the treatment differences plus the other *unexplained* variability. Using either formula, the data from Example 13.1 produce

$$\eta^2 = \frac{50}{58} = 0.862 \ (\text{or } 86.2\%)$$

This result means that 86.2% of the variability in the data (except for the individual differences) is accounted for by the differences between treatments.

IN THE LITERATURE

REPORTING THE RESULTS OF A REPEATED-MEASURES ANOVA

As described in Chapter 12 (p. 409), the format for reporting ANOVA results in journal articles consists of

1. A summary of descriptive statistics (at least treatment means and standard deviations, and tables or graphs as needed)
2. A concise statement of the outcome of the ANOVA

For the study in Example 13.1, the report could state:

> The means and variances for the four television viewing distances are shown in Table 1. A repeated-measures analysis of variance indicated significant mean differences in the participants' ratings of the four distances, $F(3, 12) = 24.88$, $p < .01$, $\eta^2 = 0.862$.
>
> **TABLE 1**
>
> Ratings of satisfaction with different television-viewing distances
>
	9 Feet	12 Feet	15 Feet	18 Feet
M	1.00	2.00	5.00	4.00
SD	1.41	1.41	1.58	1.22

POST HOC TESTS WITH REPEATED-MEASURES ANOVA

Recall that ANOVA provides an overall test of significance for the mean differences between treatments. When the null hypothesis is rejected, it indicates only that there is a difference between at least two of the treatment means. If $k = 2$, it is obvious which two treatments are different. However, when k is greater than 2, the situation becomes more complex. To determine exactly where significant differences exist, the researcher must follow the ANOVA with post hoc tests. In Chapter 12, we used Tukey's HSD and the Scheffé test to make these multiple comparisons among treatment means. These two procedures attempt to control the overall alpha level by making adjustments for the number of potential comparisons.

For a repeated-measures ANOVA, Tukey's HSD and the Scheffé test can be used in the exact same manner as was done for the independent-measures ANOVA, *provided* that you substitute MS_{error} in place of $MS_{\text{within treatments}}$ in the formulas and use df_{error} in place of $df_{\text{within treatments}}$ when locating the critical value in a statistical table. Note that statisticians are not in complete agreement about the appropriate error term in post hoc tests for repeated-measures designs (for a discussion, see Keppel, 1973, or Keppel & Zedeck, 1989).

ASSUMPTIONS OF THE REPEATED-MEASURES ANOVA

The basic assumptions for the repeated-measures ANOVA are identical to those required for the independent-measures ANOVA.

1. The observations within each treatment condition must be independent (see p. 254).
2. The population distribution within each treatment must be normal. (As before, the assumption of normality is important only with small samples.)
3. The variances of the population distributions for each treatment should be equivalent.

For the repeated-measures ANOVA, there is an additional assumption, called homogeneity of covariance. Basically, it refers to the requirement that the relative standing of each subject be maintained in each treatment condition. This assumption is violated if the effect of the treatment is not consistent for all of the subjects or if order effects exist for some, but not other, subjects. This issue is very complex and is beyond the scope of this book. However, methods do exist for dealing with violations of this assumption (for a discussion, see Keppel, 1973).

If there is reason to suspect that one of the assumptions for the repeated-measures ANOVA has been violated, an alternative analysis known as the Friedman test can be used. The Friedman test is presented in Appendix E. It requires that the original scores be transformed into ranks before evaluating the differences between treatment conditions.

LEARNING CHECK

1. Explain how SS_{error} is computed in the repeated-measures ANOVA.
2. A repeated-measures study is used to evaluate the mean differences among three treatment conditions using a sample of $n = 8$ participants. What are the df values for the F-ratio?
3. For the following data, compute $SS_{between\ treatments}$ and $SS_{between\ subjects}$.

	Treatment				
Subject	1	2	3	4	
A	2	2	2	2	$G = 32$
B	4	0	0	4	$\Sigma X^2 = 96$
C	2	0	2	0	
D	4	2	2	4	
	$T = 12$ $SS = 4$	$T = 4$ $SS = 4$	$T = 6$ $SS = 3$	$T = 10$ $SS = 11$	

4. A research report includes a repeated-measures F-ratio with $df = 3, 24$. How many treatment conditions were compared, and how many individuals participated in the study? (See Box 13.2.)

ANSWERS

1. $SS_{error} = SS_{within} - SS_{between\ subjects}$ Variability from individual differences is subtracted from the within-treatments variability.
2. $df = 2, 14$
3. $SS_{between\ treatments} = 10$, $SS_{between\ subjects} = 8$
4. There were 4 treatment conditions ($k - 1 = 3$) and 9 participants ($n - 1 = 8$).

13.4 ADVANTAGES AND DISADVANTAGES OF THE REPEATED-MEASURES DESIGN

When we first encountered the repeated-measure design (Chapter 11), we noted that this type of research study has certain advantages and disadvantages (pp. 366−369). On the bright side, a repeated-measures study may be desirable if the supply of participants is limited. A repeated-measures study is economical in that the research requires relatively few participants. Also, a repeated-measures design eliminates or minimizes most of the problems associated with individual differences. However, disadvantages also exist. These take the form of order effects, such as fatigue, that can make the interpretation of the data difficult.

Now that we have introduced the repeated-measures ANOVA, we can examine one of the primary advantages of this design—namely, the elimination of variability caused by individual differences. Consider the structure of the *F*-ratio for both the independent-and the repeated-measures designs.

$$F = \frac{\text{treatment effects} + \text{random, unsystematic differences}}{\text{random, unsystematic differences}}$$

In each case, the goal of the analysis is to determine whether the data provide evidence for a treatment effect. If there is no treatment effect, then the numerator and denominator are both measuring the same random, unsystematic variance and the *F*-ratio should produce a value near 1.00. On the other hand, the existence of a treatment effect should make the numerator substantially larger than the denominator and result in a large value for the *F*-ratio.

For the independent-measures design, the unsystematic differences include individual differences as well as other random sources of error. Thus, for the independent-measures ANOVA, the *F*-ratio has the following structure:

$$F = \frac{\text{treatment effect} + \text{individual differences and other error}}{\text{individual differences and other error}}$$

For the repeated-measures design, the individual differences are eliminated or subtracted out, and the resulting *F*-ratio is structured as follows:

$$F = \frac{\text{treatment effect} + \text{error (excluding individual differences)}}{\text{error (excluding individual differences)}}$$

The removal of individual differences from the analysis becomes an advantage in situations in which very large individual differences exist among the participants being studied.

When individual differences are large, the presence of a treatment effect may be masked if an independent-measures study is performed. In this case, a repeated-measures design would be more sensitive in detecting a treatment effect because individual differences do not influence the value of the *F*-ratio.

This point will become evident in the following example. Suppose that we know how much variability is accounted for by the different sources of variance. For example,

$$\text{treatment effect} = 10 \text{ units of variance}$$

$$\text{individual differences} = 10 \text{ units of variance}$$

$$\text{other error} = 1 \text{ unit of variance}$$

Notice that a large amount of the variability in the experiment is caused by individual differences. By comparing the F-ratios for an independent- and a repeated-measures analysis, we are able to see a fundamental difference between the two types of experimental designs. For an independent-measures experiment, we obtain

$$F = \frac{\text{treatment effect} + \text{individual differences} + \text{error}}{\text{individual differences} + \text{error}}$$
$$= \frac{10+10+1}{10+1} = \frac{21}{11} = 1.91$$

Thus, the independent-measures ANOVA produces an F-ratio of $F = 1.91$. Recall that the F-ratio is structured to produce $F = 1.00$ if there is no treatment effect whatsoever. In this case, the F-ratio is near to 1.00 and strongly suggests that there is little or no treatment effect. If you check the F-distribution table in Appendix B, you will find that it is almost impossible for an F-ratio as small as 1.91 to be significant. For the independent-measures ANOVA, the 10-point treatment effect is overwhelmed by all of the other variance.

Now consider what happens with a repeated-measures ANOVA. With the individual differences removed, the F-ratio becomes:

$$F = \frac{\text{treatment effect} + \text{error}}{\text{error}}$$
$$= \frac{10 + 1}{1} = \frac{11}{1} = 11$$

For the repeated-measures ANOVA, the numerator of the F-ratio (which includes the treatment effect) is 11 times larger than the denominator (which has no treatment effect). This result strongly indicates that there is a substantial treatment effect. In this example, the F-ratio is much larger for the repeated-measures study because the individual differences, which are extremely large, have been removed. In the independent-measures ANOVA, the presence of a treatment effect is obscured by the influence of individual differences. This problem is eliminated by the repeated-measures design, in which variability caused by individual differences is partitioned out of the analysis. When the individual differences are large, a repeated-measures experiment may provide a more sensitive test for a treatment effect. In statistical terms, a repeated-measures test has more *power* than an independent-measures test; that is, it is more likely to detect a real treatment effect.

INDIVIDUAL DIFFERENCES AND THE CONSISTENCY OF THE TREATMENT EFFECTS

As we have demonstrated, one major advantage of a repeated-measures design is that it removes individual differences from the denominator of the F-ratio, which usually increases the likelihood of obtaining a significant result. However, removing individual differences is an advantage only when the treatment effects are reasonably consistent for all of the participants. If the treatment effects are not consistent across participants, the individual differences tend to disappear and value in the denominator is not noticeably reduced by removing them. This phenomenon is demonstrated in the following example.

EXAMPLE 13.2

Table 13.4 presents hypothetical data from a repeated-measures research study. We constructed the data specifically to demonstrate the relationship between consistent treatment effects and large individual differences.

TABLE 13.4

Data from a repeated-measures study comparing three treatments. The data show consistent treatment effects from one participant to another, which produce consistent and relatively large differences in the individual *P* totals.

	Treatment			
Person	I	II	III	
A	0	1	2	$P = 3$
B	1	2	3	$P = 6$
C	2	4	6	$P = 12$
D	3	5	7	$P = 15$
	$T = 6$	$T = 12$	$T = 18$	
	$SS = 5$	$SS = 10$	$SS = 17$	

First, notice the consistency of the treatment effects. Treatment II has the same effect on every participant, increasing everyone's score by 1 or 2 points compared to treatment I. Also, treatment III produces a consistent increase of 1 or 2 points compared to treatment II. One consequence of the consistent treatment effects is that the individual differences are maintained in all of the treatment conditions. For example, participant A has the lowest score in all three treatments, and participant D always has the highest score. The participant totals (*P* values) reflect the consistent differences. For example, participant D has the largest score in every treatment and, therefore, has the largest *P* value. Also notice that there are big differences between the *P* totals from one individual to the next. For these data, $SS_{\text{between subjects}} = 30$ points.

Now consider the data in Table 13.5. To construct these data we started with the same numbers within each treatment that were used in Table 13.4. However, we scrambled the numbers within each column to eliminate the consistency of the treatment effects. In Table 13.5, for example, two participants show an increase in scores as they go from treatment I to treatment II, and two show a decrease. The data also show an inconsistent treatment effect as the participants go from treatment II to treatment III. One consequence of the inconsistent treatment effects is that there are no consistent individual differences between participants. Participant C, for example, has the lowest score in treatment II and the highest score in treatment III. As a result, there are no longer consistent differences between the individual participants. All of the *P* totals are about the same. For these data, $SS_{\text{between subjects}} = 3.33$ points. Because the two sets of data (Tables 13.4 and 13.5) have the same treatment totals (*T* values) and *SS* values, they have the same $SS_{\text{between treatments}}$ and $SS_{\text{within treatments}}$. For both sets of data,

$$SS_{\text{between treatments}} = 18 \text{ and } SS_{\text{within treatments}} = 32$$

However, there is a huge difference between the two sets of data when you compute SS_{error} for the denominator of the *F*-ratio. For the data in Table 13.4, with consistent treatment effects and large individual differences,

$$\begin{aligned} SS_{\text{error}} &= SS_{\text{within treatments}} - SS_{\text{between subjects}} \\ &= 32 - 30 \\ &= 2 \end{aligned}$$

For the data in Table 13.5, with no consistent treatment effects and relatively small differences between the individual *P* totals,

$$\begin{aligned} SS_{\text{error}} &= SS_{\text{within treatments}} - SS_{\text{between subjects}} \\ &= 32 - 3.33 \\ &= 28.67 \end{aligned}$$

TABLE 13.5

Data from a repeated-measures study comparing three treatments. The data show treatment effects that are inconsistent from one participant to another and, as a result, produce relatively small differences in the individual *P* totals. Note that the data have exactly the same scores within each treatment as the data in Table 13.5, however, the scores have been scrambled to eliminate the consistency of the treatment effects.

	Treatment			
Person	I	II	III	
A	0	4	3	$P = 7$
B	1	5	2	$P = 8$
C	2	1	7	$P = 10$
D	3	2	6	$P = 11$
	$T = 6$	$T = 12$	$T = 18$	
	$SS = 5$	$SS = 10$	$SS = 17$	

Thus, consistent treatment effects tend to produce a relatively small error term for the *F*-ratio. As a result, consistent treatment effects are more likely to be statistically significant (reject the null hypothesis). For the examples we have been considering, the data in Table 13.4 produce an *F*-ratio of $F = 27.0$. With $df = 2, 6$, this *F*-ratio is well into the critical region for $\alpha = .05$ or .01 and we conclude that there are significant differences among the three treatments. On the other hand, the same mean differences in Table 13.5 produce $F = 1.88$. With $df = 2, 6$, this value is not in the critical region for $\alpha = .05$ or .01, and we conclude that there are no significant differences.

In summary, when treatment effects are consistent from one individual to another, the individual differences also tend to be consistent and relatively large. The large individual differences get subtracted from the denominator of the *F*-ratio producing a larger value for *F* and increasing the likelihood that the *F*-ratio will be in the critical region.

13.5 REPEATED-MEASURES ANOVA AND REPEATED-MEASURES *t* TEST

As we noted in Chapter 12 (pp. 420–421), whenever you are evaluating the difference between two sample means, you can use either a *t* test or ANOVA. In Chapter 12, we demonstrated that the two tests are related in many respects, including:

1. The two tests always reach the same conclusion about the null hypothesis.
2. The basic relationship between the two test statistics is $F = t^2$.
3. The *df* value for the *t* statistic is identical to the *df* value for the denominator of the *F*-ratio.
4. If you square the critical value for the two-tailed *t* test, you obtain the critical value for the *F*-ratio. Again, the basic relationship is $F = t^2$.

In Chapter 12, these relationships were demonstrated for the independent-measures tests, but they are also true for repeated-measures designs comparing two treatment conditions. The following example demonstrates the relationships.

EXAMPLE 13.3

The following table shows the data from a repeated-measures study comparing two treatment conditions. We have structured the data in a format that is compatible with the repeated-measures *t* test. Note that the calculations for the *t* test are based on the difference scores (*D* values) in the final column.

	Treatment			
Participant	I	II	D	
A	3	5	2	
B	4	14	10	
C	5	7	2	
D	4	6	2	
				$M_D = 4$
				$SS_D = 48$

The repeated-measures *t* test The null hypothesis for the *t* test states that, for the general population, there is no mean difference between the two treatment conditions.

$$H_0: \mu_D = 0$$

With $n = 4$ participants, the test has $df = 3$ and the critical boundaries for a two-tailed test with $\alpha = .05$ are $t = \pm 3.182$.

For these data, the sample mean difference is $M_D = 4$, the variance for the difference scores is $s^2 = 16$, and the standard error is $s_{M_D} = 2$ points. These values produce a *t* statistic of

$$t = \frac{M_D - \mu_D}{s_{M_D}} = \frac{4 - 0}{2} = 2.00$$

The *t* value is not in the critical region so we fail to reject H_0 and conclude that there is no significant difference between the two treatments.

The repeated-measures ANOVA Now we reorganize the data into a format that is compatible with a repeated-measures ANOVA. Notice that the ANOVA uses the original scores (not the difference scores) and requires the *P* totals for each participant.

	Treatment			
Participant	I	II	*P*	
A	3	5	8	$G = 48$
B	4	14	18	$\Sigma X^2 = 372$
C	5	7	12	$N = 8$
D	4	6	10	

Again, the null hypothesis states that, for the general population, there is no mean difference between the two treatment conditions.

$$H_0: \mu_1 = \mu_2$$

For this study, $df_{\text{between treatments}} = 1$, $df_{\text{within treatments}} = 6$, $df_{\text{between subjects}} = 3$, which produce $df_{\text{error}} = (6 - 3) = 3$. Thus, the F-ratio has $df = 1, 3$ and the critical value for $\alpha = .05$ is $F = 10.13$. Note that the denominator of the F-ratio has the same df value as the t statistic ($df = 3$) and that the critical value for F is equal to the squared critical value for t ($10.13 = 3.182^2$).

$$\begin{aligned} \text{For these data, } SS_{\text{total}} &= 84, \\ SS_{\text{within}} &= 52 \\ SS_{\text{between treatments}} &= (84 - 52) = 32 \\ SS_{\text{between subjects}} &= 28 \\ SS_{\text{error}} &= (52 - 28) = 24 \end{aligned}$$

The two variances in the F-ratio are

$$MS_{\text{between treatments}} = \frac{SS_{\text{between treatments}}}{df_{\text{between treatments}}} = \frac{32}{1} = 32$$

$$\text{and } MS_{\text{error}} = \frac{SS_{\text{error}}}{df_{\text{error}}} = \frac{24}{3} = 8$$

$$\text{and the } F\text{-ratio is } F = \frac{MS_{\text{between treatments}}}{MS_{\text{error}}} = \frac{32}{8} = 4.00$$

Notice that the F-ratio and the t statistic are related by the equation $F = t^2$ ($4 = 2^2$). The F-ratio (like the t statistic) is not in the critical region so, once again, we fail to reject H_0 and conclude that there is no significant difference between the two treatments.

SUMMARY

1. The repeated-measures ANOVA is used to evaluate the mean differences obtained in a research study comparing two or more treatment conditions using the same sample of individuals in each condition. The test statistic is an F-ratio, in which the numerator measures the variance (differences) between treatments and the denominator measures the variance (differences) that is expected without any treatment effects or individual differences.

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{error}}}$$

2. The first stage of the repeated-measures ANOVA is identical to the independent-measures ANOVA and separates the total variability into two components: between-treatments and within-treatments. Because a repeated-measures design uses the same subjects in every treatment condition, the differences between treatments cannot be caused by individual differences. Thus, individual differences are automatically eliminated from the between-treatments variance in the numerator of the F-ratio.

3. In the second stage of the repeated-measures analysis, individual differences are computed and removed from the denominator of the F-ratio. To remove the individual differences, you first compute the variability between subjects (SS and df) and then subtract these values from the corresponding within-treatments values. The residual provides a measure of error excluding individual differences, which is the appropriate denominator for the repeated-measures F-ratio. The equations for analyzing SS and df for the repeated-measures ANVOA are presented in Figure 13.3.

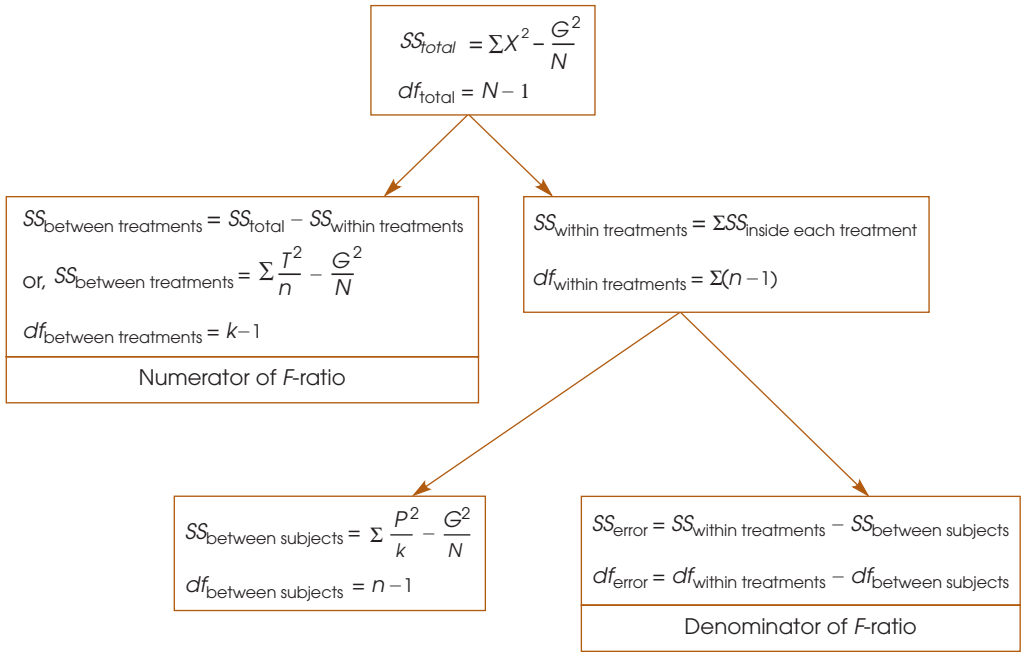


FIGURE 13.3
Formulas for the repeated-measures ANOVA.

4. Effect size for the repeated-measures ANOVA is measured by computing eta squared, the percentage of variance accounted for by the treatment effect. For the repeated-measures ANOVA

$$\eta^2 = \frac{SS_{between\ treatments}}{SS_{total} - SS_{between\ subjects}}$$

$$= \frac{SS_{between\ treatments}}{SS_{between\ treatments} + SS_{error}}$$

Because part of the variability (the *SS* caused by individual differences) is removed before computing η^2, this measure of effect size is often called a partial eta squared.

5. When the obtained *F*-ratio is significant (that is, H_0 is rejected), it indicates that a significant difference lies between at least two of the treatment conditions. To determine exactly where the difference lies, post hoc comparisons may be made. Post hoc tests, such as Tukey's HSD, use MS_{error} rather than $MS_{within\ treatments}$ and df_{error} instead of $df_{within\ treatments}$.

6. A repeated-measures ANOVA eliminates the influence of individual differences from the analysis. If individual differences are extremely large, then a treatment effect might be masked in an independent-measures experiment. In this case, a repeated-measures design might be a more sensitive test for a treatment effect.

KEY TERMS

individual differences (437)
between-treatments variance(437)
error variance (438)
between-subjects variance (440)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 13 on the book companion website.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform the Single-Factor, **Repeated-Measures Analysis of Variance (ANOVA)** presented in this chapter.

Data Entry

Enter the scores for each treatment condition in a separate column, with the scores for each individual in the same row. All of the scores for the first treatment go in the VAR00001 column, the second treatment scores go in the VAR00002 column, and so on.

Data Analysis

1. Click **Analyze** on the tool bar, select **General Linear Model,** and click on **Repeated-Measures.**
2. SPSS presents a box entitled **Repeated-Measures Define Factors.** Within the box, the Within-Subjects Factor Name should already contain **Factor 1.** If not, type in Factor 1.
3. Enter the **Number of levels** (number of different treatment conditions) in the next box.
4. Click **Add.**
5. Click **Define.**
6. One by one, move the column labels for your treatment conditions into the **Within Subjects Variables** box. (Highlight the column label on the left and click the arrow to move it into the box.)
7. If you want descriptive statistics for each treatment, click on the **Options** box, select **Descriptives,** and click **Continue.**
8. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the television viewing study in Example 13.1 and portions of the program output are shown in Figure 13.4. Note that large portions of the SPSS output are not relevant for our purposes and are not included in Figure 13.1. The first item of interest is the table of **Descriptive Statistics,** which presents the mean, standard deviation, and number of scores for each treatment. Next, we skip to the table showing **Tests of Within-Subjects Effects.** The top line of the factor 1 box (Sphericity Assumed) shows the between-treatments sum of squares, degrees of freedom, and mean square that form the numerator of the *F*-ratio. The same line reports the value of the *F*-ratio and the level of significance (the *p* value or alpha level). Similarly, the top line of the Error (factor 1) box shows the sum of squares, the degrees of freedom, and the mean square for the error term (the denominator of the *F*-ratio). The final box in the output (not shown in Figure 13.4) is labeled **Tests of Between-Subjects Effects** and the bottom line (Error) reports the between-subjects sum of squares and degrees of freedom (ignore the mean square and *F*-ratio, which are not part of the repeated-measures ANOVA).

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00001	1.0000	1.41421	5
VAR00002	2.0000	1.41421	5
VAR00003	5.0000	1.58114	5
VAR00004	4.0000	1.22474	5

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
factor 1	Sphericity Assumed	50.000	3	16.667	25.000	.000
	Greenhouse-Geisser	50.000	1.600	31.250	25.000	.001
	Huynh-Feldt	50.000	2.500	20.000	25.000	.000
	Lower-bound	50.000	1.000	50.000	25.000	.007
Error (factor 1)	Sphericity Assumed	8.000	12	.667		
	Greenhouse-Geisser	8.000	6.400	1.250		
	Huynh-Feldt	8.000	10.000	.800		
	Lower-bound	8.000	4.000	2.000		

FIGURE 13.4

Portions of the SPSS output for the repeated-measures ANOVA for the television viewing study in Example 13.1.

FOCUS ON PROBLEM SOLVING

1. Before you begin a repeated-measures ANOVA, complete all of the preliminary calculations needed for the ANOVA formulas. This requires that you find the total for each treatment (Ts), the total for each person (Ps), the grand total (G), the SS for each treatment condition, and ΣX^2 for the entire set of N scores. As a partial check on these calculations, be sure that the T values add up to G and that the P values have a sum of G.
2. To help remember the structure of repeated-measures ANOVA, keep in mind that a repeated-measures experiment eliminates the contribution of individual differences. There are no individual differences contributing to the numerator of the F-ratio ($MS_{\text{between treatments}}$) because the same individuals are used for all treatments. Therefore, you must also eliminate individual differences in the denominator. This is accomplished by partitioning within-treatments variability into two components: between-subjects variability and error variability. It is the MS value for error variability that is used in the denominator of the F-ratio.

DEMONSTRATION 13.1

REPEATED-MEASURES ANOVA

The following data were obtained from a research study examining the effect of sleep deprivation on motor-skills performance. A sample of five participants was tested on a motor-skills task after 24 hours of sleep deprivation, tested again after 36 hours, and tested once more after 48 hours. The dependent variable is the number of errors made on the motor-skills task. Do these data indicate that the number of hours of sleep deprivation has a significant effect on motor skills performance?

Participant	24 Hours	36 Hours	48 Hours	P totals	
A	0	0	6	6	$N = 15$
B	1	3	5	9	$G = 45$
C	0	1	5	6	$\Sigma X^2 = 245$
D	4	5	9	18	
E	0	1	5	6	
	$T = 5$	$T = 10$	$T = 30$		
	$SS = 12$	$SS = 16$	$SS = 12$		

STEP 1 **State the hypotheses, and specify alpha.** The null hypothesis states that, for the general population, there are no differences among the three deprivation conditions. Any differences that exist among the samples are simply the result of chance or error. In symbols,

$$H_0: \mu_1 = \mu_2 = \mu_3$$

The alternative hypothesis states that there are differences among the conditions.

H_1: At least one of the treatment means is different.

We use $\alpha = .05$.

STEP 2 **The repeated-measures analysis.** Rather than compute the *df* values and look for a critical value for *F* at this time, we proceed directly to the ANOVA.

STAGE 1 The first stage of the analysis is identical to the independent-measures ANOVA presented in Chapter 12.

$$SS_{\text{total}} = \Sigma X^2 - \frac{G^2}{N} = 245 - \frac{45^2}{15} = 110$$

$$SS_{\text{within}} = \Sigma SS_{\text{inside each treatment}} = 12 + 16 + 12 = 40$$

$$SS_{\text{between}} = \Sigma \frac{T^2}{n} - \frac{G^2}{N} = \frac{5^2}{5} + \frac{10^2}{5} + \frac{30^2}{5} - \frac{45^2}{15} = 70$$

and the corresponding degrees of freedom are

$$df_{\text{total}} = N = -1 = 14$$

$$df_{\text{within}} = \Sigma df = 4 + 4 + 4 = 12$$

$$df_{\text{between}} = k - 1 = 2$$

STAGE 2 The second stage of the repeated-measures analysis measures and removes the individual differences from the denominator of the *F*-ratio.

$$\begin{aligned} SS_{\text{between subjects}} &= \Sigma \frac{P^2}{k} - \frac{G^2}{N} \\ &= \frac{6^2}{3} + \frac{9^2}{3} + \frac{6^2}{3} + \frac{18^2}{3} + \frac{6^2}{3} - \frac{45^2}{15} \\ &= 36 \end{aligned}$$

$$\begin{aligned} SS_{\text{error}} &= SS_{\text{within}} = SS_{\text{between subjects}} \\ &= 40 - 36 \\ &= 4 \end{aligned}$$

and the corresponding *df* values are

$$df_{\text{between subjects}} = n - 1 = 4$$

$$\begin{aligned} df_{\text{error}} &= df_{\text{within}} - df_{\text{between subjects}} \\ &= 12 - 4 \\ &= 8 \end{aligned}$$

The mean square values that form the *F*-ratio are as follows:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{70}{2} = 35$$

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{df_{\text{error}}} = \frac{4}{8} = 0.50$$

Finally, the F-ratio is

$$F = \frac{MS_{\text{between}}}{MS_{\text{error}}} = \frac{35}{0.50} = 70.00$$

STEP 3 **Make a decision and state a conclusion.** With $df = 2, 8$ and $\alpha = .05$, the critical value is $F = 4.46$. Our obtained F-ratio ($F = 70.00$) is well into the critical region, so our decision is to reject the null hypothesis and conclude that there are significant differences among the three levels of sleep deprivation.

DEMONSTRATION 13.2

EFFECT SIZE FOR THE REPEATED-MEASURES ANOVA

We compute η^2, the percentage of variance explained by the treatment differences, for the data in Demonstration 13.1. Using Equation 13.11 we obtain

$$\eta^2 = \frac{SS_{\text{between treatments}}}{SS_{\text{between treatments}} + SS_{\text{error}}} = \frac{70}{70 + 4} = \frac{70}{74} = 0.95 \quad \text{(or 95\%)}$$

PROBLEMS

1. How does the denominator of the F-ratio (the error term) differ for a repeated-measures ANOVA compared to an independent-measures ANOVA?
2. The repeated-measures ANOVA can be viewed as a two-stage process. What is the purpose of the second stage?
3. A researcher conducts an experiment comparing three treatment conditions with $n = 10$ scores in each condition.
   a. If the researcher uses an independent-measures design, how many individuals are needed for the study and what are the df values for the F-ratio?
   b. If the researcher uses a repeated-measures design, how many individuals are needed for the study and what are the df values for the F-ratio?
4. A researcher conducts a repeated-measures experiment using a sample of $n = 8$ subjects to evaluate the differences among four treatment conditions. If the results are examined with an ANOVA, what are the df values for the F-ratio?
5. A researcher uses a repeated-measures ANOVA to evaluate the results from a research study and reports an F-ratio with $df = 2, 30$.
   a. How many treatment conditions were compared in the study?
   b. How many individuals participated in the study?
6. A published report of a repeated-measures research study includes the following description of the statistical analysis. "The results show significant differences among the treatment conditions, $F(2, 20) = 6.10$, $p < .01$."
   a. How many treatment conditions were compared in the study?
   b. How many individuals participated in the study?
7. The following data were obtained from a repeated-measures study comparing three treatment conditions. Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether there are significant mean differences among the three treatments.

	Treatments				
Person	I	II	III	Person Totals	
A	0	4	2	$P = 6$	
B	1	5	6	$P = 12$	$N = 18$
C	3	3	3	$P = 9$	$G = 48$
D	0	1	5	$P = 6$	$\Sigma X^2 = 184$
E	0	2	4	$P = 6$	
F	2	3	4	$P = 9$	
	$M = 1$	$M = 3$	$M = 4$		
	$T = 6$	$T = 18$	$T = 24$		
	$SS = 8$	$SS = 10$	$SS = 10$		

8. The following data were obtained from a repeated-measures study comparing two treatment conditions. Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether there are significant mean differences between the two treatments.

	Treatments			
Person	I	II	Person Totals	
A	3	5	$P = 8$	
B	5	9	$P = 14$	$N = 16$
C	1	5	$P = 6$	$G = 80$
D	1	7	$P = 8$	$\Sigma X^2 = 500$
E	5	9	$P = 14$	
F	3	7	$P = 10$	
G	2	6	$P = 8$	
H	4	8	$P = 12$	
	$M = 3$	$M = 7$		
	$T = 24$	$T = 56$		
	$SS = 18$	$SS = 18$		

9. The following data were obtained from a repeated-measures study comparing three treatment conditions.

a. Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether there are significant mean differences among the three treatments.

b. Compute η^2, the percentage of variance accounted for by the mean differences, to measure the size of the treatment effects.

c. Write a sentence demonstrating how a research report would present the results of the hypothesis test and the measure of effect size.

	Treatments				
Person	I	II	III	Person Totals	
A	1	1	4	$P = 6$	
B	3	4	8	$P = 15$	$N = 15$
C	0	2	7	$P = 9$	$G = 45$
D	0	0	6	$P = 6$	$\Sigma X^2 = 231$
E	1	3	5	$P = 9$	
	$M = 1$	$M = 2$	$M = 6$		
	$T = 5$	$T = 10$	$T = 30$		
	$SS = 6$	$SS = 10$	$SS = 10$		

10. For the data in problem 9,

a. Compute SS_{total} and $SS_{between\ treatments}$.

b. Eliminate the mean differences between treatments by adding 2 points to each score in treatment I, adding 1 point to each score in treatment II, and subtracting 3 points from each score in treatment III. (All three treatments should end up with $M = 3$ and $T = 15$.)

c. Calculate SS_{total} for the modified scores. (*Caution:* You first must find the new value for ΣX^2.)

d. Because the treatment effects were eliminated in part b, you should find that SS_{total} for the modified scores is smaller than SS_{total} for the original scores. The difference between the two SS values should be exactly equal to the value of $SS_{between\ treatments}$ for the original scores.

11. The following data were obtained from a repeated-measures study comparing three treatment conditions.

	Treatment				
Subject	I	II	III	P	
A	6	8	10	24	$G = 48$
B	5	5	5	15	$\Sigma X^2 = 294$
C	1	2	3	6	
D	0	1	2	3	
	$T = 12$	$T = 16$	$T = 20$		
	$SS = 26$	$SS = 30$	$SS = 38$		

Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether these data are sufficient to demonstrate significant differences between the treatments.

12. In Problem 11 the data show large and consistent differences between subjects. For example, subject A has the largest score in every treatment and subject D always has the smallest score. In the second stage of the ANOVA, the large individual differences are subtracted out of the denominator of the F-ratio, which results in a larger value for F.

The following data were created by using the same numbers that appeared in Problem 11. However, we eliminated the consistent individual differences by scrambling the scores within each treatment.

	Treatment				
Subject	I	II	III	P	
A	6	2	3	11	$G = 48$
B	5	1	5	11	$\Sigma X^2 = 294$
C	0	5	10	15	
D	1	8	2	11	
	$T = 12$	$T = 16$	$T = 20$		
	$SS = 26$	$SS = 30$	$SS = 38$		

a. Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether these data are sufficient to demonstrate significant differences between the treatments.

b. Explain how the results of this analysis compare with the results from Problem 11.

13. One of the primary advantages of a repeated-measures design, compared to an independent-measures design, is that it reduces the overall variability by removing variance caused by individual differences. The following data are from a research study comparing three treatment conditions.

a. Assume that the data are from an independent-measures study using three separate samples, each with $n = 6$ participants. Ignore the column of P totals and use an independent-measures ANOVA with $\alpha = .05$ to test the significance of the mean differences.

b. Now assume that the data are from a repeated-measures study using the same sample of $n = 6$ participants in all three treatment conditions. Use a repeated-measures ANOVA with $\alpha = .05$ to test the significance of the mean differences.

c. Explain why the two analyses lead to different conclusions.

Treatment 1	Treatment 2	Treatment 3	P	
6	9	12	27	
8	8	8	24	$N = 18$
5	7	9	21	$G = 108$
0	4	8	12	$\Sigma X^2 = 800$
2	3	4	9	
3	5	7	15	
$M = 4$	$M = 6$	$M = 8$		
$T = 24$	$T = 36$	$T = 48$		
$SS = 42$	$SS = 28$	$SS = 34$		

14. The following data are from an experiment comparing three different treatment conditions:

A	B	C	
0	1	2	$N = 15$
2	5	5	$\Sigma X^2 = 354$
1	2	6	
5	4	9	
2	8	8	
$T = 10$	$T = 20$	$T = 30$	
$SS = 14$	$SS = 30$	$SS = 30$	

a. If the experiment uses an *independent-measures design,* can the researcher conclude that the treatments are significantly different? Test at the .05 level of significance.

b. If the experiment is done with a *repeated-measures design,* should the researcher conclude that the treatments are significantly different? Set alpha at .05 again.

c. Explain why the analyses in parts a and b lead to different conclusions.

15. A researcher is evaluating customer satisfaction with the service and coverage of two phone carriers. Each individual in a sample of $n = 25$ uses one carrier for two weeks and then switches to the other. Each participant then rates the two carriers. The following table presents the results from the repeated-measures ANOVA comparing the average ratings. Fill in the missing values in the table. (*Hint:* Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	____	____	2	$F =$ ____
Within treatments	____	____		
Between subjects	____	____		
Error	12	____	____	
Total	23	____		

16. The following summary table presents the results from a repeated-measures ANOVA comparing three treatment conditions with a sample of $n = 11$ subjects. Fill in the missing values in the table. (*Hint:* Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	____	____	____	$F = 5.00$
Within treatments	80	____		
Between subjects	____	____		
Error	60	____	____	
Total	____	____		

17. The following summary table presents the results from a repeated-measures ANOVA comparing four treatment conditions, each with a sample of $n = 12$ participants. Fill in the missing values in the table. (*Hint:* Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	54	____	20	$F =$ ____
Within treatments	____	____		
Between subjects	____	____		
Error	____	____	3	
Total	194	____		

18. A recent study indicates that simply giving college students a pedometer can result in increased walking (Jackson & Howton, 2008). Students were given pedometers for a 12-week period, and asked to record the average number of steps per day during weeks 1, 6, and 12. The following data are similar to the results obtained in the study.

Number of steps (x1000)					
	Week				
Participant	1	6	12	*P*	
A	6	8	10	24	
B	4	5	6	15	
C	5	5	5	15	$G = 72$
D	1	2	3	6	$\Sigma X^2 = 400$
E	0	1	2	3	
F	2	3	4	9	
	$T = 18$	$T = 24$	$T = 30$		
	$SS = 28$	$SS = 32$	$SS = 40$		

a. Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether the mean number of steps changes significantly from one week to another.

b. Compute η^2 to measure the size of the treatment effect.

c. Write a sentence demonstrating how a research report would present the results of the hypothesis test and the measure of effect size.

19. A repeated-measures experiment comparing only two treatments can be evaluated with either a *t* statistic or an ANOVA. As we found with the independent-measures design, the *t* test and the ANOVA produce equivalent conclusions, and the two test statistics are related by the equation $F = t^2$. The following data are from a repeated-measures study:

Subject	Treatment 1	Treatment 2	Difference
1	2	4	+2
2	1	3	+2
3	0	10	+10
4	1	3	+2

a. Use a repeated-measures *t* statistic with $\alpha = .05$ to determine whether the data provide evidence of a significant difference between the two treatments. (*Caution:* ANOVA calculations are done with the *X* values, but for *t* you use the difference scores.)

b. Use a repeated-measures ANOVA with $\alpha = .05$ to evaluate the data. (You should find $F = t^2$.)

20. For either independent-measures or repeated-measures designs comparing two treatments, the mean difference can be evaluated with either a *t* test or an ANOVA. The two tests are related by the equation $F = t^2$. For the following data,

a. Use a repeated-measures *t* test with $\alpha = .05$ to determine whether the mean difference between treatments is statistically significant.

b. Use a repeated-measures ANOVA with $\alpha = .05$ to determine whether the mean difference between treatments is statistically significant. (You should find that $F = t^2$.)

Person	Treatment 1	Treatment 2	Difference
A	4	7	3
B	2	11	9
C	3	6	3
D	7	10	3
	$M = 4$	$M = 8.5$	$M_D = 4.5$
	$T = 16$	$T = 34$	
	$SS = 14$	$SS = 17$	$SS = 27$

21. In the Preview section for this chapter, we presented an example of a delayed discounting study in which people are willing to settle for a smaller reward today in exchange for a larger reward in the future. The following data represent the typical results from one of these studies. The participants are asked how much they would take today instead of waiting for a specific delay period to receive $1000. Each participant responds to all 5 of the delay periods. Use a repeated-measures ANOVA with $\alpha = .01$ to determine whether there are significant differences among the 5 delay periods for the following data:

Participant	1 month	6 months	1 year	2 years	5 years
A	950	850	800	700	550
B	800	800	750	700	600
C	850	750	650	600	500
D	750	700	700	650	550
E	950	900	850	800	650
F	900	900	850	750	650

22. The endorphins released by the brain act as natural painkillers. For example, Gintzler (1970) monitored endorphin activity and pain thresholds in pregnant rats during the days before they gave birth. The data showed an increase in pain threshold as the pregnancy progressed. The change was gradual until 1 or 2 days

before birth, at which point there was an abrupt increase in pain threshold. Apparently a natural painkilling mechanism was preparing the animals for the stress of giving birth. The following data represent pain-threshold scores similar to the results obtained by Gintzler. Do these data indicate a significant change in pain threshold? Use a repeated-measures ANOVA with $\alpha = .01$.

	Days Before Giving Birth			
Subject	7	5	3	1
A	39	40	49	52
B	38	39	44	55
C	44	46	50	60
D	40	42	46	56
E	34	33	41	52

CHAPTER

14

Two-Factor Analysis of Variance (Independent Measures)

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Introduction to analysis of variance (Chapter 12)
  - The logic of analysis of variance
  - ANOVA notation and formulas
  - Distribution of *F*-ratios

Preview

Imagine that you are seated at your desk, ready to take the final exam in statistics. Just before the exams are handed out, a television crew appears and sets up a camera and lights aimed directly at you. They explain that they are filming students during exams for a television special. You are told to ignore the camera and go ahead with your exam.

Would the presence of a TV camera affect your performance on an exam? For some of you, the answer to this question is "definitely yes" and for others, "probably not." In fact, both answers are right; whether the TV camera affects performance depends on your personality. Some of you would become terribly distressed and self-conscious, while others really could ignore the camera and go on as if everything were normal.

In an experiment that duplicates the situation we have described, Shrauger (1972) tested participants on a concept-formation task. Half of the participants worked alone (no audience), and half worked with an audience of people who claimed to be interested in observing the experiment. Shrauger also divided the participants into two groups on the basis of personality: those high in self-esteem and those low in self-esteem. The dependent variable for this experiment was the number of errors on the concept formation task. Data similar to those obtained by Shrauger are shown in Figure 14.1. Notice that the audience had no effect on the high-self-esteem participants. However, the low-self-esteem participants made nearly twice as many errors with an audience as when working alone.

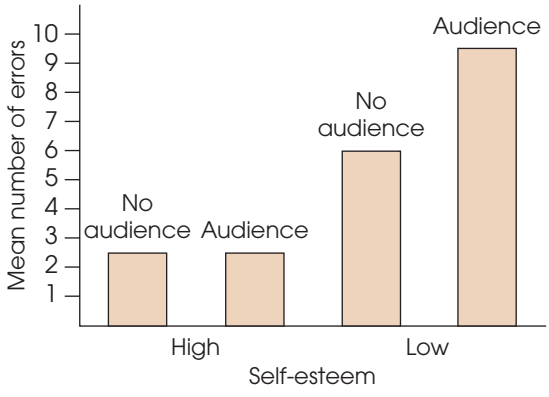


FIGURE 14.1

Results of an experiment examining the effect of an audience on the number of errors made on a concept formation task for participants who are rated either high or low in self-esteem. Notice that the effect of the audience depends on the self-esteem of the participants.

Shrauger, J. S. (1972). Self-esteem and reactions to being observed by others. *Journal of Personality and Social Psychology*, *23*, 192–200. Copyright 1972 by the American Psychological Association. Adapted by permission of the author.

The Problem: Shrauger's study is an example of research that involves two independent variables in the same study. The independent variables are:

1. Audience (present or absent)
2. Self-esteem (high or low)

The results of the study indicate that the effect of one variable (audience) *depends on* another variable (self-esteem).

You should realize that it is quite common to have two variables that interact in this way. For example, a drug may have a profound effect on some patients and have no effect whatsoever on others. Some children survive abusive environments and live normal, productive lives, while others show serious difficulties. To observe how one variable interacts with another, it is necessary to study both variables simultaneously in one study. However, the analysis of variance (ANOVA) procedures introduced in Chapters 12 and 13 are limited to evaluating mean differences produced by one independent variable and are not appropriate for mean differences involving two (or more) independent variables.

The Solution: ANOVA is a very flexible hypothesis testing procedure and can be modified again to evaluate the mean differences produced in a research study with two (or more) independent variables. In this chapter we introduce the *two-factor* ANOVA, which tests the significance of each independent variable acting alone as well as the interaction between variables.

14.1 AN OVERVIEW OF THE TWO-FACTOR, INDEPENDENT-MEASURES ANOVA

In most research situations, the goal is to examine the relationship between two variables. Typically, the research study attempts to isolate the two variables to eliminate or reduce the influence of any outside variables that may distort the relationship being studied. A typical experiment, for example, focuses on one independent variable (which is expected to influence behavior) and one dependent variable (which is a measure of the behavior). In real life, however, variables rarely exist in isolation. That is, behavior usually is influenced by a variety of different variables acting and interacting simultaneously. To examine these more complex, real-life situations, researchers often design research studies that include more than one independent variable. Thus, researchers systematically change two (or more) variables and then observe how the changes influence another (dependent) variable.

An independent variable is a manipulated variable in an experiment. A quasi-independent variable is not manipulated but defines the groups of scores in a nonexperimental study.

In Chapters 12 and 13, we examined ANOVA for *single-factor* research designs—that is, designs that included only one independent variable or only one quasi-independent variable. When a research study involves more than one factor, it is called a *factorial design*. In this chapter, we consider the simplest version of a factorial design. Specifically, we examine ANOVA as it applies to research studies with exactly two factors. In addition, we limit our discussion to studies that use a separate sample for each treatment condition—that is, independent-measures designs. Finally, we consider only research designs for which the sample size (n) is the same for all treatment conditions. In the terminology of ANOVA, this chapter examines *two-factor, independent-measures, equal* n *designs.*

We use Shrauger's audience and self-esteem study described in the Chapter Preview to introduce the two-factor research design. Table 14.1 shows the structure of Shrauger's study. Note that the study involves two separate factors: One factor is manipulated by the researcher, changing from no-audience to audience, and the second factor is self-esteem, which varies from high to low. The two factors are used to create a *matrix* with the different levels of self-esteem defining the rows and the different audience conditions defining the columns. The resulting two-by-two matrix shows four different combinations of the variables, producing four different conditions. Thus, the research study would require four separate samples, one for each *cell*, or box, in the matrix. The dependent variable for the study is the number of errors on the concept-formation task for people observed in each of the four conditions.

TABLE 14.1

The structure of a two-factor experiment presented as a matrix. The two factors are self-esteem and presence/absence of an audience, with two levels for each factor.

		Factor *B*: Audience Condition	
		No Audience	Audience
Factor *A*: Self-Esteem	Low	Scores for a group of participants who are classified as low self-esteem and are tested with no audience.	Scores for a group of participants who are classified as low self-esteem and are tested with an audience.
	High	Scores for a group of participants who are classified as high self-esteem and are tested with no audience.	Scores for a group of participants who are classified as high self-esteem and are tested with an audience.

The two-factor ANOVA tests for mean differences in research studies that are structured like the audience-and-self-esteem example in Table 14.1. For this example, the two-factor ANOVA evaluates three separate sets of mean differences:

1. What happens to the mean number of errors when the audience is added or taken away?
2. Is there a difference in the mean number of errors for participants with high self-esteem compared to those with low self-esteem?
3. Is the mean number of errors affected by specific combinations of self-esteem and audience? (For example, an audience may have a large effect on participants with low self-esteem but only a small effect for those with high self-esteem.)

Thus, the two-factor ANOVA allows us to examine three types of mean differences within one analysis. In particular, we conduct three separate hypotheses tests for the same data, with a separate F-ratio for each test. The three F-ratios have the same basic structure:

$$F = \frac{\text{variance (differences) between treatments}}{\text{variance (differences) expected if there is no treatment effect}}$$

In each case, the numerator of the F-ratio measures the actual mean differences in the data, and the denominator measures the differences that would be expected if there is no treatment effect. As always, a large value for the F-ratio indicates that the sample mean differences are greater than would be expected by chance alone, and, therefore, provides evidence of a treatment effect. To determine whether the obtained F-ratios are *significant*, we need to compare each F-ratio with the critical values found in the F-distribution table in Appendix B.

14.2 MAIN EFFECTS AND INTERACTIONS

As noted in the previous section, a two-factor ANOVA actually involves three distinct hypothesis tests. In this section, we examine these three tests in more detail.

Traditionally, the two independent variables in a two-factor experiment are identified as factor A and factor B. For the study presented in Table 14.1, self-esteem is factor A, and the presence or absence of an audience is factor B. The goal of the study is to evaluate the mean differences that may be produced by either of these factors acting independently or by the two factors acting together.

MAIN EFFECTS

One purpose of the study is to determine whether differences in self-esteem (factor A) result in differences in performance. To answer this question, we compare the mean score for all of the participants with low self-esteem with the mean for those with high self-esteem. Note that this process evaluates the mean difference between the top row and the bottom row in Table 14.1.

To make this process more concrete, we present a set of hypothetical data in Table 14.2. The table shows the mean score for each of the treatment conditions (cells) as well as the overall mean for each column (each audience condition) and the overall mean for each row (each self-esteem group). These data indicate that the low self-esteem participants (the top row) had an overall mean of $M = 8$ errors. This overall mean was obtained by computing the average of the two means in the top row. In

TABLE 14.2

Hypothetical data for an experiment examining the effect of an audience on participants with different levels of self-esteem.

	No Audience	Audience	
Low	$M = 7$	$M = 9$	$M = 8$
High	$M = 3$	$M = 5$	$M = 4$
	$M = 5$	$M = 7$	

contrast, the high self-esteem participants had an overall mean of $M = 4$ errors (the mean for the bottom row). The difference between these means constitutes what is called the *main effect* for self-esteem, or the *main effect for factor* A.

Similarly, the main effect for factor *B* (audience condition) is defined by the mean difference between the columns of the matrix. For the data in Table 14.2, the two groups of participants tested with no audience had an overall mean score of $M = 5$ errors. Participants tested with an audience committed an overall average of $M = 7$ errors. The difference between these means constitutes the *main effect* for the audience conditions, or the *main effect for factor* B.

DEFINITION

The mean differences among the levels of one factor are referred to as the **main effect** of that factor. When the design of the research study is represented as a matrix with one factor determining the rows and the second factor determining the columns, then the mean differences among the rows describe the main effect of one factor, and the mean differences among the columns describe the main effect for the second factor.

The mean differences between columns or rows simply *describe* the main effects for a two-factor study. As we have observed in earlier chapters, the existence of sample mean differences does not necessarily imply that the differences are *statistically significant*. In general, two samples are not expected to have exactly the same means. There are always small differences from one sample to another, and you should not automatically assume that these differences are an indication of a systematic treatment effect. In the case of a two-factor study, any main effects that are observed in the data must be evaluated with a hypothesis test to determine whether they are statistically significant effects. Unless the hypothesis test demonstrates that the main effects are significant, you must conclude that the observed mean differences are simply the result of sampling error.

The evaluation of main effects accounts for two of the three hypothesis tests in a two-factor ANOVA. We state hypotheses concerning the main effect of factor *A* and the main effect of factor *B* and then calculate two separate *F*-ratios to evaluate the hypotheses.

For the example we are considering, factor *A* involves the comparison of two different levels of self-esteem. The null hypothesis would state that there is no difference between the two levels; that is, self-esteem has no effect on performance. In symbols,

$$H_0: \quad \mu_{A_1} = \mu_{A_2}$$

The alternative hypothesis is that the two different levels of self-esteem do produce different scores:

$$H_1: \quad \mu_{A_1} \neq \mu_{A_2}$$

To evaluate these hypotheses, we compute an F-ratio that compares the actual mean differences between the two self-esteem levels versus the amount of difference that would be expected without any systematic treatment effects.

$$F = \frac{\text{variance (differences) between the means for factor } A}{\text{variance (differences) expected if there is no treatment effect}}$$

$$F = \frac{\text{variance (differences) between the row means}}{\text{variance (differences) expected if there is no treatment effect}}$$

Similarly, factor B involves the comparison of the two different audience conditions. The null hypothesis states that there is no difference in the mean number of errors between the two conditions. In symbols,

$$H_0: \quad \mu_{B_1} = \mu_{B_2}$$

As always, the alternative hypothesis states that the means are different:

$$H_1: \quad \mu_{B_1} \neq \mu_{B_2}$$

Again, the F-ratio compares the obtained mean difference between the two audience conditions versus the amount of difference that would be expected if there is no systematic treatment effect.

$$F = \frac{\text{variance (differences) between the means for factor } B}{\text{variance (differences) expected if there is no treatment effect}}$$

$$F = \frac{\text{variance (differences) between the column means}}{\text{variance (differences) expected if there is no treatment effect}}$$

INTERACTIONS

In addition to evaluating the main effect of each factor individually, the two-factor ANOVA allows you to evaluate other mean differences that may result from unique combinations of the two factors. For example, specific combinations of self-esteem and an audience acting together may have effects that are different from the effects of self-esteem or an audience acting alone. Any "extra" mean differences that are not explained by the main effects are called an *interaction,* or an *interaction between factors*. The real advantage of combining two factors within the same study is the ability to examine the unique effects caused by an interaction.

DEFINITION

An **interaction** between two factors occurs whenever the mean differences between individual treatment conditions, or cells, are different from what would be predicted from the overall main effects of the factors.

To make the concept of an interaction more concrete, we reexamine the data shown in Table 14.2. For these data, there is no interaction; that is, there are no extra mean differences that are not explained by the main effects. For example, within each audience condition (each column of the matrix) the average number of errors for the low self-esteem participants is 4 points higher than the average for the high self-esteem participants. This 4-point mean difference is exactly what is predicted by the overall main effect for self-esteem.

Now consider a different set of data shown in Table 14.3. These new data show exactly the same main effects that existed in Table 14.2 (the column means and the row

TABLE 14.3

Hypothetical data for an experiment examining the effect of an audience on participants with different levels of self-esteem. The data show the same main effects as the values in Table 14.5 but the individual treatment means have been modified to create an interaction.

	No Audience	Audience	
Low	$M = 6$	$M = 10$	$M = 8$
High	$M = 4$	$M = 4$	$M = 4$
	$M = 5$	$M = 7$	

means have not been changed). But now there is an interaction between the two factors. For example, for the low self-esteem participants (top row), there is a 4-point difference in the number of errors committed with an audience and without an audience. This 4-point difference cannot be explained by the 2-point main effect for the audience factor. Also, for the high self-esteem participants (bottom row), the data show no difference between the two audience conditions. Again, the zero difference is not what would be expected based on the 2-point main effect for the audience factor. Mean differences that are not explained by the main effects are an indication of an interaction between the two factors.

The data in Table 14.3 show the same pattern of results that was obtained in Shrauger's research study.

To evaluate the interaction, the two-factor ANOVA first identifies mean differences that are not explained by the main effects. The extra mean differences are then evaluated by an F-ratio with the following structure:

$$F = \frac{\text{variance (mean differences) not explained by main effects}}{\text{variance (differences) expected if there is no treatment effects}}$$

The null hypothesis for this F-ratio simply states that there is no interaction:

> H_0: There is no interaction between factors A and B. All of the mean differences between treatment conditions are explained by the main effects of the two factors.

The alternative hypothesis is that there is an interaction between the two factors:

> H_1: There is an interaction between factors. The mean differences between treatment conditions are not what would be predicted from the overall main effects of the two factors.

MORE ABOUT INTERACTIONS

In the previous section, we introduced the concept of an interaction as the unique effect produced by two factors working together. This section presents two alternative definitions of an interaction. These alternatives are intended to help you understand the concept of an interaction and to help you identify an interaction when you encounter one in a set of data. You should realize that the new definitions are equivalent to the original and simply present slightly different perspectives on the same concept.

The first new perspective on the concept of an interaction focuses on the notion of independence for the two factors. More specifically, if the two factors are independent, so that one factor does not influence the effect of the other, then there is no interaction. On the other hand, when the two factors are not independent, so that the effect of one factor *depends on* the other, then there is an interaction. The notion of dependence between factors is consistent with our earlier discussion of interactions. If one factor influences the effect of the other, then unique combinations of the factors produce unique effects.

DEFINITION

When the effect of one factor depends on the different levels of a second factor, then there is an **interaction** between the factors.

This definition of an interaction should be familiar in the context of a "drug interaction." Your doctor and pharmacist are always concerned that the effect of one medication may be altered or distorted by a second medication that is being taken at the same time. Thus, the effect of one drug (factor *A*) depends on a second drug (factor *B*), and you have an interaction between the two drugs.

Returning to Table 14.2, notice that the size of the audience effect (first column versus second column) *does not depend* on the self-esteem of the participants. For these data, adding an audience produces the same 2-point increase in errors for both groups of participants. Thus, the audience effect does not depend on self-esteem, and there is no interaction. Now consider the data in Table 14.3. This time, the effect of adding an audience *depends on* the self-esteem of the participants. For example, there is a 4-point increase in errors for the low-self-esteem participants but adding an audience has no effect on the errors for the high-self-esteem participants. Thus, the audience effect depends on the level of self-esteem, which means that there is an interaction between the two factors.

The second alternative definition of an interaction is obtained when the results of a two-factor study are presented in a line graph. In this case, the concept of an interaction can be defined in terms of the pattern displayed in the graph. Figure 14.2 shows the two sets of data we have been considering. The original data from Table 14.2, where there is no interaction, are presented in Figure 14.2(a). To construct this figure, we selected one of the factors to be displayed on the horizontal axis; in this case, the different levels of the audience factor. The dependent variable, the number of errors, is shown on the vertical axis. Note that the figure actually contains two separate graphs: The top line shows the relationship between the audience factor and errors for the low-self-esteem

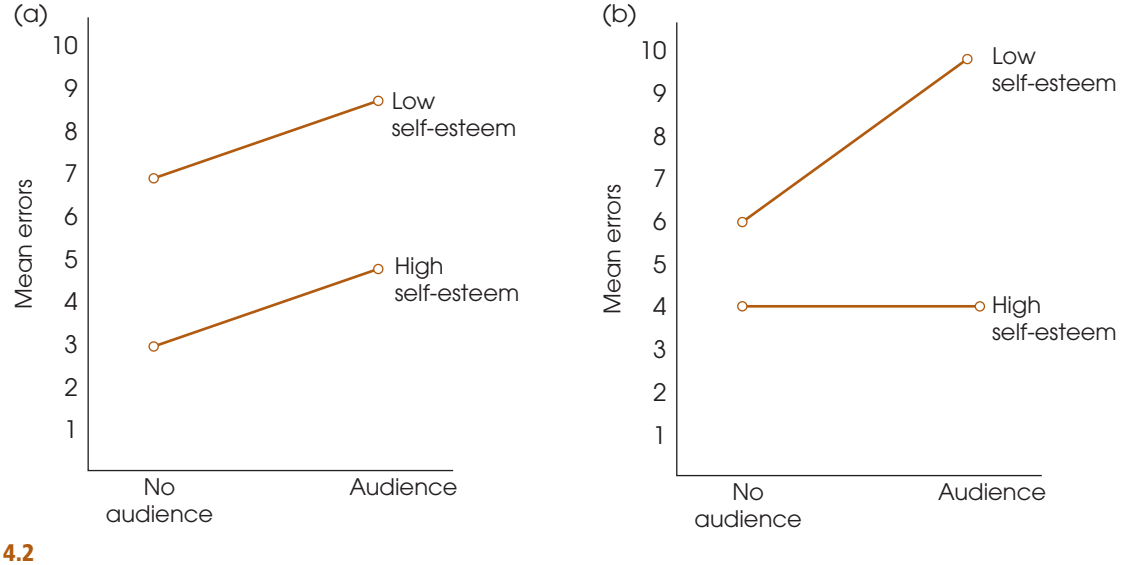


FIGURE 14.2

(a) Graph showing the treatment means from Table 14.2, for which there is no reaction. (b) Graph for Table 14.3, for which there is an interaction.

participants, and the bottom line shows the relationship for the high-self-esteem participants. In general, the picture in the graph matches the structure of the data matrix; the columns of the matrix appear as values along the *X*-axis, and the rows of the matrix appear as separate lines in the graph (Box 14.1).

For the original set of data, Figure 14.2(a), note that the two lines are parallel; that is, the distance between lines is constant. In this case, the distance between lines reflects the 2-point difference in mean errors between low- and high-self-esteem participants, and this 2-point difference is the same for both audience conditions.

Now look at a graph that is obtained when there is an interaction in the data. Figure 14.2(b) shows the data from Table 14.3. This time, note that the lines in the graph are not parallel. The distance between the lines changes as you scan from left to right. For these data, the distance between the lines corresponds to the self-esteem effect—that is, the mean difference in errors for low- versus high-self-esteem participants. The fact that this difference depends on the audience condition is an indication of an interaction between the two factors.

DEFINITION

When the results of a two-factor study are presented in a graph, the existence of nonparallel lines (lines that cross or converge) indicates an **interaction** between the two factors.

The $A \times B$ interaction typically is called the "*A* by *B*" interaction. If there is an interaction between an audience and self-esteem, it may be called the "audience by self-esteem" interaction.

For many students, the concept of an interaction is easiest to understand using the perspective of interdependency; that is, an interaction exists when the effects of one variable *depend* on another factor. However, the easiest way to identify an interaction within a set of data is to draw a graph showing the treatment means. The presence of nonparallel lines is an easy way to spot an interaction.

BOX 14.1 GRAPHING RESULTS FROM A TWO-FACTOR DESIGN

One of the best ways to get a quick overview of the results from a two-factor study is to present the data in a line graph. Because the graph must display the means obtained for *two* independent variables (two factors), constructing the graph can be a bit more complicated than constructing the single-factor graphs we presented in Chapter 3 (pp. 93–95).

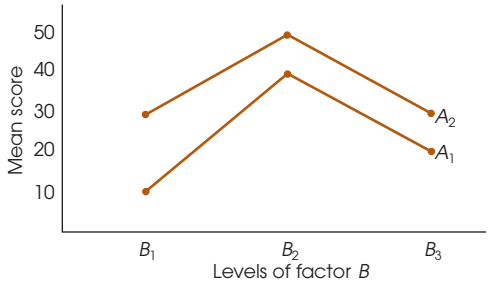
Figure 14.3 shows a line graph presenting the results from a two-factor study with 2 levels of factor *A* and 3 levels of factor *B*. With a 2×3 design, there are a total of 6 different treatment means, which are shown in the following matrix.

		Factor *B*		
		B1	*B2*	*B3*
Factor *A*	A_1	10	40	20
	A_2	30	50	30

In the graph, note that values for the dependent variable (the treatment means) are shown on the vertical axis. Also note that the levels for one factor (we selected factor *B*) are displayed on the horizontal axis. Directly above the B_1 value on the horizontal axis, we have placed two dots corresponding to the two means in the B_1 column of the data matrix. Similarly, we have placed two dots above B_2 and another two dots above B_3. Finally, we have drawn a line connecting the three dots corresponding to level 1 of factor *A* (the three means in the top row of the data matrix). We have also drawn a second line that connects the three dots corresponding to level 2 of factor *A*. These lines are labeled A_1 and A_2 in the figure.

FIGURE 14.3

A line graph showing the results from a two-factor experiment.



INDEPENDENCE OF MAIN EFFECTS AND INTERACTIONS

The two-factor ANOVA consists of three hypothesis tests, each evaluating specific mean differences: the A effect, the B effect, and the $A \times B$ interaction. As we have noted, these are three *separate* tests, but you should also realize that the three tests are *independent*. That is, the outcome for any one of the three tests is totally unrelated to the outcome for either of the other two. Thus, it is possible for data from a two-factor study to display any possible combination of significant and/or nonsignificant main effects and interactions. The data sets in Table 14.4 show several possibilities.

Table 14.4(a) shows data with mean differences between levels of factor A (an A effect) but no mean differences for factor B and no interaction. To identify the A effect, notice that the overall mean for A_1 (the top row) is 10 points higher than the overall mean for A_2 (the bottom row). This 10-point difference is the main effect for factor A. To evaluate the B effect, notice that both columns have exactly the same overall mean, indicating no difference between levels of factor B; hence, there is no B effect. Finally, the absence of an interaction is indicated by the fact that the overall A effect (the 10-point difference) is constant within each column; that is, the A effect *does not depend* on the levels of factor B. (Another indication is that the data indicate that the overall B effect is constant within each row.)

Table 14.4(b) shows data with an A effect and a B effect but no interaction. For these data, the A effect is indicated by the 10-point mean difference between rows, and the B effect is indicated by the 20-point mean difference between columns. The fact that the 10-point A effect is constant within each column indicates no interaction.

Finally, Table 14.4(c) shows data that display an interaction but no main effect for factor A or for factor B. For these data, there is no mean difference between rows (no A effect) and no mean difference between columns (no B effect). However, within each row (or within each column), there are mean differences. The "extra" mean differences within the rows and columns cannot be explained by the overall main effects and, therefore, indicate an interaction.

TABLE 14.4

Three sets of data showing different combinations of main effects and interaction for a two-factor study. (The numerical value in each cell of the matrices represents the mean value obtained for the sample in that treatment condition.)

(a) Data showing a main effect for factor A but no B effect and no interaction

	B_1	B_2	
A_1	20	20	A_1 mean = 20
A_2	10	10	A_2 mean = 10
	B_1 mean = 15	B_2 mean = 15	

10-point difference

No difference

(b) Data showing main effects for both factor *A* and factor *B* but no interaction

	B_1	B_2		
A_1	10	30	A_1 mean = 20	10-point difference
A_2	20	40	A_2 mean = 30	
	B_1 mean = 15	B_2 mean = 35		

20-point difference

(c) Data showing no main effect for either factor but an interaction

	B_1	B_2		
A_1	10	20	A_1 mean = 15	No difference
A_2	20	10	A_2 mean = 15	
	B_1 mean = 15	B_2 mean = 15		

No difference

LEARNING CHECK

1. Each of the following matrices represents a possible outcome of a two-factor experiment. For each experiment:

a. Describe the main effect for factor *A*.

b. Describe the main effect for factor *B*.

c. Does there appear to be an interaction between the two factors?

Experiment I

	B_1	B_2
A_1	$M = 10$	$M = 20$
A_2	$M = 30$	$M = 40$

Experiment II

	B_1	B_2
A_1	$M = 10$	$M = 30$
A_2	$M = 20$	$M = 20$

2. In a graph showing the means from a two-factor experiment, parallel lines indicate that there is no interaction. (True or false?)

3. A two-factor ANOVA consists of three hypothesis tests. What are they?

4. It is impossible to have an interaction unless you also have main effects for at least one of the two factors. (True or false?)

ANSWERS

1. For Experiment I:

a. There is a main effect for factor *A*; the scores in A_2 average 20 points higher than in A_1.

b. There is a main effect for factor *B*; the scores in B_2 average 10 points higher than in B_1.

c. There is no interaction; there is a constant 20-point difference between A_1 and A_2 that does not depend on the levels of factor *B*.

For Experiment II:

a. There is no main effect for factor A; the scores in A_1 and in A_2 both average 20.

b. There is a main effect for factor B; on average, the scores in B_2 are 10 points higher than in B_1.

c. There is an interaction. The difference between A_1 and A_2 depends on the level of factor B. (There is a $+10$ difference in B_1 and a -10 difference in B_2.)

2. True.

3. The two-factor ANOVA evaluates the main effect for factor A, the main effect for factor B, and the interaction between the two factors.

4. False. Main effects and interactions are completely independent.

14.3 NOTATION AND FORMULAS FOR THE TWO-FACTOR ANOVA

The two-factor ANOVA is composed of three distinct hypothesis tests:

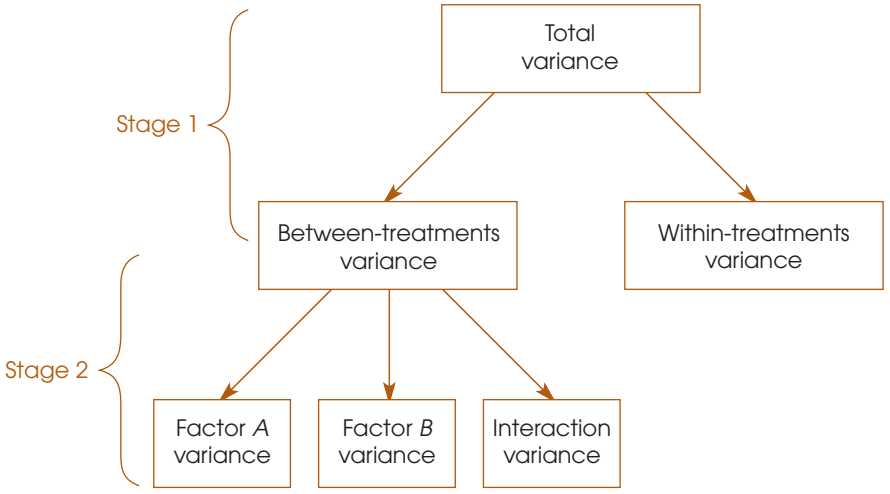
1. The main effect for factor A (often called the A-effect). Assuming that factor A is used to define the rows of the matrix, the main effect for factor A evaluates the mean differences between rows.
2. The main effect for factor B (called the B-effect). Assuming that factor B is used to define the columns of the matrix, the main effect for factor B evaluates the mean differences between columns.
3. The interaction (called the $A \times B$ interaction). The interaction evaluates mean differences between treatment conditions that are not predicted from the overall main effects from factor A and factor B.

For each of these three tests, we are looking for mean differences between treatments that are larger than would be expected if there are no treatment effects. In each case, the significance of the treatment effect is evaluated by an F-ratio. All three F-ratios have the same basic structure:

$$F = \frac{\text{variance (mean differences) between treatments}}{\text{variance (mean differences) expected if there are no treatment effects}} \qquad (14.1)$$

The general structure of the two-factor ANOVA is shown in Figure 14.4. Note that the overall analysis is divided into two stages. In the first stage, the total variability is separated into two components: between-treatments variability and within-treatments variability. This first stage is identical to the single-factor ANOVA introduced in Chapter 12, with each cell in the two-factor matrix viewed as a separate treatment condition. The within-treatments variability that is obtained in stage 1 of the analysis is used to compute the denominator for the F-ratios. As we noted in Chapter 12, within each treatment, all of the participants are treated exactly the same. Thus, any differences that exist within the treatments cannot be caused by treatment effects. As a result,

FIGURE 14.4
Structure of the analysis for a two-factor ANOVA.



the within-treatments variability provides a measure of the differences that exist when there are no systematic treatment effects influencing the scores (see Equation 14.1).

The between-treatments variability obtained in stage 1 of the analysis combines all of the mean differences produced by factor *A*, factor *B*, and the interaction. The purpose of the second stage is to partition the differences into three separate components: differences attributed to factor *A*, differences attributed to factor *B*, and any remaining mean differences that define the interaction. These three components form the numerators for the three *F*-ratios in the analysis.

The goal of this analysis is to compute the variance values needed for the three *F*-ratios. We need three between-treatments variances (one for factor *A*, one for factor *B*, and one for the interaction), and we need a within-treatments variance. Each of these variances (or mean squares) is determined by a sum of squares value (*SS*) and a degrees of freedom value (*df*):

Remember that in ANOVA a variance is called a mean square, or *MS*.

$$\text{mean square} = MS = \frac{SS}{df}$$

EXAMPLE 14.1

We use the data shown in Table 14.5 to demonstrate the two-factor ANOVA. The data are representative of many studies examining the relationship between arousal and performance. The general result of these studies is that increasing the level of arousal (or motivation) tends to improve the level of performance. (You probably have tried to "psych yourself up" to do well on a task.) For very difficult tasks, however, increasing arousal beyond a certain point tends to lower the level of performance. (Your friends have probably advised you to "calm down and stay focused" when you get overanxious about doing well.) This relationship between arousal and performance is known as the Yerkes-Dodson law.

The data are displayed in a matrix with the two levels of task difficulty (factor *A*) making up the rows and the three levels of arousal (factor *B*) making up

TABLE 14.5

Data for a two-factor research study comparing two levels of task difficulty (easy and hard) and three levels of arousal (low, medium, and high). The study involves a total of six different treatment conditions with $n = 5$ participants in each condition.

Factor A: Task Difficulty	Factor B: Arousal Level — Low	Medium	High	
Easy	3	1	10	$T_{ROW1} = 90$
	1	4	10	
	1	8	14	
	6	6	7	
	4	6	9	
	$M = 3$	$M = 5$	$M = 10$	
	$T = 15$	$T = 25$	$T = 50$	
	$SS = 18$	$SS = 28$	$SS = 26$	$N = 30$
				$G = 120$
				$\Sigma X^2 = 860$
Difficult	0	2	1	
	2		7q	
	0	2	1	
	0	2	6	$T_{ROW2} = 30$
	3	2	1	
	$M = 1$	$M = 3$	$M = 2$	
	$T = 5$	$T = 15$	$T = 10$	
	$SS = 8$	$SS = 20$	$SS = 20$	
	$T_{COL1} = 20$	$T_{COL2} = 40$	$T_{COL3} = 60$	

the columns. For the easy task, note that performance scores increase consistently as arousal increases. For the difficult task, on the other hand, performance peaks at a medium level of arousal and drops when arousal is increased to a high level. Note that the data matrix has a total of six *cells,* or treatment conditions, with a separate sample of $n = 5$ subjects in each condition. Most of the notation should be familiar from the single-factor ANOVA presented in Chapter 12. Specifically, the treatment totals are identified by T values, the total number of scores in the entire study is $N = 30$, and the grand total (sum) of all 30 scores is $G = 120$. In addition to these familiar values, we have included the totals for each row and for each column in the matrix. The goal of the ANOVA is to determine whether the mean differences observed in the data are significantly greater than would be expected if there are no treatment effects.

STAGE 1 OF THE TWO-FACTOR ANOVA

The first stage of the two-factor ANOVA separates the total variability into two components: between-treatments and within-treatments. The formulas for this stage are identical to the formulas used in the single-factor ANOVA in Chapter 12 with the provision that each cell in the two-factor matrix is treated as a separate treatment condition. The formulas and the calculations for the data in Table 14.5 are as follows:

Total variability

$$SS_{total} = \Sigma X^2 - \frac{G^2}{N} \tag{14.2}$$

For these data,

$$SS_{\text{total}} = 860 - \frac{120^2}{30}$$

$$= 860 - 480$$

$$= 380$$

This SS value measures the variability for all $N = 30$ scores and has degrees of freedom given by

$$df_{\text{total}} = N - 1 \tag{14.3}$$

For the data in Table 14.5, $df_{\text{total}} = 29$.

Within-treatments variability To compute the variance within treatments, we first compute SS and $df = n - 1$ for each of the individual treatment conditions. Then the within-treatments SS is defined as

$$SS_{\text{within treatments}} = \Sigma SS_{\text{each treatment}} \tag{14.4}$$

And the within-treatments df is defined as

$$df_{\text{within treatments}} = \Sigma df_{\text{each treatment}} \tag{14.5}$$

For the six treatment conditions in Table 14.4,

$$SS_{\text{within treatments}} = 18 + 28 + 26 + 8 + 20 + 20$$

$$= 120$$

$$df_{\text{within treatments}} = 4 + 4 + 4 + 4 + 4 + 4$$

$$= 24$$

Between-treatments variability Because the two components in stage 1 must add up to the total, the easiest way to find $SS_{\text{between treatments}}$ is by subtraction.

$$SS_{\text{between treatments}} = SS_{\text{total}} - SS_{\text{within}} \tag{14.6}$$

For the data in Table 14.4, we obtain

$$SS_{\text{between treatments}} = 380 - 120 = 260$$

However, you can also use the computational formula to calculate $SS_{\text{between treatments}}$ directly.

$$SS_{\text{between treatments}} = \Sigma \frac{T^2}{n} - \frac{G^2}{N} \tag{14.7}$$

For the data in Table 14.4, there are six treatments (six T values), each with $n = 5$ scores, and the between-treatments SS is

$$SS_{\text{between treatments}} = \frac{15^2}{5} + \frac{25^2}{5} + \frac{50^2}{5} + \frac{5^2}{5} + \frac{15^2}{5} + \frac{10^2}{5} - \frac{120^2}{30}$$

$$= 45 + 125 + 500 + 5 + 45 + 20 - 480$$

$$= 260$$

The between-treatments *df* value is determined by the number of treatments (or the number of *T* values) minus one. For a two-factor study, the number of treatments is equal to the number of cells in the matrix. Thus,

$$df_{\text{between treatments}} = \text{number of cells} - 1 \tag{14.8}$$

For these data, $df_{\text{between treatments}} = 5$.

This completes the first stage of the analysis. Note that the two components, when added, equal the total for both *SS* values and *df* values.

$$SS_{\text{between treatments}} + SS_{\text{within treatments}} = SS_{\text{total}}$$

$$240 + 120 = 360$$

$$df_{\text{between treatments}} + df_{\text{within treatments}} = df_{\text{total}}$$

$$5 + 24 = 29$$

STAGE 2 OF THE TWO-FACTOR ANOVA

The second stage of the analysis determines the numerators for the three *F*-ratios. Specifically, this stage determines the between-treatments variance for factor *A*, factor *B*, and the interaction.

1. Factor *A*. The main effect for factor *A* evaluates the mean differences between the levels of factor *A*. For this example, factor *A* defines the rows of the matrix, so we are evaluating the mean differences between rows. To compute the *SS* for factor *A*, we calculate a between-treatment *SS* using the row totals in exactly the same way that we computed $SS_{\text{between treatments}}$ using the treatment totals (*T* values) earlier. For factor *A*, the row totals are 90 and 30, and each total was obtained by adding 15 scores.

   Therefore,

$$SS_A = \Sigma \frac{T_{ROW}^2}{n_{ROW}} - \frac{G^2}{N} \tag{14.9}$$

   For our data,

$$SS_A = \frac{90^2}{15} + \frac{30^2}{15} - \frac{120^2}{30}$$

$$= 540 + 60 - 480$$

$$= 120$$

   Factor *A* involves two treatments (or two rows), easy and difficult, so the *df* value is

$$df_A = \text{number of rows} - 1 \tag{14.10}$$

$$= 2 - 1$$

$$= 1$$

2. Factor *B*. The calculations for factor *B* follow exactly the same pattern that was used for factor *A*, except for substituting columns in place of rows. The main

effect for factor B evaluates the mean differences between the levels of factor B, which define the columns of the matrix.

$$SS_B = \Sigma \frac{T_{COL}^2}{n_{COL}} - \frac{G^2}{N} \tag{14.11}$$

For our data, the column totals are 20, 40, and 60, and each total was obtained by adding 10 scores. Thus,

$$\begin{aligned} SS_B &= \frac{20^2}{10} + \frac{40^2}{10} + \frac{60^2}{10} - \frac{120^2}{30} \\ &= 40 + 160 + 360 - 480 \\ &= 80 \end{aligned}$$

$$\begin{aligned} df_B &= \text{number of columns} - 1 \\ &= 3 - 1 \\ &= 2 \end{aligned} \tag{14.12}$$

3. The $A \times B$ Interaction. The $A \times B$ interaction is defined as the "extra" mean differences not accounted for by the main effects of the two factors. We use this definition to find the SS and df values for the interaction by simple subtraction. Specifically, the between-treatments variability is partitioned into three parts: the A effect, the B effect, and the interaction (see Figure 14.4). We have already computed the SS and df values for A and B, so we can find the interaction values by subtracting to find out how much is left. Thus,

$$SS_{A\times B} = SS_{\text{between treatments}} - SS_A - SS_B \tag{14.13}$$

For our data,

$$\begin{aligned} SS_{A\times B} &= 260 - 120 - 80 \\ &= 60 \end{aligned}$$

Similarly,

$$\begin{aligned} df_{A\times B} &= df_{\text{between treatments}} - df_A - df_B \\ &= 5 - 1 - 2 \\ &= 2 \end{aligned} \tag{14.14}$$

The two-factor ANOVA consists of three separate hypothesis tests with three separate F-ratios. The denominator for each F-ratio is intended to measure the variance (differences) that would be expected if there are no treatment effects. As we saw in Chapter 12, the within-treatments variance is the appropriate denominator for an independent-measures design. Remember that inside each treatment all of the individuals are treated exactly the same, which means that the differences that exist were not caused by any systematic treatment effects (see Chapter 12, p. 393). The within-treatments variance is called a *mean square,* or MS, and is computed as follows:

$$MS_{\text{within treatments}} = \frac{SS_{\text{within treatments}}}{df_{\text{within treatments}}}$$

For the data in Table 14.4,

$$MS_{\text{within treatments}} = \frac{120}{24} = 5.00$$

This value forms the denominator for all three *F*-ratios.

The numerators of the three *F*-ratios all measured variance or differences between treatments: differences between levels of factor *A*, differences between levels of factor *B*, and extra differences that are attributed to the $A \times B$ interaction. These three variances are computed as follows:

$$MS_A = \frac{SS_A}{df_A} \quad MS_B = \frac{SS_B}{df_B} \quad MS_{A\times B} = \frac{SS_{A\times B}}{df_{A\times B}}$$

For the data in Table 14.5, the three *MS* values are

$$MS_A = \frac{SS_A}{df_A} = \frac{120}{1} = 120 \quad MS_B = \frac{SS_B}{df_B} = \frac{80}{2} = 40$$

$$MS_{A\times B} = \frac{SS_{A\times B}}{df_{A\times B}} = \frac{60}{2} = 30$$

Finally, the three *F*-ratios are

$$F_A = \frac{MS_A}{MS_{\text{within treatments}}} = \frac{120}{5} = 24.00$$

$$F_B = \frac{MS_B}{MS_{\text{within treatments}}} = \frac{40}{5} = 8.00$$

$$F_{A\times B} = \frac{MS_{A\times B}}{MS_{\text{within treatments}}} = \frac{30}{5} = 6.00$$

To determine the significance of each *F*-ratio, we must consult the *F* distribution table using the *df* values for each of the individual *F*-ratios. For this example, the *F*-ratio for factor *A* has $df = 1$ for the numerator and $df = 24$ for the denominator. Checking the table with $df = 1, 24$, we find a critical value of 4.26 for $\alpha = .05$ and a critical value of 7.82 for $\alpha = .01$. Our obtained *F*-ratio, $F = 24.00$ exceeds both of these values, so we conclude that there is a significant difference between the levels of factor *A*. That is, performance on the easy task (top row) is significantly different from performance on the difficult task (bottom row).

The *F*-ratio for factor *B* has $df = 2, 24$. The critical values obtained from the table are 3.40 for $\alpha = .05$ and 5.61 for $\alpha = .01$. Again, our obtained *F*-ratio, $F = 8.00$, exceeds both values, so we can conclude that there are significant differences among the levels of factor *B*. For this study, the three levels of arousal result in significantly different levels of performance.

Finally, the *F*-ratio for the $A \times B$ interaction has $df = 2, 24$ (the same as factor *B*). With critical values of 3.40 for $\alpha = .05$ and 5.61 for $\alpha = .01$, our obtained *F*-ratio of $F = 6.00$ is sufficient to conclude that there is a significant interaction between task difficulty and level of arousal.

Table 14.6 is a summary table for the complete two-factor ANOVA from Example 14.1. Although these tables are no longer commonly used in research reports, they provide a concise format for displaying all of the elements of the analysis.

LEARNING CHECK

1. Explain why the within-treatment variability is the appropriate denominator for the two-factor independent-measures F-ratios.
2. The following data summarize the results from a two-factor independent-measures experiment:

		Factor B		
		B_1	B_2	B_3
Factor A	A_1	$n = 10$ $T = 0$ $SS = 30$	$n = 10$ $T = 10$ $SS = 40$	$n = 10$ $T = 20$ $SS = 50$
	A_2	$n = 10$ $T = 40$ $SS = 60$	$n = 10$ $T = 30$ $SS = 50$	$n = 10$ $T = 20$ $SS = 40$

   a. Calculate the totals for each level of factor A, and compute SS for factor A.

   b. Calculate the totals for factor B, and compute SS for this factor. (*Note:* You should find that the totals for B are all the same, so there is no variability for this factor.)

   c. Given that the between-treatments (or between-cells) SS is equal to 100, what is the SS for the interaction?

ANSWERS

1. Within each treatment condition, all individuals are treated exactly the same. Therefore, the within-treatment variability measures the differences that exist between one score and another when there is no treatment effect causing the scores to be different. This is exactly the variance that is needed for the denominator of the F-ratios.
2. **a.** The totals for factor A are 30 and 90, and each total is obtained by adding 30 scores. $SS_A = 60$.

   b. All three totals for factor B are equal to 40. Because they are all the same, there is no variability, and $SS_B = 0$.

   c. The interaction is determined by differences that remain after the main effects have been accounted for. For these data,

$$\begin{aligned} SS_{A\times B} &= SS_{\text{between treatments}} - SS_A - SS_B \\ &= 100 - 60 - 0 \\ &= 40 \end{aligned}$$

TABLE 14.6

A summary table for the two-factor ANOVA for the data from Example 14.1.

Source	SS	df	MS	F
Between treatments	260	5		
Factor A (difficulty)	120	1	120	$F(1, 24) = 24.00$
Factor B (arousal)	80	2	40	$F(2, 24) = 8.00$
$A \times B$	60	2	30	$F(2, 24) = 6.00$
Within treatments	120	24	5	
Total	380	29		

MEASURING EFFECT SIZE FOR THE TWO-FACTOR ANOVA

The general technique for measuring effect size with an ANOVA is to compute a value for η^2, the percentage of variance that is explained by the treatment effects. For a two-factor ANOVA, we compute three separate values for eta squared: one measuring how much of the variance is explained by the main effect for factor *A*, one for factor *B*, and a third for the interaction. As we did with the repeated-measures ANOVA (p. 446), we remove any variability that can be explained by other sources before we calculate the percentage for each of the three specific effects. Thus, for example, before we compute the η^2 for factor *A*, we remove the variability that is explained by factor *B* and the variability explained by the interaction. The resulting equation is,

$$\text{for factor } A,\ \eta^2 = \frac{SS_A}{SS_{\text{total}} - SS_B - SS_{A\times B}} \tag{14.15}$$

Note that the denominator of Equation 14.15 consists of the variability that is explained by factor *A* and the other *unexplained* variability. Thus, an equivalent version of the equation is,

$$\text{for factor } A,\ \eta^2 = \frac{SS_A}{SS_A + SS_{\text{within treatments}}} \tag{14.16}$$

Similarly, the η^2 formulas for factor *B* and for the interaction are as follows:

$$\text{for factor } B,\ \eta^2 = \frac{SS_B}{SS_{\text{total}} - SS_A - SS_{A\times B}} = \frac{SS_B}{SS_B + SS_{\text{within treatments}}} \tag{14.17}$$

$$\text{for } A \times B,\ \eta^2 = \frac{SS_{A\times B}}{SS_{\text{total}} - SS_A - SS_B} = \frac{SS_{A\times B}}{SS_{A\times B} + SS_{\text{within treatments}}} \tag{14.18}$$

Because each of the η^2 equations computes a percentage that is not based on the total variability of the scores, the results are often called *partial* eta squares. For the data in Example 14.1, the equations produce the following values:

$$\eta^2 \text{ for factor } A \text{ (difficulty)} = \frac{120}{380-80-60} = \frac{120}{240} = 0.50 \ (50\%)$$

$$\eta^2 \text{ for factor } B \text{ (arousal)} = \frac{80}{380-120-60} = \frac{80}{200} = 0.40 \ (40\%)$$

$$\eta^2 \text{ for the interaction} = \frac{60}{380-120-80} = \frac{60}{180} = 0.33 \ (33\%)$$

IN THE LITERATURE

REPORTING THE RESULTS OF A TWO-FACTOR ANOVA

The APA format for reporting the results of a two-factor ANOVA follows the same basic guidelines as the single-factor report. First, the means and standard deviations are reported. Because a two-factor design typically involves several treatment conditions, these descriptive statistics usually are presented in a table or a graph.

Next, the results of all three hypothesis tests (F-ratios) are reported. The results for the study in Example 14.1 could be reported as follows:

The means and standard deviations for all treatment conditions are shown in Table 1. The two-factor analysis of variance showed a significant main effect for task difficulty, $F(1, 24) = 24.00$, $p < .01$, $\eta^2 = 0.50$; a significant main effect for arousal, $F(2, 24) = 8.00$, $p < .01$, $\eta^2 = 0.40$; and a significant interaction between difficulty and arousal, $F(2, 24) = 6.00$, $p < .01$, $\eta^2 = 0.33$.

TABLE 1

Mean performance score for each treatment condition.

		Level of Arousal		
		Low	Medium	High
Difficulty	Easy	$M = 3$ $SD = 2.12$	$M = 5$ $SD = 2.65$	$M = 10$ $SD = 2.55$
	Hard	$M = 1$ $SD = 1.41$	$M = 3$ $SD = 2.24$	$M = 2$ $SD = 2.24$

INTERPRETING THE RESULT FROM A TWO-FACTOR ANOVA

Because the two-factor ANOVA involves three separate tests, you must consider the overall pattern of results rather than focusing on the individual main effects or the interaction. In particular, whenever there is a significant interaction, you should be cautious about accepting the main effects at face value (whether they are significant or not). Remember, an interaction means that the effect of one factor *depends on* the level of the second factor. Because the effect changes from one level to the next, there is no consistent "main effect."

Figure 14.5 shows the sample means obtained from the task difficulty and arousal study. Recall that the analysis showed that both main effects and the interaction were

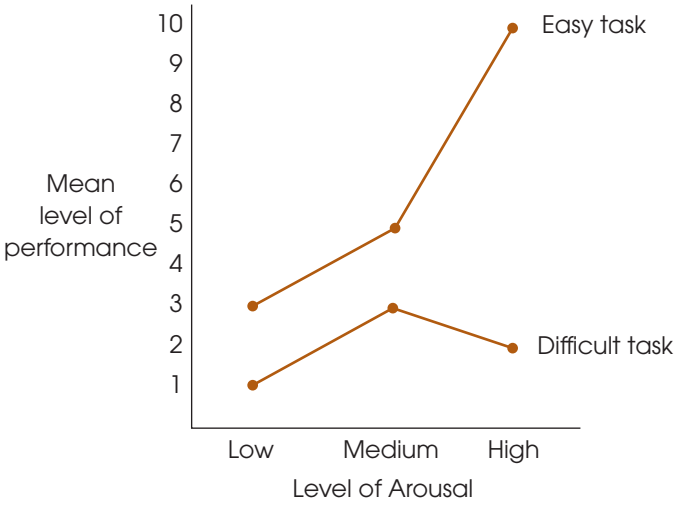


FIGURE 14.5

Sample means for the data in Example 14.1. The data are hypothetical results for a two-factor study examining how performance is related to task difficulty and level of arousal.

significant. The main effect for factor *A* (task difficulty) can be seen by the fact that the scores on the easy task are generally higher than scores on the difficult task.

The main effect for factor *B* (arousal) is based on the general tendency for the scores to increase as the level of arousal increases. However, this is not a completely consistent trend. In fact, the scores on the difficult task show a sharp *decrease* when arousal is increased from moderate to high. This is an example of the complications that can occur when you have a significant interaction. Remember that an interaction means that a factor does not have a consistent effect. Instead, the effect of one factor *depends on* the other factor. For the data in Figure 14.5, the effect of increasing arousal depends on the task difficulty. For the easy task, increasing arousal produces increased performance. For the difficult task, however, increasing arousal beyond a moderate level produces decreased performance. Thus, the consequences of increasing arousal *depend on* the difficulty of the task. This interdependence between factors is the source of the significant interaction.

TESTING SIMPLE MAIN EFFECTS

The existence of a significant interaction indicates that the effect (mean differences) for one factor depends on the levels of the second factor. When the data are presented in a matrix showing treatment means, a significant interaction indicates that the mean differences within one column (or row) show a different pattern than the mean differences within another column (or row). In this case, a researcher may want to perform a separate analysis for each of the individual columns (or rows). In effect, the researcher is separating the two-factor experiment into a series of separate single-factor experiments. The process of testing the significance of mean differences within one column (or one row) of a two-factor design is called testing *simple main effects.* To demonstrate this process, we once again use the data from the task-difficulty and arousal study (Example 14.1), which are summarized in Figure 14.5.

EXAMPLE 14.2

For this demonstration, we test for significant mean differences within each column of the two-factor data matrix. That is, we test for significant mean differences between the two levels of task difficulty for the low level of arousal, then repeat the test for the medium level of arousal, and once more for the high level. In terms of the two-factor notation system, we test the simple main effect of factor *A* for each level of factor *B*.

For the low level of arousal We begin by considering only the low level of arousal. Because we are restricting the data to the first column of the data matrix, the data effectively have been reduced to a single-factor study comparing only two treatment conditions. Therefore, the analysis is essentially a single-factor ANOVA duplicating the procedure presented in Chapter 12. To facilitate the change from a two-factor to a

Low Level of Arousal		
Easy Task	Difficult Task	
$n = 5$	$n = 5$	$N = 10$
$M = 3$	$M = 1$	$G = 20$
$T = 15$	$T = 5$	

single-factor ANOVA, the data for the low level of arousal (first column of the matrix) are reproduced using the notation for a single-factor study.

STEP 1 State the hypothesis. For this restricted set of the data, the null hypothesis would state that there is no difference between the mean for the easy task condition and the mean for the difficult task condition. In symbols,

$$H_0: \mu_{\text{easy}} = \mu_{\text{difficult}} \qquad \text{for the low level of arousal}$$

STEP 2 To evaluate this hypothesis, we use an F-ratio for which the numerator, $MS_{\text{between treatments}}$, is determined by the mean differences between these two groups and the denominator consists of $MS_{\text{within treatments}}$ from the original ANOVA. Thus, the F-ratio has the structure

$$F = \frac{\text{variance (differences) for the means in column 1}}{\text{variance (differences) expected if there are no treatment effects}}$$

$$= \frac{MS_{\text{between treatments}} \text{ for the two treatments in column 1}}{MS_{\text{within treatments}} \text{ from the original ANOVA}}$$

To compute the $MS_{\text{between treatments}}$, we begin with the two treatment totals $T = 15$ and $T = 5$. Each of these totals is based on $n = 5$ scores, and the two totals add up to a grand total of $G = 20$. The $SS_{\text{between treatments}}$ for the two treatments is

$$SS_{\text{between treatments}} = \Sigma\frac{T^2}{n} - \frac{G^2}{N}$$

$$= \frac{15^2}{5} + \frac{5^2}{5} - \frac{20^2}{10}$$

$$= 45 + 5 - 40$$

$$= 10$$

$$MS_{\text{between treatments}} = \frac{SS}{df} = \frac{10}{1} = 10$$

Remember that the F-ratio uses $MS_{\text{within treatments}}$ from the original ANOVA. This $MS = 5$ with $df = 24$. Because this SS value is based on only two treatments, it has $df = 1$. Therefore,

Using $MS_{\text{within treatments}} = 5$ from the original two-factor analysis, the final F-ratio is

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{within treatments}}} = \frac{10}{5} = 2.00$$

Note that this F-ratio has the same df values (1, 24) as the test for factor A main effects (easy versus difficult) in the original ANOVA. Therefore, the critical value for the F-ratio is the same as that in the original ANOVA. With $df = 1, 24$ the critical value is 4.26. In this case, our F-ratio fails to reach the critical value, so we conclude that there is no significant difference between the two tasks, easy and difficult, at a low level of arousal.

For the medium level of arousal The test for the medium level of arousal follows the same process. The data for the medium level are as follows:

Medium Level of Arousal		
Easy Task	Difficult Task	
$n = 5$	$n = 5$	$N = 10$
$M = 5$	$M = 3$	$G = 40$
$T = 25$	$T = 15$	

Note that these data show a 2-point mean difference between the two conditions ($M = 5$ and $M = 3$), which is exactly the same as the 2-point difference that we evaluated for the low level of arousal ($M = 3$ and $M = 1$). Because the mean difference is the same for these two levels of arousal, the F-ratios are also identical. For the low level of arousal, we obtained $F(1, 24) = 2.00$, which was not significant. This test also produces $F(1, 24) = 2.00$ and again we conclude that there is no significant difference. (*Note:* You should be able to complete the test to verify this decision.)

For the high level of arousal The data for the high level are as follows:

High Level of Arousal		
Easy Task	Difficult Task	
$n = 5$	$n = 5$	$N = 10$
$M = 10$	$M = 2$	$G = 60$
$T = 50$	$Y = 10$	

For these data,

$$SS_{\text{between treatments}} = \Sigma\frac{T^2}{n} - \frac{G^2}{N}$$
$$= \frac{50^2}{5} + \frac{10^2}{5} + \frac{60^2}{10}$$
$$= 500 + 20 - 360$$
$$= 160$$

Again, we are comparing only two treatment conditions, so $df = 1$ and

$$MS_{\text{between treatments}} = \frac{SS}{df} = \frac{160}{1} = 160$$

Thus, for the high level of arousal, the final F-ratio is

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{within treatments}}} = \frac{160}{5} = 32.00$$

As before, this F-ratio has $df = 1, 24$ and is compared with the critical value $F = 4.26$. This time the F-ratio is far into the critical region and we conclude that

there is a significant difference between the easy task and the difficult task for the high level of arousal.

As a final note, we should point out that the evaluation of simple main effects accounts for the interaction as well as the overall main effect for one factor. In Example 14.1, the significant interaction indicates that the effect of task difficulty (factor *A*) depends on the level of arousal (factor *B*). The evaluation of the simple main effects demonstrates this dependency. Specifically, task difficulty has no significant effect on performance when arousal level is low or medium, but does have a significant effect when arousal level is high. Thus, the analysis of simple main effects provides a detailed evaluation of the effects of one factor *including its interaction with a second factor.*

The fact that the simple main effects for one factor encompass both the interaction and the overall main effect of the factor can be seen if you consider the *SS* values. For this demonstration,

Simple Main Effects for Arousal	Interaction and Main Effect for Arousal
$SS_{\text{low arousal}} = 10$ $SS_{\text{medium arousal}} = 10$ $SS_{\text{high arousal}} = 160$	$SS_{A \times B} = 60$ $SS_A = 120$
Total $SS = 180$	Total $SS = 180$

Notice that the total variability from the simple main effects of difficulty (factor *A*) completely accounts for the total variability of factor *A* and the $A \times B$ interaction.

14.4 USING A SECOND FACTOR TO REDUCE VARIANCE CAUSED BY INDIVIDUAL DIFFERENCES

As we noted in Chapters 10 and 12, a concern for independent-measures designs is the variance that exists within each treatment condition. Specifically, large variance tends to reduce the size of the *t* statistic or *F*-ratio and, therefore, reduces the likelihood of finding significant mean differences. Much of the variance in an independent-measures study comes from individual differences. Recall that individual differences are the characteristics, such as age or gender, that differ from one participant to the next and can influence the scores obtained in the study.

Occasionally, there are consistent individual differences for one specific participant characteristic. For example, the males in a study may consistently have lower scores than the females. Or, the older participants may have consistently higher scores than the younger participants. For example, suppose that a researcher compares two treatment conditions using a separate group of children for each condition. Each group of participants contains a mix of boys and girls. Hypothetical data for this study are shown in Table 14.7(a), with each child's gender noted with an M or an F. While examining the results, the researcher notices that the girls tend to have higher scores than the boys, which produces big individual differences and high variance within each group. Fortunately, there is a relatively simple solution to the problem of high variance. The solution involves using the specific variable, in this case gender,

TABLE 14.7

A single-factor study comparing two treatments (a) can be transformed into a two-factor study (b) by using a participant characteristic (gender) as a second factor. This process creates smaller, more homogeneous groups, which reduces the variance within groups.

(a)

Treatment I	Treatment II
3 (M)	8 (F)
4 (F)	4 (F)
4 (F)	1 (M)
0 (M)	10 (F)
6 (F)	5 (M)
1 (M)	5 (M)
2 (F)	10 (F)
4 (M)	5 (M)
$M = 3$	$M = 6$
$SS = 50$	$SS = 68$

(b)

	Treatment I	Treatment II
Males	3	1
	0	5
	1	5
	4	5
	$M = 2$	$M = 4$
	$SS = 10$	$SS = 12$
Females	4	8
	4	4
	6	10
	2	10
	$M = 4$	$M = 8$
	$SS = 8$	$SS = 24$

as a second factor. Instead of one group in each treatment, the researcher divides the participants into two separate groups within each treatment: a group of boys and a group of girls. This process creates the two-factor study shown in Table 14.7(b), with one factor consisting of the two treatments (I and II) and the second factor consisting of the gender (male and female).

By adding a second factor and creating four groups of participants instead of only two, the researcher has greatly reduced the individual differences (gender differences) within each group. This should produce a smaller variance within each group and, therefore, increase the likelihood of obtaining a significant mean difference. This process is demonstrated in the following example.

EXAMPLE 14.3

We use the data in Table 14.7 to demonstrate how the variance caused by individual differences can be reduced by adding a participant characteristic, such as age or gender, as a second factor. For the single-factor study in Table 14.7(a), the two treatments produce $SS_{\text{within treatments}} = 50 + 68 = 118$. With $n = 8$ in each treatment, we obtain $df_{\text{within treatments}} = 7 + 7 = 14$. These values produce $MS_{\text{within treatments}} = \frac{118}{14} = 8.43$, which is the denominator of the F-ratio evaluating the mean difference between treatments. For the two-factor study in Table 14.7(b), the four treatments produce $SS_{\text{within treatments}} = 10 + 12 + 8 + 24 = 54$. With $n = 4$ in each treatment, we obtain $df_{\text{within treatments}} = 3 + 3 + 3 + 3 = 12$. These value produce $MS_{\text{within treatments}} = \frac{54}{12} = 4.50$, which is the denominator of the F-ratio evaluating the main effect for the treatments. Notice that the error term for the single-factor F is nearly twice as big as the error term for the two-factor F. Reducing the individual differences within each group has greatly reduced the within-treatment variance that forms the denominator of the F-ratio.

Both designs, single-factor and two-factor, evaluate the difference between the two treatment means, $M = 3$ and $M = 6$, with $n = 8$ in each treatment. These values produce $SS_{\text{between treatments}} = 36$ and, with $k = 2$ treatments, we obtain $df_{\text{between treatments}} = 1$. Thus, $MS_{\text{between treatments}} = \frac{36}{1} = 36$. (For the two-factor design, this is the MS for the main effect of the treatment factor.) With different denominators, however, the two designs produce very different F-ratios. For the single-factor design, we obtain

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{within treatments}}} = \frac{36}{8.43} = 4.27$$

With $df = 1, 14$, the critical value for $\alpha = .05$ is $F = 4.60$. Our F-ratio is not in the critical region, so we fail to reject the null hypothesis and must conclude that there is no significant difference between the two treatments.

For the two-factor design, however, we obtain

$$F = \frac{MS_{\text{between treatments}}}{MS_{\text{within treatments}}} = \frac{36}{4.50} = 8.88$$

With $df = 1, 14$, the critical value for $\alpha = .05$ is $F = 4.75$. Our F-ratio is well beyond this value, so we reject the null hypothesis and conclude that there is a significant difference between the two treatments.

For the single-factor study in Example 14.3, the individual differences caused by gender are part of the variance within each treatment condition. This increased variance reduces the F-ratio and results in a conclusion of no significant difference between treatments. In the two-factor analysis, the individual differences caused by gender are measured by the main effect for gender, which is a between-groups factor. Because the gender differences are now between-groups rather than within-groups, they no longer contribute to the variance.

The two-factor ANOVA has other advantages beyond reducing the variance. Specifically, it allows you to evaluate mean differences between genders as well as differences between treatments, and it reveals any interaction between treatment and gender.

14.5 ASSUMPTIONS FOR THE TWO-FACTOR ANOVA

The validity of the ANOVA presented in this chapter depends on the same three assumptions we have encountered with other hypothesis tests for independent-measures designs (the t test in Chapter 10 and the single-factor ANOVA in Chapter 12):

1. The observations within each sample must be independent (see p. 254).
2. The populations from which the samples are selected must be normal.
3. The populations from which the samples are selected must have equal variances (homogeneity of variance).

As before, the assumption of normality generally is not a cause for concern, especially when the sample size is relatively large. The homogeneity of variance assumption is more important, and if it appears that your data fail to satisfy this requirement, you should conduct a test for homogeneity before you attempt the ANOVA. Hartley's F-max test (see p. 338) allows you to use the sample variances from your data to determine whether there is evidence for any differences among the population variances. Remember, for the two-factor ANOVA, there is a separate sample for each cell in the data matrix. The test for homogeneity applies to all of these samples and the populations that they represent.

SUMMARY

1. A research study with two independent variables is called a two-factor design. Such a design can be diagramed as a matrix with the levels of one factor defining the rows and the levels of the other factor defining the columns. Each cell in the matrix corresponds to a specific combination of the two factors.
2. Traditionally, the two factors are identified as factor *A* and factor *B*. The purpose of the ANOVA is to determine whether there are any significant mean differences among the treatment conditions or cells in the experimental matrix. These treatment effects are classified as follows:
   a. The *A*-effect: Overall mean differences among the levels of factor *A*.
   b. The *B*-effect: Overall mean differences among the levels of factor *B*.
   c. The $A \times B$ interaction: Extra mean differences that are not accounted for by the main effects.
3. The two-factor ANOVA produces three *F*-ratios: one for factor *A*, one for factor *B*, and one for the $A \times B$ interaction. Each *F*-ratio has the same basic structure:

$$F = \frac{MS_{\text{treatment effect}}(\text{either } A \text{ or } B \text{ or } A \times B)}{MS_{\text{within treatments}}}$$

The formulas for the *SS*, *df*, and *MS* values for the two-factor ANOVA are presented in Figure 14.6.

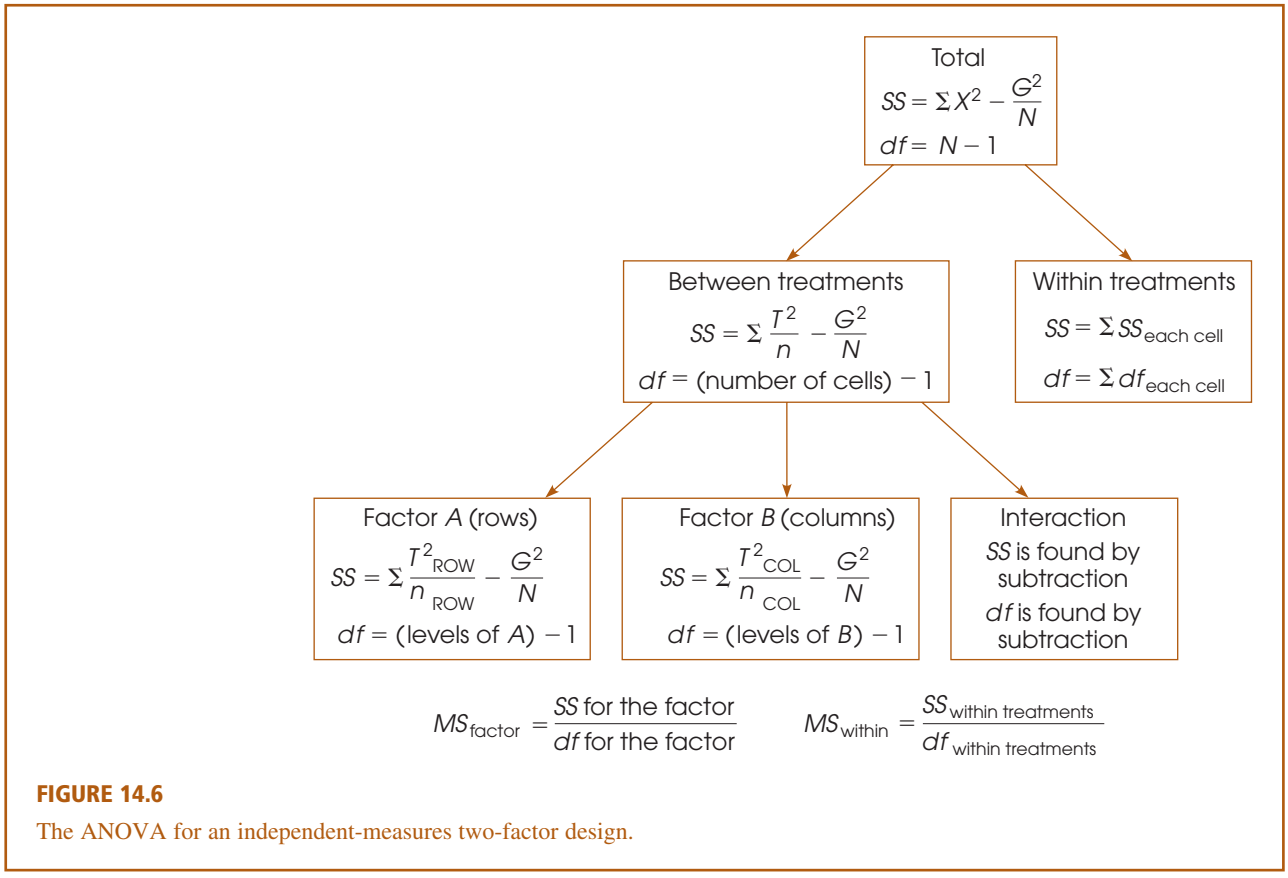


FIGURE 14.6
The ANOVA for an independent-measures two-factor design.

KEY TERMS

two-factor design (467)
matrix (467)
cells (467)
main effect (469)
interaction (470)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 14 on the book companion website. The website also provides access to two workshops entitled *Two Way ANOVA* and *Factorial ANOVA* that both review the two-factor analysis presented in this chapter.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGE brain .com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform the **Two-Factor, Independent-Measures Analysis of Variance (ANOVA)** presented in this chapter.

Data Entry

1. The scores are entered into the SPSS data editor in a *stacked format,* which means that all of the scores from all of the different treatment conditions are entered in a single column (VAR00001).
2. In a second column (VAR00002), enter a code number to identify the level of factor *A* for each score. If factor *A* defines the rows of the data matrix, enter a 1 beside each score from the first row, enter a 2 beside each score from the second row, and so on.
3. In a third column (VAR00003), enter a code number to identify the level of factor *B* for each score. If factor *B* defines the columns of the data matrix, enter a 1 beside each score from the first column, enter a 2 beside each score from the second column, and so on.

   Thus, each row of the SPSS data editor has one score and two code numbers, with the score in the first column, the code for factor *A* in the second column, and the code for factor *B* in the third column.

Data Analysis

1. Click **Analyze** on the tool bar, select **General Linear Model,** and click on **Univariant.**
2. Highlight the column label for the set of scores (VAR0001) in the left box and click the arrow to move it into the **Dependent Variable** box.
3. One by one, highlight the column labels for the two factor codes and click the arrow to move them into the **Fixed Factors** box.
4. If you want descriptive statistics for each treatment, click on the **Options** box, select **Descriptives,** and click **Continue.**
5. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the arousal-and-task-difficulty study in Example 14.1 and part of the program output is shown in Figure 14.7. The output begins with a table listing the factors (not shown in Figure 14.7), followed by a table showing descriptive statistics, including the mean and standard deviation for each cell, or treatment condition. The results of the ANOVA are shown in the table labeled **Tests of Between-Subjects Effects.** The top row (*Corrected Model*) presents the between-treatments *SS* and *df* values. The second row (*Intercept*) is not relevant for our purposes. The next three rows present the two main effects and the interaction (the *SS*, *df*, and *MS* values, as well as the *F*-ratio and the level of significance), with each factor identified by its column number from the SPSS data editor. The next row (*Error*) describes the error term (denominator of the *F*-ratio), and the final row (*Corrected Total*) describes the total variability for the entire set of scores. (Ignore the row labeled *Total.*)

FOCUS ON PROBLEM SOLVING

1. Before you begin a two-factor ANOVA, take time to organize and summarize the data. It is best if you summarize the data in a matrix with rows corresponding to the levels of one factor and columns corresponding to the levels of the other factor. In each cell of the matrix, show the number of scores (n), the total and mean for the cell, and the *SS* within the cell. Also compute the row totals and column totals that are needed to calculate main effects.
2. For a two-factor ANOVA, there are three separate *F*-ratios. These three *F*-ratios use the same error term in the denominator (MS_{within}). On the other hand, these *F*-ratios have different numerators and may have different *df* values associated with each of these numerators. Therefore, you must be careful when you look up the critical *F* values in the table. The two factors and the interaction may have different critical *F* values.

DEMONSTRATION 14.1

TWO-FACTOR ANOVA

The following data are representative of the results obtained in a research study examining the relationship between eating behavior and body weight (Schachter, 1968). The two factors in this study were:

1. The participant's weight (normal or obese)
2. The participant's state of hunger (full stomach or empty stomach)

Descriptive Statistics

Dependent Variable: VAR00001

VAR00002	VAR00003	Mean	Std. Deviation	N
1.00	1.00	3.0000	2.12132	5
	2.00	5.0000	2.64575	5
	3.00	10.0000	2.54951	5
	Total	6.0000	3.79850	15
2.00	1.00	1.0000	1.41421	5
	2.00	3.0000	2.23607	5
	3.00	2.0000	2.23607	5
	Total	2.0000	2.03540	15
1.00	1.00	2.0000	2.00000	10
	2.00	4.0000	2.53859	10
	3.00	6.0000	4.78423	10
	Total	4.0000	3.61987	30

Tests of Between-Subjects Effects

Dependent Variable: VAR00001

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	260.000	5	52.000	10.400	.000
Intercept	480.000	1	480.000	96.000	.000
VAR00002	120.000	1	120.000	24.000	.000
VAR00003	80.000	2	40.000	8.000	.002
VAR00002 * VAR00003	60.000	2	30.000	6.000	.008
Error	120.000	24	5.000		
Total	860.000	30			
Corrected Total	380.000	29			

FIGURE 14.7

Portions of the SPSS output for the two-factor ANOVA for the arousal-and-task-difficulty study in Example 14.1.

All participants were led to believe that they were taking part in a taste test for several types of crackers, and they were allowed to eat as many crackers as they wanted. The dependent variable was the number of crackers eaten by each participant. There were two specific predictions for this study. First, it was predicted that normal participants' eating behavior would be determined by their state of hunger. That is, people with empty

stomachs would eat more and people with full stomachs would eat less. Second, it was predicted that eating behavior for obese participants would not be related to their state of hunger. Specifically, it was predicted that obese participants would eat the same amount whether their stomachs were full or empty. Note that the researchers are predicting an interaction: The effect of hunger will be different for the normal participants and the obese participants. The data are as follows:

		Factor *B*: Hunger			
		Empty stomach	Full stomach		
Factor *A*: Weight	Normal	$n = 20$ $M = 22$ $T = 440$ $SS = 1540$	$n = 20$ $M = 15$ $T = 300$ $SS = 1270$	$T_{normal} = 740$	$G = 1440$ $N = 80$ $\Sigma X^2 = 31{,}836$
	Obese	$n = 20$ $M = 17$ $T = 340$ $SS = 1320$	$n = 20$ $M = 18$ $T = 360$ $SS = 1266$	$T_{normal} = 700$	
		$T_{empty} = 780$	$T_{full} = 660$		

STEP 1 **State the hypotheses, and select alpha.** For a two-factor study, there are three separate hypotheses, the two main effects and the interaction.

For factor *A,* the null hypothesis states that there is no difference in the amount eaten for normal participants versus obese participants. In symbols,

$$H_0: \mu_{normal} = \mu_{obese}$$

For factor *B,* the null hypothesis states that there is no difference in the amount eaten for full-stomach versus empty-stomach conditions. In symbols,

$$H_0: \mu_{full} = \mu_{empty}$$

For the $A \times B$ interaction, the null hypothesis can be stated two different ways. First, the difference in eating between the full-stomach and empty-stomach conditions will be the same for normal and obese participants. Second, the difference in eating between the normal and obese participants will be the same for the full-stomach and empty-stomach conditions. In more general terms,

> H_0: The effect of factor *A* does not depend on the levels of factor *B* (and *B* does not depend on *A*).

We use $\alpha = .05$ for all tests.

STEP 2 **The two-factor analysis.** Rather than compute the *df* values and look up critical values for *F* at this time, we proceed directly to the ANOVA.

STAGE 1 The first stage of the analysis is identical to the independent-measures ANOVA presented in Chapter 12, where each cell in the data matrix is considered a separate treatment condition.

$$SS_{\text{total}} = \Sigma X^2 - \frac{G^2}{N}$$
$$= 31{,}836 - \frac{1440^2}{80} = 5916$$

$$SS_{\text{within treatments}} = \Sigma SS_{\text{inside each treatment}} = 1540 + 1270 + 1320 + 1266 = 5396$$

$$SS_{\text{between treatments}} = \Sigma \frac{T^2}{n} - \frac{G^2}{N}$$
$$= \frac{440^2}{20} + \frac{300^2}{20} + \frac{340^2}{20} + \frac{360^2}{20} - \frac{1440^2}{80}$$
$$= 520$$

The corresponding degrees of freedom are

$$df_{\text{total}} = N - 1 = 79$$
$$df_{\text{within treatments}} = \Sigma df = 19 + 19 + 19 + 19 = 76$$
$$df_{\text{between treatments}} = \text{number of treatments} - 1 = 3$$

STAGE 2 The second stage of the analysis partitions the between-treatments variability into three components: the main effect for factor A, the main effect for factor B, and the $A \times B$ interaction.

For factor A (normal/obese),

$$SS_A = \Sigma \frac{T_{\text{ROWS}}^2}{n_{\text{ROWS}}} - \frac{G^2}{N}$$
$$= \frac{740^2}{40} + \frac{700^2}{40} - \frac{1440^2}{80}$$
$$= 20$$

For factor B (full/empty),

$$SS_B = \Sigma \frac{T_{\text{COLS}}^2}{n_{\text{COLS}}} - \frac{G^2}{N}$$
$$= \frac{780^2}{40} + \frac{660^2}{40} - \frac{1440^2}{80}$$
$$= 180$$

For the $A \times B$ interaction,

$$SS_{A\times B} = SS_{\text{between treatments}} - SS_A \times SS_B$$
$$= 520 - 20 - 180$$
$$= 320$$

The corresponding degrees of freedom are

$$df_A = \text{number of rows} - 1 = 1$$

$$df_B = \text{number of columns} - 1 = 1$$

$$df_{A \times B} = df_{\text{between treatments}} - df_A - df_B$$

$$= 3 - 1 - 1$$

$$= 1$$

The *MS* values needed for the *F*-ratios are

$$MS_A = \frac{SS_A}{df_A} = \frac{20}{1} = 20$$

$$MS_B = \frac{SS_B}{df_B} = \frac{180}{1} = 180$$

$$MS_{A \times B} = \frac{SS_{A \times B}}{df_{A \times B}} = \frac{320}{1} = 320$$

$$MS_{\text{within treatments}} = \frac{SS_{\text{within treatments}}}{df_{\text{within treatments}}} = \frac{5396}{76} = 71$$

Finally, the *F*-ratios are

$$F_A = \frac{MS_A}{MS_{\text{within treatments}}} = \frac{20}{71} = 0.28$$

$$F_B = \frac{MS_B}{MS_{\text{within treatments}}} = \frac{180}{71} = 2.54$$

$$F_{A \times B} = \frac{MS_{A \times B}}{MS_{\text{within treatments}}} = \frac{320}{71} = 4.51$$

STEP 3 **Make a decision and state a conclusion.** All three *F*-ratios have $df = 1, 76$. With $\alpha = .05$, the critical *F* value is 3.98 for all three tests.

For these data, factor *A* (weight) has no significant effect; $F(1, 76) = 0.28$. Statistically, there is no difference in the number of crackers eaten by normal versus obese participants.

Similarly, factor *B* (fullness) has no significant effect; $F(1, 76) = 2.54$. Statistically, the number of crackers eaten by full participants is no different from the number eaten by hungry participants. (*Note:* This conclusion concerns the combined group of normal and obese participants. The interaction concerns these two groups separately.)

These data produce a significant interaction; $F(1, 76) = 4.51$, $p < .05$. This means that the effect of fullness does depend on weight. A closer look at the original data shows that the degree of fullness did affect the normal participants, but it had no effect on the obese participants.

DEMONSTRATION 14.2

MEASURING EFFECT SIZE FOR THE TWO-FACTOR ANOVA

Effect size for each main effect and for the interaction is measured by eta squared (η^2), the percentage of variance explained by the specific main effect or interaction. In each case, the variability that is explained by other sources is removed before the percentage is computed. For the two-factor ANOVA in Demonstration 14.1,

$$\text{For factor } A,\ \eta^2 = \frac{SS_A}{SS_{\text{total}} - SS_B - SS_{A\times B}} = \frac{20}{5916 - 180 - 320} = 0.004 \left(\text{or } 0.4\%\right)$$

$$\text{For factor } B,\ \eta^2 = \frac{SS_B}{SS_{\text{total}} - SS_A - SS_{A\times B}} = \frac{180}{5916 - 20 - 320} = 0.032 \left(\text{or } 3.2\%\right)$$

$$\text{For } A\times B,\ \eta^2 = \frac{SS_{A\times B}}{SS_{\text{total}} - SS_A - SS_B} = \frac{320}{5916 - 20 - 180} = 0.056 \left(\text{or } 5.6\%\right)$$

PROBLEMS

1. Define each of the following terms:
- **a.** Factor
- **b.** Level
- **c.** Two-factor study

2. The structure of a two-factor study can be presented as a matrix with the levels of one factor determining the rows and the levels of the second factor determining the columns. With this structure in mind, describe the mean differences that are evaluated by each of the three hypothesis tests that make up a two-factor ANOVA.

3. Briefly explain what happens during the second stage of the two-factor ANOVA.

4. For the data in the following matrix:

	No Treatment	Treatment	
Male	$M = 5$	$M = 3$	Overall $M = 4$
Female	$M = 9$	$M = 13$	Overall $M = 11$
	overall $M = 7$	overall $M = 8$	

- **a.** Which two means are compared to describe the treatment main effect?
- **b.** Which two means are compared to describe the gender main effect?
- **c.** Is there an interaction between gender and treatment? Explain your answer.

5. The following matrix presents the results from an independent-measures, two-factor study with a sample of $n = 10$ participants in each treatment condition. Note that one treatment mean is missing.

		Factor B	
		B_1	B_2
Factor A	A_1	$M = 20$	$M = 30$
	A_2	$M = 40$	

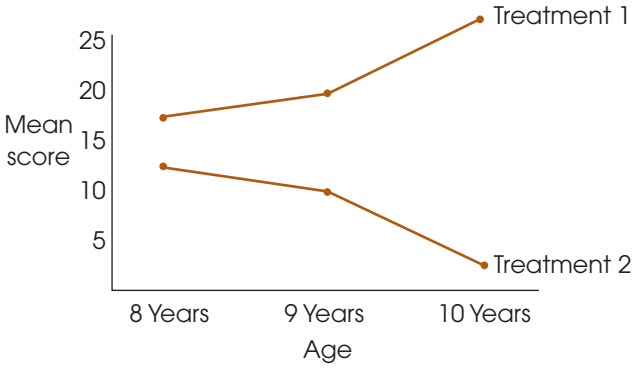
- **a.** What value for the missing mean would result in no main effect for factor A?
- **b.** What value for the missing mean would result in no main effect for factor B?
- **c.** What value for the missing mean would result in no interaction?

6. The following matrix presents the results of a two-factor study with $n = 10$ scores in each of the six treatment conditions. Note that one of the treatment means is missing.

		Factor B		
		B_1	B_2	B_3
Factor A	A_1	$M = 10$	$M = 20$	$M = 40$
	A_2	$M = 20$	$M = 30$	

a. What value for the missing mean would result in no main effect for factor *A*?
b. What value for the missing mean would result in no interaction?

7. For the data in the following graph:
a. Is there a main effect for the treatment factor?
b. Is there a main effect for the age factor?
c. Is there an interaction between age and treatment?



8. A researcher conducts an independent-measures, two-factor study using a separate sample of $n = 15$ participants in each treatment condition. The results are evaluated using an ANOVA and the researcher reports an *F*-ratio with $df = 1, 84$ for factor *A*, and an *F*-ratio with $df = 2, 84$ for factor *B*.
a. How many levels of factor *A* were used in the study?
b. How many levels of factor *B* were used in the study?
c. What are the *df* values for the *F*-ratio evaluating the interaction?

9. The following results are from an independent-measures, two-factor study with $n = 10$ participants in each treatment condition.

		Factor *B*	
		B_1	B_2
Factor *A*	A_1	T = 40 $M = 4$ $SS = 50$	$T = 10$ $M = 1$ $SS = 30$
	A_2	$T = 50$ $M = 5$ $SS = 60$	$T = 20$ $M = 2$ $SS = 40$

$N = 40$
$G = 120$
$\Sigma X^2 = 640$

a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the main effects and the interaction.
b. Compute η^2 to measure the effect size for each of the main effects and the interaction.

10. The following results are from an independent-measures, two-factor study with $n = 5$ participants in each treatment condition.

		Factor *B*		
		B_1	B_2	B_3
Factor *A*	A_1	T = 25 $M = 5$ $SS = 30$	$T = 40$ $M = 8$ $SS = 38$	$T = 70$ $M = 14$ $SS = 46$
	A_2	$T = 15$ $M = 3$ $SS = 22$	$T = 20$ $M = 4$ $SS = 26$	$T = 40$ $M = 8$ $SS = 30$

$N = 40$
$G = 120$
$\Sigma X^2 = 640$

a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the main effects and the interaction.
b. Test the simple main effects using $\alpha = .05$ to evaluate the mean difference between treatment A_1 and A_2 for each level of factor *B*.

11. A researcher conducts an independent-measures, two-factor study with two levels of factor *A* and three levels of factor *B*, using a sample of $n = 12$ participants in each treatment condition.
a. What are the *df* values for the *F*-ratio evaluating the main effect of factor *A*?
b. What are the *df* values for the *F*-ratio evaluating the main effect of factor *B*?
c. What are the *df* values for the *F*-ratio evaluating the interaction?

12. Most sports injuries are immediate and obvious, like a broken leg. However, some can be more subtle, like the neurological damage that may occur when soccer players repeatedly head a soccer ball. To examine long-term effects of repeated heading, Downs and Abwender (2002) examined two different age groups of soccer players and swimmers. The dependent variable was performance on a conceptual thinking task. Following are hypothetical data, similar to the research results.
a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the main effects and interaction.
b. Calculate the effects size (η^2) for the main effects and the interaction.
c. Briefly describe the outcome of the study.

		Factor *B*: Age	
		College	Older
Factor *A*: Sport	Soccer	$n = 20$ $M = 9$ $T = 180$ $SS = 380$	$n = 20$ $M = 4$ $T = 80$ $SS = 390$
	Swimming	$n = 20$ $M = 9$ $T = 180$ $SS = 350$	$n = 20$ $M = 8$ $T = 160$ $SS = 400$

$\Sigma X^2 = 6360$

13. Some people like to pour beer gently down the side of the glass to preserve bubbles. Others splash it down the center to release the bubbles into a foamy head and free the aromas. Champagne, however is best when the bubbles remain concentrated in the wine. A group of French scientists recently verified the difference between the two pouring methods by measuring the amount of bubbles in each glass of champagne poured two different ways and at three different temperatures (Liger-Belair, 2010). The following data present the pattern of results obtained in the study.

	Champagne Temperature (°F)		
	40°	46°	52°
Gentle Pour	$n = 10$ $M = 7$ $SS = 64$	$n = 10$ $M = 3$ $SS = 57$	$n = 10$ $M = 2$ $SS = 47$
Splashing Pour	$n = 10$ $M = 5$ $SS = 56$	$n = 10$ $M = 1$ $SS = 54$	$n = 10$ $M = 0$ $SS = 46$

a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the mean differences.

b. Briefly explain how temperature and pouring influence the bubbles in champagne according to this pattern of results.

14. The following table summarizes the results from a two-factor study with 2 levels of factor *A* and 3 levels of factor *B* using a separate sample of $n = 8$ participants in each treatment condition. Fill in the missing values. (*Hint:* Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	60	____		
Factor *A*	____	____	5	$F =$ ____
Factor *B*	____	____	____	$F =$ ____
A × *B* Interaction	25	____	____	$F =$ ____
Within treatments	____	____	2.5	
Total	____	____		

15. The following table summarizes the results from a two-factor study with 3 levels of factor *A* and 3 levels of factor *B* using a separate sample of $n = 9$ participants in each treatment condition. Fill in the missing values. (*Hint:* Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	144	____		
Factor *A*	____	____	18	$F =$ ____
Factor *B*	____	____	____	$F =$ ____
A × *B* Interaction	____	____	____	$F = 7.0$
Within treatments	____	____	____	
Total	360	____		

16. The Preview section for this chapter described a two-factor study examining performance under two audience conditions (factor *B*) for high and low self-esteem participants (factor *A*). The following summary table presents possible results from the analysis of that study. Assuming that the study used a separate sample of $n = 15$ participants in each treatment condition (each cell), fill in the missing values in the table. (*Hint*: Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	67	____		
Audience	____	____	____	$F =$ ____
Self-esteem	29	____	____	$F =$ ____
Interaction	____	____	____	$F = 5.50$
Within treatments	____	____	4	
Total	____	____		

17. The following table summarizes the results from a two-factor study with 2 levels of factor *A* and 3 levels of factor *B* using a separate sample of $n = 11$ participants in each treatment condition. Fill in the missing values. (*Hint:* Start with the *df* values.)

Source	*SS*	*df*	*MS*	
Between treatments	____	____		
Factor *A*	____	____	____	$F = 7$
Factor *B*	____	____	____	$F = 8$
A × *B* Interaction	____	____	____	$F = 3$
Within treatments	240	____	____	
Total	____	____		

18. The following data are from a two-factor study examining the effects of two treatment conditions on males and females.

a. Use an ANOVA with $\alpha = .05$ for all tests to evaluate the significance of the main effects and the interaction.

b. Compute η^2 to measure the size of the effect for each main effect and the interaction.

Factor *A*: Gender	Treatment I	Treatment II	
Male	3 8 9 4 $M = 6$ $T = 24$ $SS = 26$	2 8 7 7 $M = 6$ $T = 24$ $SS = 22$	$T_{male} = 48$
Female	0 0 2 6 $M = 2$ $T = 8$ $SS = 24$	12 6 9 13 $M = 10$ $T = 40$ $SS = 30$	$T_{female} = 48$
	$T_I = 32$	$T_{II} = 64$	

$N = 16$
$G = 96$
$\Sigma X^2 = 806$

19. The following data are from a two-factor study examining the effects of three treatment conditions on males and females.

a. Use an ANOVA with $\alpha = .05$ for all tests to evaluate the significance of the main effects and the interaction.

b. Test the simple main effects using $\alpha = .05$ to evaluate the mean difference between males and females for each of the three treatments.

Factor *A*: Gender	Factor *B* Treatment I	Factor *B* Treatment II	Factor *B* Treatment III	
Male	1 2 6 $M = 3$ $T = 9$ $SS = 14$	7 2 9 $M = 6$ $T = 18$ $SS = 26$	9 11 7 $M = 9$ $T = 27$ $SS = 8$	$T_{male} = 54$
Female	3 1 5 $M = 3$ $T = 9$ $SS = 8$	10 11 15 $M = 12$ $T = 36$ $SS = 14$	16 18 11 $M = 15$ $T = 45$ $SS = 26$	$T_{female} = 90$

$N = 18$
$G = 144$
$\Sigma X^2 = 1608$

20. Mathematics word problems can be particularly difficult, especially for primary-grade children. A recent study investigated a combination of techniques for teaching students to master these problems (Fuchs, Fuchs, Craddock, Hollenbeck, Hamlett, & Schatschneider, 2008). The study investigated the effectiveness of small-group tutoring and the effectiveness of a classroom instruction technique known as "hot math." The hot-math program teaches students to recognize types or categories of problems so that they can generalize skills from one problem to another. The following data are similar to the results obtained in the study. The dependent variable is a math test score for each student after 16 weeks in the study.

	No Tutoring	With Tutoring
Traditional Instruction	3 6 2 2 4 7	9 4 5 8 4 6
Hot-Math Instruction	7 7 2 6 8 6	8 12 9 13 9 9

a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the significance of the main effects and the interaction.
b. Calculate the η^2 values to measure the effect size for the two main effects.
c. Describe the pattern of results. (Is tutoring significantly better than no tutoring? Is traditional classroom instruction significantly different from hot math? Does the effect of tutoring depend on the type of classroom instruction?)

21. In Chapter 12 (p. 432), we described a study reporting that college students who are on Facebook (or have it running in the background) while studying had lower grades than students who did not use the social network (Kirschner & Karpinski, 2010). A researcher would like to know if the same result extends to students in lower grade levels. The researcher planned a two-factor study comparing Facebook users with non-users for middle school students, high school students, and college students. For consistency across groups, grades were converted into six categories, numbered 0 to 5 from low to high. The results are presented in the following matrix.

a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the mean differences.
b. Describe the pattern of results.

	Middle School	High School	College
User	3 5 5 3	5 5 2 4	5 4 2 5
Non-user	5 3 2 2	1 2 3 2	1 0 0 3

22. In Chapter 11, we described a research study in which the color red appeared to increase men's attraction to women (Elliot & Niesta, 2008). The same researchers have published other results showing that red also increases women's attraction to men but does not appear to affect judgments of same sex individuals (Elliot, et al., 2010). Combining these results into one study produces a two-factor design in which men judge photographs of both women and men, which are shown on both red and white backgrounds. The dependent variable is a rating of attractiveness for the person shown in the photograph. The study uses a separate group of participants for each condition. The following table presents data similar to the results from previous research.

a. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the main effects and the interaction.

		Person Shown in Photograph	
		Female	Male
Background Color for Photograph	White	$n = 10$ $M = 4.5$ $SS = 6$	$n = 10$ $M = 4.4$ $SS = 7$
	Red	$n = 10$ $M = 7.5$ $SS = 9$	$n = 10$ $M = 4.6$ $SS = 8$

b. Describe the effect of background color on judgments of males and females.

23. In the Preview section of this chapter, we presented an experiment that examined the effect of an audience on the performance of two different personality types. Data from this experiment are as follows. The dependent variable is the number of errors made by each participant.

		No Audience	Audience
Self-Esteem	High	3 6 2 2 4 7	9 4 5 8 4 6
	Low	7 7 2 6 8 6	10 14 11 15 11 11

a. Use an ANOVA with $\alpha = .05$ to evaluate the data. Describe the effect of the audience and the effect of self-esteem on performance.
b. Calculate the effect size (η^2) for each main effect and for the interaction.

PART IV

REVIEW

After completing this part, you should be able to perform an ANOVA to evaluate the significance of mean differences in three research situations. These include:

1. The single-factor independent-measures design introduced in Chapter 12.
2. The single-factor repeated-measures design introduced in Chapter 13.
3. The two-factor independent-measures design introduced in Chapter 14.

In this part we introduce three applications of ANOVA that use an F-ratio statistic to evaluate the mean differences among two or more populations. In each case, the F-ratio has the following structure:

$$F = \frac{\text{variance between treatments}}{\text{variance from random unsystematic sources}}$$

The numerator of the *F*-ratio measures the mean differences that exist from one treatment condition to another, including any systematic differences caused by the treatments. The denominator measures the differences that exist when there are no systematic factors that cause one score to be different from another. The *F*-ratio is structured so that the numerator and denominator are measuring exactly the same variance when the null hypothesis is true and there are no systematic treatment effects. In this case, the *F*-ratio should have a value near 1.00. Thus, an *F*-ratio near 1.00 is evidence that the null hypothesis is true. Similarly, an *F*-ratio that is much larger than 1.00 provides evidence that a systematic treatment effect does exist and the null hypothesis should be rejected.

For independent-measures designs, either single-factor or two-factor, the denominator of the *F*-ratio is obtained by computing the variance within treatments. Inside each treatment condition all participants are treated exactly the same, so there are no systematic treatment effects that cause the scores to vary.

For a repeated-measures design, the same individuals are used in every treatment condition, so any differences between treatments cannot be caused by individual differences. Thus, the numerator of the *F*-ratio does not include any individual differences. Therefore, individual differences must also be eliminated from the denominator to balance the *F*-ratio. As a result, the repeated-measures ANOVA is a two-stage process. The first stage separates the between-treatments variance (numerator) and the within-treatments variance. The second stage removes the systematic individual differences from the within-treatments variance to produce the appropriate denominator for the *F*-ratio.

For a two-factor design, the mean differences between treatments can be caused by either of the two factors or by specific combinations of factors. The goal of the ANOVA is to separate these possible treatment effects so that each can be evaluated independent of the others. To accomplish this, the two-factor ANOVA is a two-stage process. The first stage separates the between-treatments variance and the within-treatments variance (denominator). The second stage analyzes the between-treatments variance into three components: the main effect from the first factor, the main effect from the second factor, and the interaction.

Note that the repeated-measures ANOVA and the two-factor ANOVA are both two-stage processes. Both begin by separating the total variance into between-treatments and within-treatments variance. However, the second stages of these two ANOVAs serve different purposes and focus on different components. The repeated-measures ANOVA focuses on the within-treatments variance and removes the individual differences to obtain the error variance. The two-factor ANOVA separates the between-treatments variance into the two main effects and the interaction.

REVIEW EXERCISES

1. Recent research indicates that the effectiveness of antidepressant medication is directly related to the severity of the depression (Khan, Brodhead, Kolts, & Brown, 2005). Based on pre-treatment depression scores, patients were divided into four groups by level of depression. After receiving the antidepressant medication, depression scores were measured again and the amount of improvement was recorded for each patient. The following data are similar to the results of the study.

a. Do the data indicate significant differences among the four levels of severity? Test with $\alpha = .05$.

b. Compute η^2, the percentage of variance explained by the group differences.

c. Write a sentence demonstrating how the outcome of the hypothesis test and the measure of effect size would appear in a research report.

Low Moderate	High Moderate	Moderately Severe	Severe	
0	1	4	5	$N = 16$
2	3	6	6	$G = 48$
2	2	2	6	$\Sigma X^2 = 204$
0	2	4	3	
$M = 1$	$M = 2$	$M = 4$	$M = 5$	
$T = 4$	$T = 8$	$T = 16$	$T = 20$	
$SS = 4$	$SS = 2$	$SS = 8$	$SS = 6$	

2. Loss of hearing can be a significant problem for older adults. Although hearing aids can correct the physical

problem, people who have lived with hearing impairment often develop poor communication strategies and social skills. To address this problem, a home education program has been developed to help people who are receiving hearing aids for the first time. The program emphasizes communication skills. To evaluate the program, overall quality of life and satisfaction were measured before treatment, again at the end of the training program, and once more at a 6-month follow-up (Kramer, Allessie, Dondorp, Zekveld, & Kapteyn, 2005). Data similar to the results obtained in the study are shown below.

	Quality-of-Life Scores			
Person	Before	After	6 Months	
A	3	7	8	$N = 12$
B	0	5	7	$G = 60$
C	4	9	5	$\Sigma X^2 = 384$
D	1	7	4	
	$T = 8$	$T = 28$	$T = 24$	
	$SS = 10$	$SS = 8$	$SS = 10$	

a. Do the data indicate a significant improvement in the quality of life following the training program? Test at the .05 level of significance.

b. Calculate η^2 to measure the size of the effect.

c. Write a sentence demonstrating how the outcome of the hypothesis test and the measure of effect size would appear in a research report.

3. Briefly describe what is meant by an interaction between factors in a two-factor research study.

4. A recent study of driving behavior suggests that self-reported measures of high driving skills and low ratings of safety skills create a dangerous combination (Sümer, Özkan, & Lajunen, 2006). (*Note:* Those who rate themselves as highly skilled drivers are probably overly confident.) Drivers were classified as high or low in self-rated skill based on responses to a driver-skill inventory, then classified as high or low in safety skill based on responses to a driver-aggression scale. An overall measure of driving risk was obtained by combining several variables such as number of accidents, tickets, tendency to speed, and tendency to pass other cars. The following data represent results similar to those obtained in the study. Use a two-factor ANOVA with $\alpha = .05$ to evaluate the results.

		Self-Rated Driving Skill		
		Low	High	
Driving Safety	Low	$n = 8$ $M = 5$ $T = 40$ $SS = 52$	$n = 8$ $M = 8.5$ $T = 68$ $SS = 71$	$N = 32$ $G = 160$ $\Sigma X^2 = 1151$
	High	$n = 8$ $M = 3$ $T = 24$ $SS = 34$	$n = 8$ $M = 3.5$ $T = 28$ $SS = 46$	

PART

V

Correlations and Nonparametric Tests

Back in Chapter 1 we stated that the primary goal of science is to establish relationships between variables. Until this point, the statistics we have presented all attempt to accomplish this goal by comparing *groups of scores* using *means and variances* as the basic statistical measures. Typically, one variable is used to define the groups, and a second variable is measured to obtain a set of scores within each group. Means and variances are then computed for the scores, and the sample means are used to test hypotheses about population means. If the hypothesis test indicates a significant mean difference, then we conclude that there is a relationship between the variables.

However, many research situations do not involve comparing groups, and many do not produce data that allow you to calculate means and variances. For example, a researcher can investigate the relationship between two variables (for example, IQ and creativity) by measuring both variables within a single group of individuals. Also, the measurement procedure may not produce numerical scores. For example, participants can indicate their color preferences by simply picking a favorite color or by ranking several choices. Without numerical scores, it is impossible to calculate means and variances. Instead, the data consist of proportions or frequencies. For example, a research study may investigate what proportion of people select red as their favorite color and whether this proportion is different for introverted people compared with extroverted people.

Notice that these new research situations are still asking questions about the relationships between variables, and they are still using sample data to make inferences about populations. However, they are no longer comparing groups and they are no longer based on means and variances. In this part, we introduce the statistical methods that have been developed for these other kinds of research.

This page intentionally left blank

CHAPTER

15

Correlation

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Sum of squares (SS) (Chapter 4)
  - Computational formula
  - Definitional formula
- *z*-scores (Chapter 5)
- Hypothesis testing (Chapter 8)

Preview

Having been a student and taken exams for much of your life, you probably have noticed a curious phenomenon. In every class, there are some students who zip through exams and turn in their papers while everyone else is still on page 1. Other students cling to their exams and are still working frantically when the instructor announces that time is up and demands that all papers be turned in. Have you wondered what grades these students receive? Are the students who finish first the best in the class, or are they completely unprepared and simply accepting their failure? Are the *A* students the last to finish because they are compulsively checking and rechecking their answers? To help answer these questions, we carefully observed a recent exam and recorded the amount of time each student spent on the exam and the grade each student received. These data are shown in Figure 15.1. Note that we have listed time along the *X*-axis and grade on the *Y*-axis. Each student is identified by a point on the graph that is located directly above the student's time and directly across from the student's grade. Also note that we have drawn a line through the middle of the data points in Figure 15.1. The line helps make the relationship between time and grade more obvious. The graph shows that the highest grades tend to go to the students who finished their exams early. Students who held their papers to the bitter end tended to have low grades.

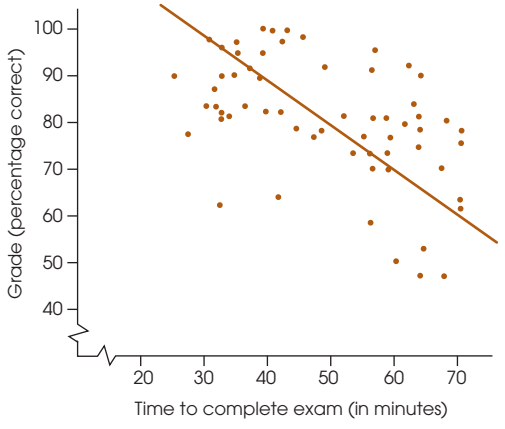


FIGURE 15.1

The relationship between exam grade and time needed to complete the exam. Notice the general trend in these data: Students who finish the exam early tend to have better grades.

The Problem: Although the data in Figure 15.1 appear to show a clear relationship, we need some procedure to measure the relationship and a hypothesis test to determine whether it is significant. In the preceding five chapters, we described relationships between variables in terms of mean differences between two or more groups of scores, and we used hypothesis tests that evaluate the significance of mean differences. For the data in Figure 15.1, there is only one group of scores, and calculating a mean is not going to help describe the relationship. To evaluate these data, a completely different approach is needed for both descriptive and inferential statistics.

The Solution: The data in Figure 15.1 are an example of the results from a correlational research study. In Chapter 1, the correlational design was introduced as a method for examining the relationship between two variables by measuring two different variables for each individual in one group of participants. The relationship obtained in a correlational study is typically described and evaluated with a statistical measure known as a *correlation*. Just as a sample mean provides a concise description of an entire sample, a correlation provides a concise description of a relationship. We look at how correlations are used and interpreted. For example, now that you have seen the relationship between time and grades, do you think it might be a good idea to start turning in your exam papers a little sooner? Wait and see.

15.1 INTRODUCTION

Correlation is a statistical technique that is used to measure and describe the relationship between two variables. Usually the two variables are simply observed as they exist naturally in the environment—there is no attempt to control or manipulate

the variables. For example, a researcher could check high school records (with permission) to obtain a measure of each student's academic performance, and then survey each family to obtain a measure of income. The resulting data could be used to determine whether there is relationship between high school grades and family income. Notice that the researcher is not manipulating any student's grade or any family's income, but is simply observing what occurs naturally. You also should notice that a correlation requires two scores for each individual (one score from each of the two variables). These scores normally are identified as *X* and *Y*. The pairs of scores can be listed in a table, or they can be presented graphically in a scatter plot (Figure 15.2). In the scatter plot, the values for the *X* variable are listed on the horizontal axis and the *Y* values are listed on the vertical axis. Each individual is then represented by a single point in the graph so that the horizontal position corresponds to the individual's *X* value and the vertical position corresponds to the *Y* value. The value of a scatter plot is that it allows you to see any patterns or trends that exist in the data. The scores in Figure 15.2, for example, show a clear relationship between family income and student grades; as income increases, grades also increase.

Person	Family Income (in $1000)	Student's Average Grade
A	31	72
B	38	86
C	42	81
D	44	78
E	49	85
F	56	80
G	58	91
H	65	89
I	70	94
J	90	83
K	92	90
L	106	97
M	135	89
N	174	95

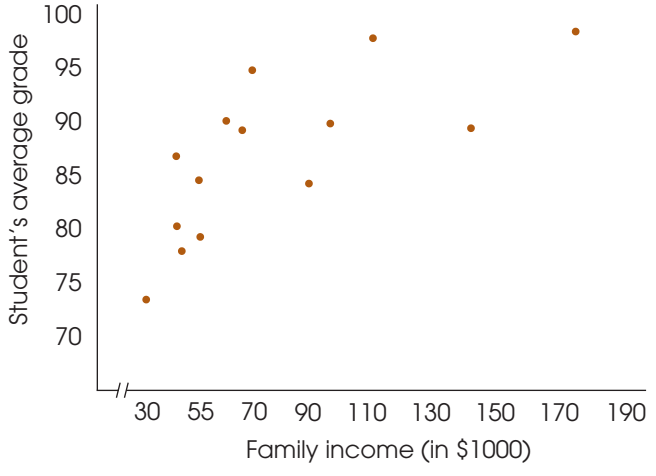


FIGURE 15.2

Correlational data showing the relationship between family income (*X*) and student grades (*Y*) for a sample of $n = 14$ high school students. The scores are listed in order from lowest to highest family income and are shown in a scatter plot.

THE CHARACTERISTICS OF A RELATIONSHIP

A correlation is a numerical value that describes and measures three characteristics of the relationship between *X* and *Y*. These three characteristics are as follows:

1. **The Direction of the Relationship.** The sign of the correlation, positive or negative, describes the direction of the relationship.

DEFINITIONS

In a **positive correlation,** the two variables tend to change in the same direction: As the value of the *X* variable increases from one individual to another, the *Y* variable also tends to increase; when the *X* variable decreases, the *Y* variable also decreases.

In a **negative correlation,** the two variables tend to go in opposite directions. As the *X* variable increases, the *Y* variable decreases. That is, it is an inverse relationship.

The following examples illustrate positive and negative relationships.

EXAMPLE 15.1

Suppose you run the drink concession at the football stadium. After several seasons, you begin to notice a relationship between the temperature at game time and the beverages you sell. Specifically, you have noted that when the temperature is low, you sell relatively little beer. However, as the temperature goes up, beer sales also go up (Figure 15.3). This is an example of a positive correlation. You also have noted a relationship between temperature and coffee sales: On cold days you sell a lot of coffee, but coffee sales go down as the temperature goes up. This is an example of a negative relationship.

2. **The Form of the Relationship.** In the preceding coffee and beer examples, the relationships tend to have a linear form; that is, the points in the scatter plot tend to cluster around a straight line. We have drawn a line through the middle

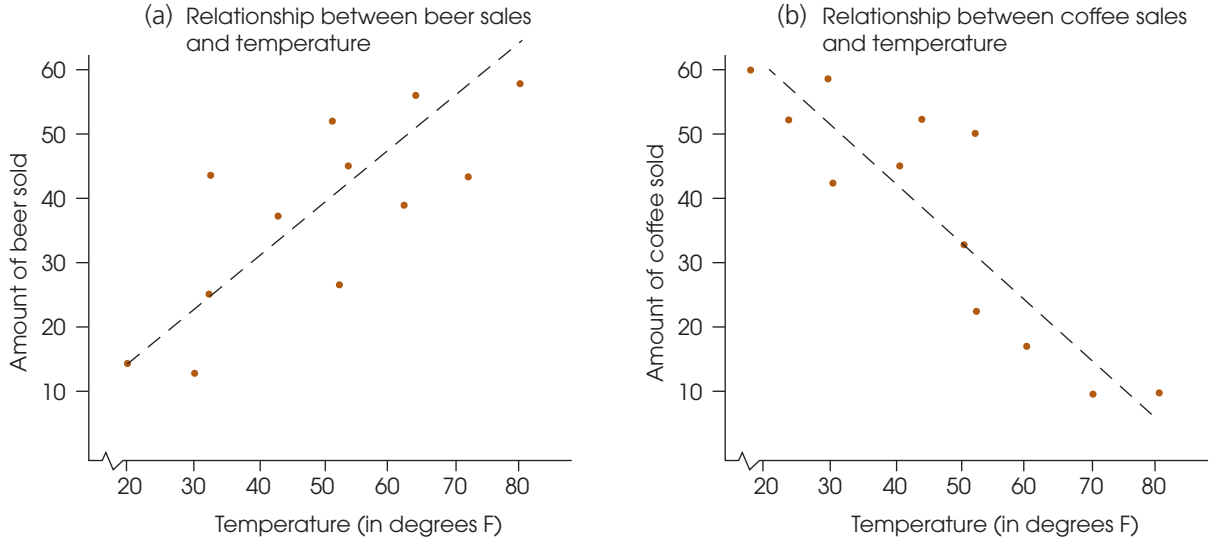


FIGURE 15.3

Examples of positive and negative relationships. (a) Beer sales are positively related to temperature. (b) Coffee sales are negatively related to temperature.

of the data points in each figure to help show the relationship. The most common use of correlation is to measure straight-line relationships. However, other forms of relationships do exist and there are special correlations used to measure them. (We examine alternatives in Section 15.5.)

3. **The Strength or Consistency of the Relationship**. Finally, the correlation measures the consistency of the relationship. For a linear elationship, for example, the data points could fit perfectly on a straight line. Every time *X* increases by one point, the value of *Y* also changes by a consistent and predictable amount. Figure 15.4(a) shows an example of a perfect linear relationship. However, relationships are usually not perfect. Although there may be a tendency for the value of *Y* to increase whenever *X* increases, the amount that *Y* changes is not always the same, and occasionally, *Y* decreases when *X* increases. In this situation, the data points do not fall perfectly on a straight line. The consistency of the relationship is measured by the numerical value of the correlation. A *perfect correlation* always is identified by a correlation of 1.00 and indicates a perfectly consistent relationship. For a correlation of 1.00 (or –1.00), each change in *X* is accompanied by a perfectly predictable change in *Y*. At the other extreme, a correlation of 0 indicates no consistency at all. For a correlation of 0, the data points are scattered randomly with no clear trend [see Figure 15.4(b)]. Intermediate values between 0 and 1 indicate the degree of consistency.

   Examples of different values for linear correlations are shown in Figure 15.4. In each example we have sketched a line around the data points. This line, called an *envelope* because it encloses the data, often helps you to see the overall trend in the data. As a rule of thumb, when the envelope is shaped roughly like a football, the correlation is around 0.7. Envelopes that are fatter than a football indicate correlations closer to 0, and narrower shapes indicate correlations closer to 1.00.

You should also note that the sign (+ or −) and the strength of a correlation are independent. For example, a correlation of 1.00 indicates a perfectly consistent relationship whether it is positive (+1.00) or negative (−1.00). Similarly, correlations of +0.80 and −0.80 are equally consistent relationships.

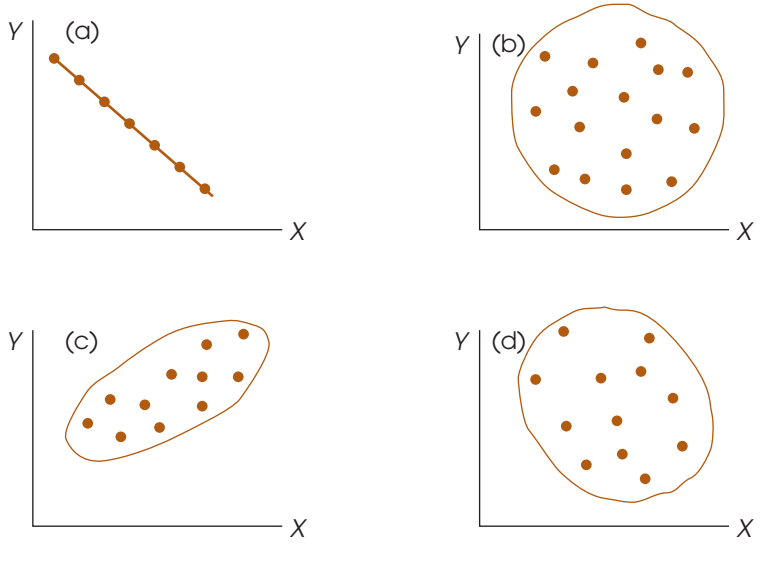


FIGURE 15.4

Examples of different values for linear correlations: (a) a perfect negative correlation, −1.00; (b) no linear trend, 0.00; (c) a strong positive relationship, approximately +0.90; (d) a relatively weak negative correlation, approximately −0.40.

LEARNING CHECK

1. For each of the following, indicate whether you would expect a positive or a negative correlation.
   - **a.** Model year and price for a used Honda
   - **b.** IQ and grade point average for high school students
   - **c.** Daily high temperature and daily energy consumption for 30 winter days in New York City
2. The data points would be clustered more closely around a straight line for a correlation of –0.80 than for a correlation of +0.05. (True or false?)
3. If the data points are clustered close to a line that slopes up from left to right, then a good estimate of the correlation would be +0.90. (True or false?)
4. If a scatter plot shows a set of data points that form a circular pattern, the correlation should be near zero. (True or false?)

ANSWERS

1. **a.** Positive: Higher model years tend to have higher prices.
   b. Positive: More intelligent students tend to get higher grades.
   c. Negative: Higher temperature tends to decrease the need for heating.
2. True. The numerical value indicates the strength of the relationship. The sign only indicates direction.
3. True.
4. True.

15.2 THE PEARSON CORRELATION

By far the most common correlation is the *Pearson correlation* (or the Pearson product–moment correlation) which measures the degree of straight-line relationship.

DEFINITION

The **Pearson correlation** measures the degree and the direction of the linear relationship between two variables.

The Pearson correlation is identified by the letter r. Conceptually, this correlation is computed by

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}}$$

$$= \frac{\text{covariability of } X \text{ and } Y}{\text{variability of } X \text{ and } Y \text{ separately}}$$

When there is a perfect linear relationship, every change in the X variable is accompanied by a corresponding change in the Y variable. In Figure 15.4(a), for example, every time the value of X increases, there is a perfectly predictable decrease in the value of Y. The result is a perfect linear relationship, with X and Y always varying together. In this case, the covariability (X and Y together) is identical to the variability of X and Y separately, and the formula produces a correlation with a magnitude of 1.00 or –1.00. At the other extreme, when there is no linear relationship, a change in the

X variable does not correspond to any predictable change in the *Y* variable. In this case, there is no covariability, and the resulting correlation is zero.

THE SUM OF PRODUCTS OF DEVIATIONS

To calculate the Pearson correlation, it is necessary to introduce one new concept: the *sum of products* of deviations, or *SP*. This new value is similar to *SS* (the sum of squared deviations), which is used to measure variability for a single variable. Now, we use *SP* to measure the amount of covariability between two variables. The value for *SP* can be calculated with either a definitional formula or a computational formula.

The *definitional formula* for the sum of products is

$$SP = \Sigma(X - M_X)(Y - M_Y) \tag{15.1}$$

where M_X is the mean for the *X* scores and M_Y is the mean for the *Y*s.

The definitional formula instructs you to perform the following sequence of operations:

1. Find the *X* deviation and the *Y* deviation for each individual.
2. Find the product of the deviations for each individual.
3. Add the products.

Notice that this process "defines" the value being calculated: the sum of the products of the deviations.

The *computational formula* for the sum of products of deviations is

$$SP = \Sigma XY - \frac{\Sigma X \Sigma Y}{n} \tag{15.2}$$

Caution: The *n* in this formula refers to the number of pairs of scores.

Because the computational formula uses the original scores (*X* and *Y* values), it usually results in easier calculations than those required with the definitional formula, especially if M_X or M_Y is not a whole number. However, both formulas always produce the same value for *SP*.

You may have noted that the formulas for *SP* are similar to the formulas you have learned for *SS* (sum of squares). The relationship between the two sets of formulas is described in Box 15.1. The following example demonstrates the calculation of *SP* with both formulas.

EXAMPLE 15.2

The same set of $n = 4$ pairs of scores is used to calculate *SP*, first using the definitional formula and then using the computational formula.

For the definitional formula, you need deviation scores for each of the *X* values and each of the *Y* values. Note that the mean for the *X*s is $M_X = 3$ and the mean for the *Y*s is $M_Y = 5$. The deviations and the products of deviations are shown in the following table:

Caution: The signs (+ and –) are critical in determining the sum of products, *SP*.

Scores		Deviations		Products
X	*Y*	$X - M_X$	$Y - M_Y$	$(X - M_X)(Y - M_Y)$
1	3	−2	−2	+4
2	6	−1	+1	−1
4	4	+1	−1	−1
5	7	+2	+2	+4
				+6 = *SP*

For these scores, the sum of the products of the deviations is $SP = +6$.

For the computational formula, you need the X value, the Y value, and the XY product for each individual. Then you find the sum of the Xs, the sum of the Ys, and the sum of the XY products. These values are as follows:

X	Y	XY	
1	3	3	
2	6	12	
4	4	16	
5	7	35	
12	20	66	Totals

Substituting the totals in the formula gives

$$SP = \Sigma XY - \frac{\Sigma X \Sigma Y}{n}$$

$$= 66 - \frac{12(20)}{4}$$

$$= 66–60$$

$$= 6$$

Both formulas produce the same result, $SP = 6$.

BOX 15.1 COMPARING THE *SP* AND *SS* FORMULAS

It will help you to learn the formulas for *SP* if you note the similarity between the two *SP* formulas and the corresponding formulas for *SS* that were presented in Chapter 4. The definitional formula for *SS* is

$$SS = \Sigma(X - M)^2$$

In this formula, you must square each deviation, which is equivalent to multiplying it by itself. With this in mind, the formula can be rewritten as

$$SS = \Sigma(X - M)(X - M)$$

The similarity between the *SS* formula and the *SP* formula should be obvious—the *SS* formula uses squares and the *SP* formula uses products. This same relationship exists for the computational formulas. For *SS,* the computational formula is

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

As before, each squared value can be rewritten so that the formula becomes

$$SS = \Sigma XX - \frac{\Sigma X \Sigma X}{n}$$

Again, note the similarity in structure between the *SS* formula and the *SP* formula. If you remember that *SS* uses squares and *SP* uses products, the two new formulas for the sum of products should be easy to learn.

CALCULATION OF THE PEARSON-CORRELATION

As noted earlier, the Pearson correlation consists of a ratio comparing the covariability of X and Y (the numerator) with the variability of X and Y separately (the denominator). In the formula for the Pearson r, we use SP to measure the covariability of X and Y. The variability of X is measured by computing SS for the X scores and the variability of Y is measured by SS for the Y scores. With these definitions, the formula for the Pearson correlation becomes

Note that you *multiply* SS for X by SS for Y in the denominator of the Pearson formula.

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} \tag{15.3}$$

The following example demonstrates the use of this formula with a simple set of scores.

EXAMPLE 15.3

X	Y
0	2
10	6
4	2
8	4
8	6

The Pearson correlation is computed for the set of $n = 5$ pairs of scores shown in the margin.

Before starting any calculations, it is useful to put the data in a scatter plot and make a preliminary estimate of the correlation. These data have been graphed in Figure 15.5. Looking at the scatter plot, it appears that there is a very good (but not perfect) positive correlation. You should expect an approximate value of $r = +0.8$ or $+0.9$. To find the Pearson correlation, we need SP, SS for X, and SS for Y. The calculations for each of these values, using the definitional formulas, are presented in Table 15.1. (Note that the mean for the X values is $M_X = 6$ and the mean for the Y scores is $M_Y = 4$.)

Using the values from Table 15.1, the Pearson correlation is

$$r = \frac{SP}{\sqrt{(SS_X)(SS_Y)}} = \frac{28}{\sqrt{(64)(16)}} = \frac{28}{32} = +0.875$$

Note that the value we obtained for the correlation is perfectly consistent with the pattern shown in Figure 15.5. First, the positive value of the correlation indicates

FIGURE 15.5
Scatter plot of the data from Example 15.3.

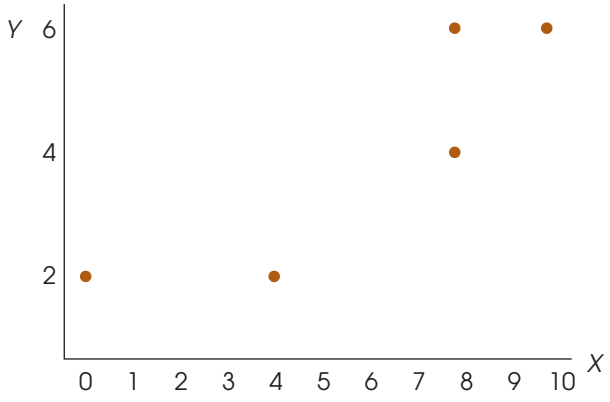


TABLE 15.1

Calculation of SS_X, SS_Y, and SP for a sample of $n = 5$ pairs of scores.

Scores		Deviations		Squared Deviations		Products
X	Y	$X - M_X$	$Y - M_Y$	$(X - M_X)^2$	$(Y - M_Y)^2$	$(X - M_X)(Y - M_Y)$
0	2	−6	−2	36	4	+12
10	6	+4	+2	16	4	+8
4	2	−2	−2	4	4	+4
8	4	+2	0	4	0	0
8	6	+2	+2	4	4	+4
				$SS_X = 64$	$SS_Y = 16$	$SP = +28$

that the points are clustered around a line that slopes up to the right. Second, the high value for the correlation (near 1.00) indicates that the points are very tightly clustered close to the line. Thus, the value of the correlation describes the relationship that exists in the data.

THE PEARSON CORRELATION AND *z*-SCORES

The Pearson correlation measures the relationship between an individual's location in the X distribution and his or her location in the Y distribution. For example, a positive correlation means that individuals who score high on X also tend to score high on Y. Similarly, a negative correlation indicates that individuals with high X scores tend to have low Y scores.

Recall from Chapter 5 that z-scores identify the exact location of each individual score within a distribution. With this in mind, each X value can be transformed into a z-score, z_X, using the mean and standard deviation for the set of Xs. Similarly, each Y score can be transformed into z_Y. If the X and Y values are viewed as a sample, the transformation is completed using the sample formula for z (Equation 5.3). If the X and Y values form a complete population, the z-scores are computed using Equation 5.1. After the transformation, the formula for the Pearson correlation can be expressed entirely in terms of z-scores.

$$\text{For a sample, } r = \frac{\Sigma z_X z_Y}{(n-1)} \tag{15.4}$$

$$\text{For a population, } \rho = \frac{\Sigma z_X z_Y}{N} \tag{15.5}$$

Note that the population value is identified with a Greek letter, in this case the letter rho (ρ), which is the Greek equivalent of the letter r.

LEARNING CHECK

1. Describe what is measured by a Pearson correlation.
2. Can SP ever have a value less than zero?
3. Calculate the sum of products of deviations (SP) for the following set of scores. Use the definitional formula and then the computational formula. Verify that you get the same answer with both formulas.

X	Y
0	1
4	3
5	3
2	2
4	1

4. For the following data:

a. Sketch a scatter plot and make an estimate of the Pearson correlation.

b. Compute the Pearson correlation.

X	Y
2	6
1	5
3	3
0	7
4	4

ANSWERS

1. The Pearson correlation measures the degree and direction of the linear relationship between two variables.

2. Yes. *SP* can be positive, negative, or zero depending on the relationship between *X* and *Y*.

3. $SP = 5$

4. $r = -8/10 = -0.80$

15.3 USING AND INTERPRETING THE PEARSON CORRELATION

WHERE AND WHY CORRELATIONS ARE USED

Although correlations have a number of different applications, a few specific examples are presented next to give an indication of the value of this statistical measure.

1. **Prediction.** If two variables are known to be related in a systematic way, then it is possible to use one of the variables to make accurate predictions about the other. For example, when you applied for admission to college, you were required to submit a great deal of personal information, including your scores on the Scholastic Achievement Test (SAT). College officials want this information so that they can predict your chances of success in college. It has been demonstrated over several years that SAT scores and college grade point averages are correlated. Students who do well on the SAT tend to do well in college; students who have difficulty with the SAT tend to have difficulty in college. Based on this relationship, college admissions officers can make a prediction about the potential success of each applicant. You should note that this prediction is not perfectly accurate. Not everyone who does poorly on the

SAT has trouble in college. That is why you also submit letters of recommendation, high school grades, and other information with your application.

2. **Validity.** Suppose that a psychologist develops a new test for measuring intelligence. How could you show that this test truly measures what it claims; that is, how could you demonstrate the validity of the test? One common technique for demonstrating validity is to use a correlation. If the test actually measures intelligence, then the scores on the test should be related to other measures of intelligence—for example, standardized IQ tests, performance on learning tasks, problem-solving ability, and so on. The psychologist could measure the correlation between the new test and each of these other measures of intelligence to demonstrate that the new test is valid.
3. **Reliability.** In addition to evaluating the validity of a measurement procedure, correlations are used to determine reliability. A measurement procedure is considered reliable to the extent that it produces stable, consistent measurements. That is, a reliable measurement procedure produces the same (or nearly the same) scores when the same individuals are measured twice under the same conditions. For example, if your IQ was measured as 113 last week, you would expect to obtain nearly the same score if your IQ was measured again this week. One way to evaluate reliability is to use correlations to determine the relationship between two sets of measurements. When reliability is high, the correlation between two measurements should be strong and positive. Further discussion of the concept of reliability is presented in Box 15.2.
4. **Theory Verification.** Many psychological theories make specific predictions about the relationship between two variables. For example, a theory may predict a relationship between brain size and learning ability; a developmental theory may predict a relationship between the parents' IQs and the child's IQ; a social psychologist may have a theory predicting a relationship between personality type and behavior in a social situation. In each case, the prediction of the theory could be tested by determining the correlation between the two variables.

INTERPRETING CORRELATIONS

When you encounter correlations, there are four additional considerations that you should bear in mind:

1. Correlation simply describes a relationship between two variables. It does not explain why the two variables are related. Specifically, a correlation should not and cannot be interpreted as proof of a cause-and-effect relationship between the two variables.
2. The value of a correlation can be affected greatly by the range of scores represented in the data.
3. One or two extreme data points, often called *outliers,* can have a dramatic effect on the value of a correlation.
4. When judging how "good" a relationship is, it is tempting to focus on the numerical value of the correlation. For example, a correlation of +0.5 is halfway between 0 and 1.00 and, therefore, appears to represent a moderate degree of relationship. However, a correlation should not be interpreted as a proportion. Although a correlation of 1.00 does mean that there is a 100% perfectly predictable relationship between X and Y, a correlation of 0.5 does not mean that you can make predictions with 50% accuracy. To describe how accurately one variable predicts the other, you must square the correlation. Thus, a

BOX 15.2 RELIABILITY AND ERROR IN MEASUREMENT

The idea of reliability of measurement is tied directly to the notion that each individual measurement includes an element of error. Expressed as an equation,

$$\text{measured score} = \text{true score} + \text{error}$$

For example, if I try to measure your intelligence with an IQ test, the score that I get is determined partially by your actual level of intelligence (your true score) but it also is influenced by a variety of other factors such as your current mood, your level of fatigue, your general health, and so on. These other factors are lumped together as *error,* and are typically a part of any measurement.

It is generally assumed that the error component changes randomly from one measurement to the next and that this causes your score to change. For example, your IQ score is likely to be higher when you are well rested and feeling good compared to a measurement that is taken when you are tired and depressed. Although your actual intelligence hasn't changed, the error component causes your score to change from one measurement to another.

As long as the error component is relatively small, then your scores will be relatively consistent from one measurement to the next, and the measurements are said to be reliable. If you are feeling especially happy and well rested, it may affect your IQ score by a few points, but it is not going to boost your IQ from 110 to 170.

On the other hand, if the error component is relatively large, then you will find huge differences from one measurement to the next and the measurements are not reliable. Measurements of reaction time, for example, tend to be very unreliable. Suppose, for example, that you are seated at a desk with your finger on a button and a light bulb in front of you. Your job is to push the button as quickly as possible when the light goes on. On some trials, you are focused on the light with your finger tensed and ready to push. On other trials, you are distracted, or day dreaming, or blink when the light goes on so that time passes before you finally push the button. As a result, there is a huge error component to the measurement and your reaction time can change dramatically from one trial to the next. When measurements are unreliable you cannot trust any single measurement to provide an accurate indication of the individual's true score. To deal with this problem, researchers typically measure reaction time repeatedly and then average it over a large number of measurements.

Correlations can be used to help researchers measure and describe reliability. By taking two measurements for each individual, it is possible to compute the correlation between the first score and the second score. A strong, positive correlation indicates a good level of reliability: people who scored high on the first measurement also scored high on the second. A weak correlation indicates that there is not a consistent relationship between the first score and the second score; that is, a weak correlation indicates poor reliability.

correlation of $r = .5$ means that one variable *partially* predicts the other, but the predictable portion is only $r^2 = 0.5^2 = 0.25$ (or 25%) of the total variability.

We now discuss each of these four points in detail.

CORRELATION AND CAUSATION

One of the most common errors in interpreting correlations is to assume that a correlation necessarily implies a cause-and-effect relationship between the two variables. (Even Pearson blundered by asserting causation from correlational data [Blum, 1978].) We constantly are bombarded with reports of relationships: Cigarette smoking is related to heart disease; alcohol consumption is related to birth defects; carrot consumption is related to good eyesight. Do these relationships mean that cigarettes cause heart disease or carrots cause good eyesight? The answer is *no*. Although there may be a causal relationship, the simple existence of a correlation does not prove it. Earlier, for example, we discussed a study showing a relationship between high school grades

and family income. However, this result does not mean that having a higher family income *causes* students to get better grades. For example, if mom gets an unexpected bonus at work, it is unlikely that her child's grades will also show a sudden increase. To establish a cause-and-effect relationship, it is necessary to conduct a true experiment (see p. 14) in which one variable is manipulated by a researcher and other variables are rigorously controlled. The fact that a correlation does not establish causation is demonstrated in the following example.

EXAMPLE 15.4

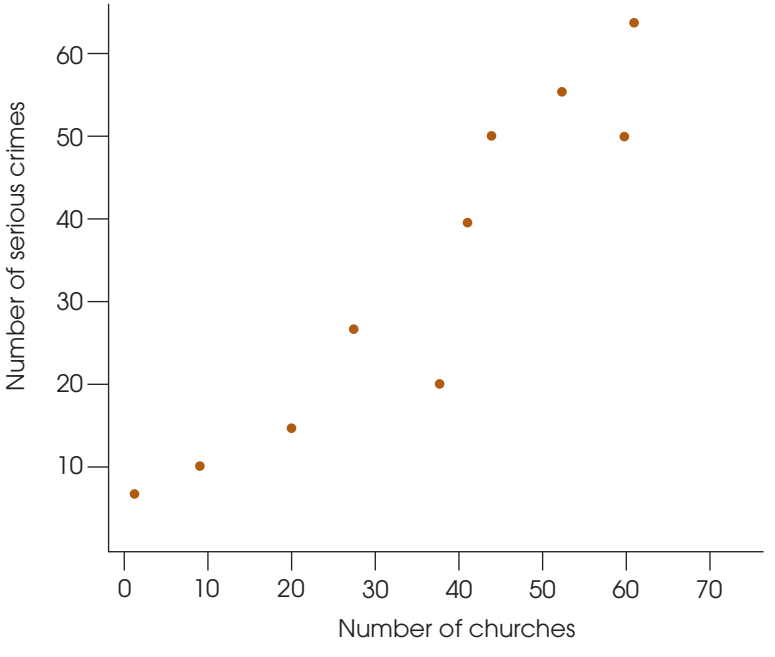
Suppose we select a variety of different cities and towns throughout the United States and measure the number of churches (*X* variable) and the number of serious crimes (*Y* variable) for each. A scatter plot showing hypothetical data for this study is presented in Figure 15.6. Notice that this scatter plot shows a strong, positive correlation between churches and crime. You also should note that these are realistic data. It is reasonable that small towns would have less crime and fewer churches and that large cities would have large values for both variables. Does this relationship mean that churches cause crime? Does it mean that crime causes churches? It should be clear that both answers are no. Although a strong correlation exists between number of churches and crime, the real cause of the relationship is the size of the population.

CORRELATION AND RESTRICTED RANGE

Whenever a correlation is computed from scores that do not represent the full range of possible values, you should be cautious in interpreting the correlation. Suppose, for example, that you are interested in the relationship between IQ and creativity. If you select a sample of your fellow college students, your data probably will represent only a limited range of IQ scores (most likely from 110 to 130). The correlation

FIGURE 15.6

Hypothetical data showing the logical relationship between the number of churches and the number of serious crimes for a sample of U.S. cities.



within this *restricted range* could be completely different from the correlation that would be obtained from a full range of IQ scores. For example, Figure 15.7 shows a strong positive relationship between X and Y when the entire range of scores is considered. However, this relationship is obscured when the data are limited to a restricted range.

To be safe, you should not generalize any correlation beyond the range of data represented in the sample. For a correlation to provide an accurate description for the general population, there should be a wide range of X and Y values in the data.

OUTLIERS

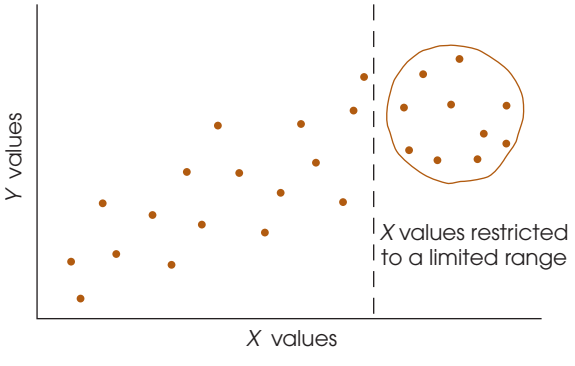
An outlier is an individual with X and/or Y values that are substantially different (larger or smaller) from the values obtained for the other individuals in the data set. The data point of a single outlier can have a dramatic influence on the value obtained for the correlation. This effect is illustrated in Figure 15.8. Figure 15.8(a) shows a set of $n = 5$ data points for which the correlation between the X and Y variables is nearly zero (actually $r = -0.08$). In Figure 15.8(b), one extreme data point (14, 12) has been added to the original data set. When this outlier is included in the analysis, a strong, positive correlation emerges (now $r = +0.85$). Note that the single outlier drastically alters the value for the correlation and, thereby, can affect one's interpretation of the relationship between variables X and Y. Without the outlier, one would conclude there is no relationship between the two variables. With the extreme data point, $r = +0.85$, which implies a strong relationship with Y increasing consistently as X increases. The problem of outliers is a good reason for looking at a scatter plot, instead of simply basing your interpretation on the numerical value of the correlation. If you only "go by the numbers," you might overlook the fact that one extreme data point inflated the size of the correlation.

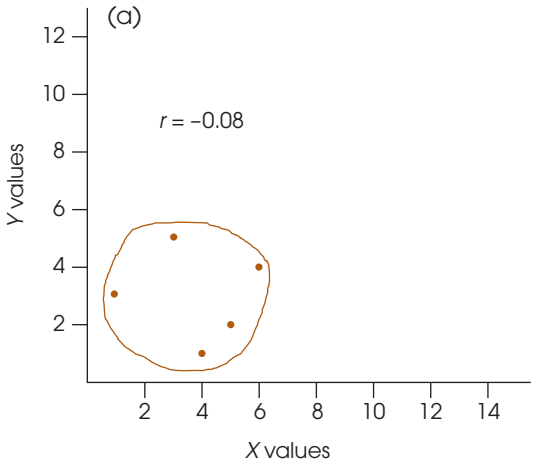
CORRELATION AND THE STRENGTH OF THE RELATIONSHIP

A correlation measures the degree of relationship between two variables on a scale from 0 to 1.00. Although this number provides a measure of the degree of relationship, many researchers prefer to square the correlation and use the resulting value to measure the strength of the relationship.

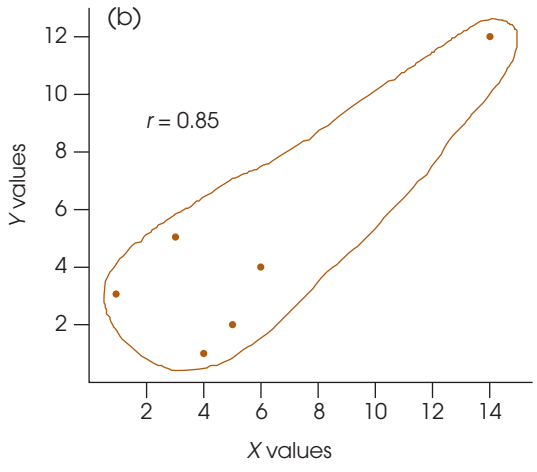
One of the common uses of correlation is for prediction. If two variables are correlated, you can use the value of one variable to predict the other. For example, college admissions officers do not just guess which applicants are likely to do well; they use other variables (SAT scores, high school grades, and so on) to predict which students are most likely to be successful. These predictions are based on correlations. By using

FIGURE 15.7
In this example, the full range of X and Y values shows a strong, positive correlation, but the restricted range of scores produces a correlation near zero.





Original Data		
Subject	*X*	*Y*
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2



Data with Outlier Included		
Subject	*X*	*Y*
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2
F	14	12

FIGURE 15.8

A demonstration of how one extreme data point (an outlier) can influence the value of a correlation.

correlations, the admissions officers expect to make more accurate predictions than would be obtained by chance. In general, the squared correlation (r^2) measures the gain in accuracy that is obtained from using the correlation for prediction. The squared correlation measures the proportion of variability in the data that is explained by the relationship between *X* and *Y*. It is sometimes called the *coefficient of determination.*

DEFINITION

The value r^2 is called the **coefficient of determination** because it measures the proportion of variability in one variable that can be determined from the relationship with the other variable. A correlation of $r = 0.80$ (or –0.80), for example, means that $r^2 = 0.64$ (or 64%) of the variability in the *Y* scores can be predicted from the relationship with *X*.

In earlier chapters (see pp. 299, 328, and 361) we introduced r^2 as a method for measuring effect size for research studies where mean differences were used to compare treatments. Specifically, we measured how much of the variance in the scores was accounted for by the differences between treatments. In experimental terminology, r^2 measures how much of the variance in the dependent variable is accounted for by the independent variable. Now we are doing the same thing, except that there is no independent or dependent variable. Instead, we simply have two variables, *X* and *Y,* and we use r^2 to measure how much of the variance in one variable can be determined from its relationship with the other variable. The following example demonstrates this concept.

EXAMPLE 15.5

Figure 15.9 shows three sets of data representing different degrees of linear relationship. The first set of data [Figure 15.9(a)] shows the relationship between IQ and shoe size. In this case, the correlation is $r = 0$ (and $r^2 = 0$), and you have no ability to predict a person's IQ based on his or her shoe size. Knowing a person's shoe size provides no information (0%) about the person's IQ. In this case, shoe size provides no help explaining why different people have different IQs.

Now consider the data in Figure 15.9(b). These data show a moderate, positive correlation, $r = +0.60$, between IQ scores and college grade point averages (GPA). Students with high IQs tend to have higher grades than students with low IQs. From this relationship, it is possible to predict a student's GPA based on his or her IQ. However, you should realize that the prediction is not perfect. Although students with high IQs *tend* to have high GPAs, this is not always true. Thus, knowing a student's IQ provides some information about the student's grades, or knowing a student's grades provides some information about the student's IQ. In this case, IQ scores help explain the fact that different students have different GPAs. Specifically, you can say that *part* of the differences in GPA are accounted for by IQ. With a correlation of $r = +0.60$, we obtain $r^2 = 0.36$, which means that 36% of the variance in GPA can be explained by IQ.

Finally, consider the data in Figure 15.9(c). This time we show a perfect linear relationship ($r = +1.00$) between monthly salary and yearly salary for a group of college employees. With $r = 1.00$ and $r^2 = 1.00$, there is 100% predictability. If you know a person's monthly salary, you can predict perfectly the person's annual salary. If two people have different annual salaries, the difference can be completely explained (100%) by the difference in their monthly salaries.

Just as r^2 was used to evaluate effect size for mean differences in Chapters 9, 10, and 11, r^2 can now be used to evaluate the size or strength of the correlation. The same standards that were introduced in Table 9.3 (p. 299), apply to both uses of the r^2 measure. Specifically, an r^2 value of 0.01 indicates a small effect or a small correlation, an r^2 value of 0.09 indicates a medium correlation, and r^2 of 0.25 or larger indicates a large correlation.

More information about the coefficient of determination (r^2) is presented in Section 15.5 and in Chapter 16. For now, you should realize that whenever two variables are consistently related, it is possible to use one variable to predict values for the

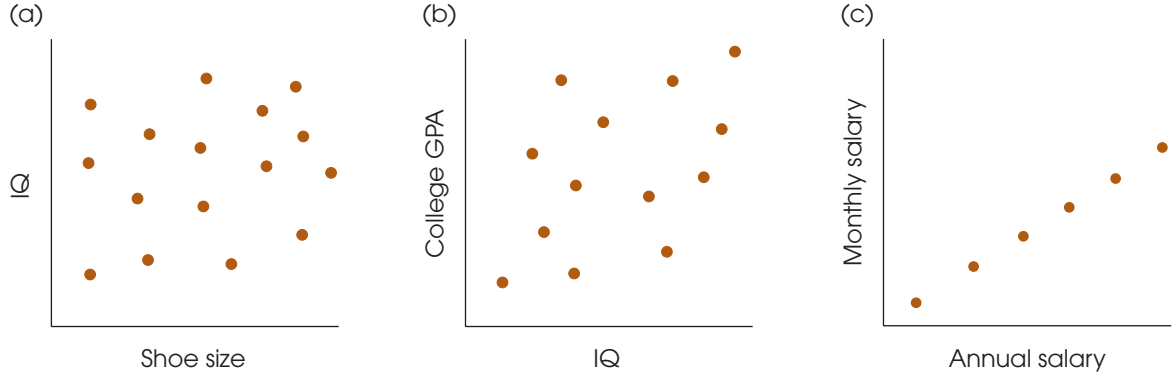


FIGURE 15.9

Three sets of data showing three different degrees of linear relationship.

second variable. One final comment concerning the interpretation of correlations is presented in Box 15.3.

BOX 15.3 REGRESSION TOWARD THE MEAN

Consider the following problem.

> Explain why the rookie of the year in major-league baseball usually does not perform as well in his second season.

Notice that this question does not appear to be statistical or mathematical in nature. However, the answer to the question is directly related to the statistical concepts of correlation and regression (Chapter 16). Specifically, there is a simple observation about correlations known as *regression toward the mean*.

DEFINITION When there is a less-than-perfect correlation between two variables, extreme scores (high or low) for one variable tend to be paired with the less extreme scores (more toward the mean) on the second variable. This fact is called **regression toward the mean.**

Figure 15.10 shows a scatter plot with a less-than-perfect correlation between two variables. The data points in this figure might represent batting averages for baseball rookies in 2010 (variable 1) and batting averages for the same players in 2011 (variable 2). Because the correlation is less than perfect, the highest scores on variable 1 are generally *not* the highest scores on variable 2. In baseball terms, the rookies who had the highest averages in 2010 do not have the highest averages in 2011.

Remember that a correlation does not explain *why* one variable is related to the other; it simply says that there is a relationship. The correlation cannot explain why the best rookie does not perform as well in his second year. But, because the correlation is not perfect, it is a statistical fact that extremely high scores in one year generally will *not* be paired with extremely high scores in the next year.

Regression toward the mean often poses a problem for interpreting experimental results. Suppose, for example, that you want to evaluate the effects of a special preschool program for disadvantaged children. You select a sample of children who score extremely low on an academic performance test. After participating in your preschool program, these children score significantly higher on the test. Why did their scores improve? One answer is that the special program helped. But an alternative answer is regression toward the mean. If there is a less-than-perfect correlation between scores on the first test and scores on the second test (which is usually the case), individuals with extremely low scores on test 1 will tend to have higher scores on test 2. It is a statistical fact of life, not necessarily the result of any special program.

Now try using the concept of regression toward the mean to explain the following phenomena:

1. You have a truly outstanding meal at a restaurant. However, when you go back with a group of friends, you find that the food is disappointing.
2. You have the highest score on exam I in your statistics class, but score only a little above average on exam II.

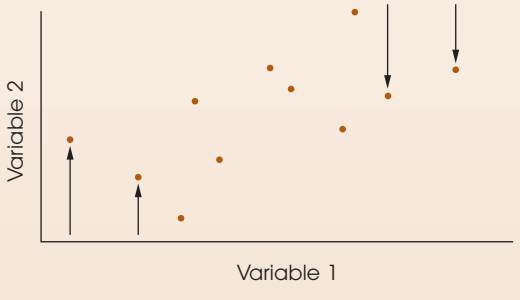


FIGURE 15.10

A demonstration of regression toward the mean. The figure shows a scatterplot for a set of data with a less-than-perfect correlation. Notice that the highest scores on variable 1 (extreme right-hand points) are not the highest scores on variable 2, but are displaced downward toward the mean. Also, the lowest scores on variable 1 (extreme left-hand points) are not the lowest scores on variable 2, but are displaced upward toward the mean.

LEARNING CHECK

1. A researcher finds a correlation of $r = (0.71$ between the time spent playing video games each week and grade point average for a group of high school boys. This means that playing video games causes students to get lower grades. (True or false?)
2. A researcher finds a correlation of $r = 0.60$ between salary and the number of years of education for a group of 40-year-old men. How much of the variance in salary is explained by the years of education?

ANSWERS

1. False. You cannot conclude that there is a cause-and-effect relationship based on a correlation.
2. $r^2 = 0.36$, or 36%

15.4 HYPOTHESIS TESTS WITH THE PEARSON CORRELATION

The Pearson correlation is generally computed for sample data. As with most sample statistics, however, a sample correlation is often used to answer questions about the corresponding population correlation. For example, a psychologist would like to know whether there is a relationship between IQ and creativity. This is a general question concerning a population. To answer the question, a sample would be selected, and the sample data would be used to compute the correlation value. You should recognize this process as an example of inferential statistics: using samples to draw inferences about populations. In the past, we have been concerned primarily with using sample means as the basis for answering questions about population means. In this section, we examine the procedures for using a sample correlation as the basis for testing hypotheses about the corresponding population correlation.

THE HYPOTHESES

The basic question for this hypothesis test is whether a correlation exists in the population. The null hypothesis is "No. There is no correlation in the population," or "The population correlation is zero." The alternative hypothesis is "Yes. There is a real, nonzero correlation in the population." Because the population correlation is traditionally represented by ρ (the Greek letter rho), these hypotheses would be stated in symbols as

H_0: $\rho = 0$ (There is no population correlation.)

H_1: $\rho \neq 0$ (There is a real correlation.)

When there is a specific prediction about the direction of the correlation, it is possible to do a directional, or one-tailed, test. For example, if a researcher is predicting a positive relationship, the hypotheses would be

H_0: $\rho \leq 0$ (The population correlation is not positive.)

H_1: $\rho > 0$ (The population correlation is positive.)

The correlation from the sample data is used to evaluate the hypotheses. For the regular, nondirectional test, a sample correlation near zero provides support for H_0 and a sample value far from zero tends to refute H_0. For a directional test, a positive value for the sample correlation would tend to refute a null hypothesis stating that the population correlation is not positive.

Although sample correlations are used to test hypotheses about population correlations, you should keep in mind that samples are not expected to be identical to the populations from which they come; there is some discrepancy (sampling error) between a sample statistic and the corresponding population parameter. Specifically, you should always expect some error between a sample correlation and the population correlation it represents. One implication of this fact is that even when there is no correlation in the population ($\rho = 0$), you are still likely to obtain a nonzero value for the sample correlation. This is particularly true for small samples. Figure 15.11 illustrates how a small sample from a population with a near-zero correlation could result in a correlation that deviates from zero. The colored dots in the figure represent the entire population and the three circled dots represent a random sample. Note that the three sample points show a relatively good, positive correlation even through there is no linear trend ($\rho = 0$) for the population.

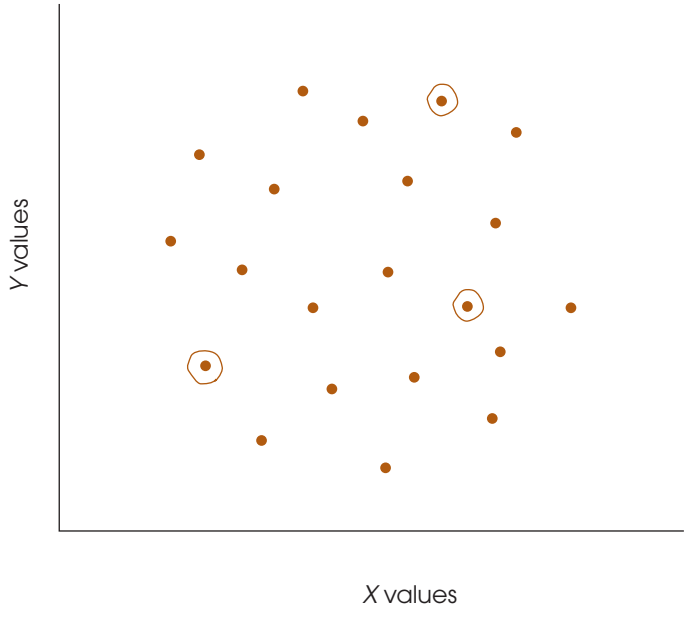
When you obtain a nonzero correlation for a sample, the purpose of the hypothesis test is to decide between the following two interpretations:

1. There is no correlation in the population ($\rho = 0$), and the sample value is the result of sampling error. Remember, a sample is not expected to be identical to the population. There always is some error between a sample statistic and the corresponding population parameter. This is the situation specified by H_0.
2. The nonzero sample correlation accurately represents a real, nonzero correlation in the population. This is the alternative stated in H_1.

The correlation from the sample helps to determine which of these two interpretations is more likely. A sample correlation near zero supports the conclusion that the population correlation is also zero. A sample correlation that is substantially different from zero supports the conclusion that there is a real, nonzero correlation in the population.

FIGURE 15.11

Scatterplot of a population of X and Y values with a near-zero correlation. However, a small sample of $n = 3$ data points from this population shows a relatively strong, positive correlation. Data points in the sample are circled.



DEGREES OF FREEDOM FOR THE CORRELATION TEST

The hypothesis test for the Pearson correlation has degrees of freedom defined by $df = n - 2$. An intuitive explanation for this value is that a sample with only $n = 2$ data points has no degrees of freedom. Specifically, if there are only two points, they will fit perfectly on a straight line, and the sample produces a perfect correlation of $r = +1.00$ or $r = -1.00$. Because the first two points always produce a perfect correlation, the sample correlation is free to vary only when the data set contains more than two points. Thus, $df = n - 2$.

THE HYPOTHESIS TEST

Although it is possible to conduct the hypothesis test by computing either a t statistic or an F-ratio, the computations for evaluating r have already been completed and are summarized in Table B.6 in Appendix B. The table is based on the concept that a sample is expected to be representative of the population from which it was obtained. In particular, a sample correlation should be similar to the population correlation. If the population correlation is zero, as specified in the null hypothesis, then the sample correlation should be near zero. Thus, a sample correlation that is close to zero provides support for H_0 and a sample correlation that is far from zero contradicts the null hypothesis.

The table lists critical values in terms of degrees of freedom: $df = n - 2$. Remember to subtract 2 when using this table.

Table B.6 identifies exactly which sample correlations are likely to be obtained from a population with $\rho = 0$ and which values are very unlikely. To use the table, you need to know the sample size (n) and the alpha level. With a sample size of $n = 20$ and an alpha level of .05, for example, you locate $df = n - 2 = 18$ in the left-hand column and the value .05 for either one tail or two tails across the top of the table. For $df = 18$ and $\alpha = .05$ for a two-tailed test, the table shows a value of 0.444. Thus, if the null hypothesis is true and there is no correlation in the population, then the sample correlation should be near to zero. According to the table, the sample correlation should have a value between +0.444 and –0.444. If H_0 is true, it is very unlikely ($\alpha = .05$) to obtain a sample correlation outside this range. Therefore, a sample correlation beyond ±0.444 leads to rejecting the null hypothesis. The following examples demonstrate the use of the table.

EXAMPLE 15.6

A researcher is using a regular, two-tailed test with $\alpha = .05$ to determine whether a nonzero correlation exists in the population. A sample of $n = 30$ individuals is obtained. With $\alpha = .05$ and $n = 30$, the table lists a value of 0.361. Thus, the sample correlation (independent of sign) must have a value greater than or equal to 0.361 to reject H_0 and conclude that there is a significant correlation in the population. Any sample correlation between 0.361 and –0.361 is considered within the realm of sampling error and, therefore, is not significant.

EXAMPLE 15.7

This time the researcher is using a directional, one-tailed test to determine whether there is a positive correlation in the population.

H_0: $\rho \leq 0$ (There is not a positive correlation.)

H_1: $\rho > 0$ (There is a positive correlation.)

With $\alpha = .05$ and a sample of $n = 30$, the table lists a value of 0.306 for a one-tailed test. To reject H_0 and conclude that there is a significant positive correlation in the population, the sample correlation must be positive (as predicted) and have a value greater than or equal to 0.306.

IN THE LITERATURE

REPORTING CORRELATIONS

When correlations are computed, the results are reported using APA format. The statement should include the sample size, the calculated value for the correlation, whether it is a statistically significant relationship, the probability level, and the type of test used (one- or two-tailed). For example, a correlation might be reported as follows:

> A correlation for the data revealed a significant relationship between amount of education and annual income, $r = +0.65$, $n = 30$, $p < .01$, two tails.

Sometimes a study might look at several variables, and correlations between all possible variable pairings are computed. Suppose, for example, that a study measured people's annual income, amount of education, age, and intelligence. With four variables, there are six possible pairings leading to six different correlations. The results from multiple correlations are most easily reported in a table called a *correlation matrix,* using footnotes to indicate which correlations are significant. For example, the report might state:

> The analysis examined the relationships among income, amount of education, age, and intelligence for $n = 30$ participants. The correlations between pairs of variables are reported in Table 1. Significant correlations are noted in the table.

TABLE 1

Correlation matrix for income, amount of education, age, and intelligence

	Education	Age	IQ
Income	+.65*	+.41**	+.27
Education		+.11	+.38**
Age			−.02

$n = 30$
*$p < .01$, two tails
**$p < .05$, two tails

LEARNING CHECK

1. A researcher obtains a correlation of $r = -0.39$ for a sample of $n = 25$ individuals. Does this sample provide sufficient evidence to conclude that there is a significant, nonzero correlation in the population? Assume a two-tailed test with $\alpha = .05$.
2. For a sample of $n = 15$, how large a correlation is needed to conclude at the .05 level of significance that there is a nonzero correlation in the population? Assume a two-tailed test.
3. As sample size gets smaller, what happens to the magnitude of the correlation necessary for significance? Explain why this occurs.

ANSWERS

1. No. For $n = 25$, the critical value is $r = 0.396$. The sample value is not in the critical region.
2. For $n = 15$, $df = 13$ and the critical value is $r = 0.514$.
3. As the sample size gets smaller, the magnitude of the correlation needed for significance gets larger. With a small sample, it is easy to get a relatively large correlation just by chance. Therefore, a small sample requires a very large correlation before you can be confident there is a real (nonzero) relationship in the population.

PARTIAL CORRELATIONS

Occasionally a researcher may suspect that the relationship between two variables is being distorted by the influence of a third variable. Earlier in the chapter, for example, we found a strong positive relationship between the number of churches and the number of serious crimes for a sample of different towns and cities (see Example 15.4, p 522). However, it is unlikely that there is a direct relationship between churches and crime. Instead, both variables are influenced by population: Large cities have a lot of churches and high crime rates compared to smaller towns, which have fewer churches and less crime. If population were controlled, there probably would be no real correlation between churches and crime.

Fortunately, there is a statistical technique, known as *partial correlation,* that allows a researcher to measure the relationship between two variables while eliminating or holding constant the influence of a third variable. Thus, a researcher could use a partial correlation to examine the relationship between churches and crime without the risk that the relationship is distorted by the size of the population.

DEFINITION

A **partial correlation** measures the relationship between two variables while controlling the influence of a third variable by holding it constant.

In a situation with three variables, X, Y, and Z, it is possible to compute three individual Pearson correlations:

1. r_{XY} measuring the correlation between X and Y
2. r_{XZ} measuring the correlation between X and Z
3. r_{YZ} measuring the correlation between Y and Z

These three individual correlations can then be used to compute a partial correlation. For example, the partial correlation between X and Y, holding Z constant, is determined by the formula

$$r_{XY-Z} = \frac{r_{XY} - \left(r_{XZ}r_{YZ}\right)}{\sqrt{\left(1-r_{XZ}^2\right)\left(1-r_{YZ}^2\right)}} \tag{15.6}$$

The following example demonstrates the calculation and interpretation of a partial correlation.

EXAMPLE 15.8

We begin with the hypothetical data shown in Table 15.2. These scores have been constructed to simulate the church/crime/population situation for a sample of $n = 15$ cities. The X variable represents the number of churches, Y represents the number of

TABLE 15.2

Hypothetical data showing the relationship between the number of churches, the number of crimes, and the population of a set of $n = 15$ cities.

Number of Churches (X)	Number of Crimes (Y)	Population (Z)
1	4	1
2	3	1
3	1	1
4	2	1
5	5	1
7	8	2
8	11	2
9	9	2
10	7	2
11	10	2
13	15	3
14	14	3
15	16	3
16	17	3
17	13	3

crimes, and Z represents the population for each city. For these scores, the individual Pearson correlations are all large and positive:

a. The correlation between churches and crime is $r_{XY} = 0.923$.

b. The correlation between churches and population is $r_{XZ} = 0.961$.

c. The correlation between crime and population is $r_{YZ} = 0.961$.

The data points for the 15 cities are shown in the scatter plot in Figure 15.12. Note that there are three categories for the size of the population (three values for Z) corresponding to small, medium, and large cities. Also note that the population variable, Z, separates the scores into three distinct groups: When $Z = 1$, the population is low and churches and crime (X and Y) are also low; when $Z = 2$, the population is moderate and churches and crime (X and Y) are also moderate; and when $Z = 3$, the population is large and churches and crime are both high. Thus, as the population increases from one city to another, the number of churches and crimes also increase, and the result is a strong positive correlation between churches and crime.

Within each of the three population categories, however, there is no linear relationship between churches and crime. Specifically, within each group, the population variable is constant and the five data points for X and Y form a circular pattern, indicating no consistent linear relationship. The partial correlation allows us to hold population constant across the entire sample and measure the underlying relationship between churches and crime without any influence from population. For these data, the partial correlation is

$$r_{XY-Z} = \frac{0.923 - 0.961(0.961)}{\sqrt{(1-0.961^2)(1-0.961^2)}}$$

$$= \frac{0}{0.076}$$

$$= 0$$

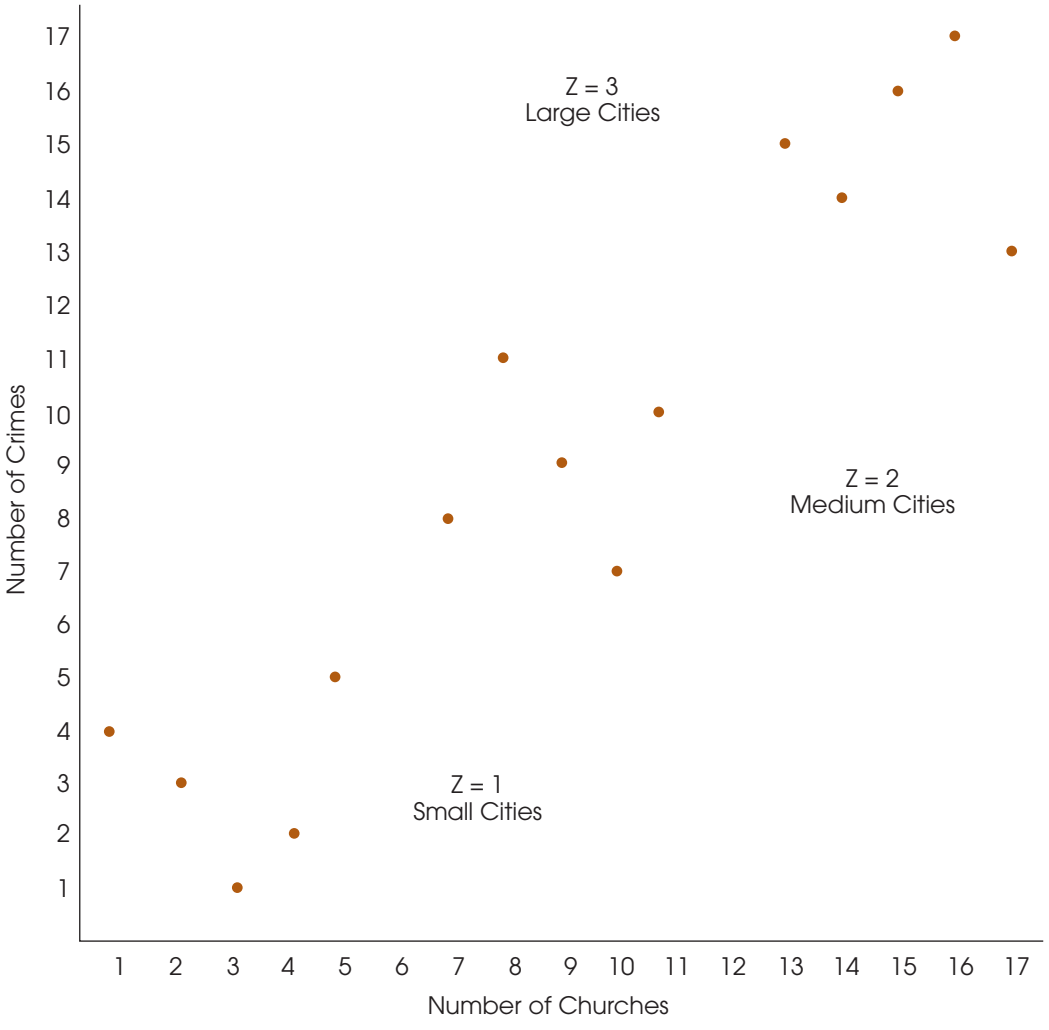


FIGURE 15.12

Hypothetical data showing the relationship between the number of churches and the number of crimes for three groups of cities. Those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

Thus, when the population differences are eliminated, there is no correlation remaining between churches and crime ($r = 0$).

In Example 15.8, the population differences, which correspond to the different values of the Z variable, were eliminated mathematically in the calculation of the partial correlation. However, it is possible to visualize how these differences are eliminated in the actual data. Looking at Figure 15.12, focus on the five points in the bottom left corner. These are the five cities with small populations, few churches, and little crime. The five points in the upper right corner represent the five cities with large populations, many churches, and a lot of crime. The partial correlation controls population size by

mathematically equalizing the populations for all 15 cities. Population is increased for the five small cities. However, increasing the population also increases churches and crime. Similarly, population is decreased for the five large cities, which also decreases churches and crime. In Figure 15.12, imagine the five points in the bottom left moving up and to the right so that they overlap with the points in the center. At the same time, the five points in the upper right move down and to the left so that they also overlap the points in the center. When population is equalized, the resulting set of 15 cities is shown in Figure 15.13. Note that controlling the population appears to have eliminated the relationship between churches and crime. This appearance is verified by the correlation for the 15 data points in Figure 15.13, which is $r = 0$, exactly the same as the partial correlation.

In Example 15.8 we used a partial correlation to demonstrate that an apparent relationship between churches and crime was actually caused by the influence of a third variable, population. It also is possible to use partial correlations to demonstrate that a relationship is not caused by the influence of a third variable. As an example, consider research examining the relationship between exposure to sexual content on television and sexual behavior among adolescents (Collins et al., 2004). The study consisted of a survey of 1,792 adolescents, 12 to 17 years old, who reported their television viewing

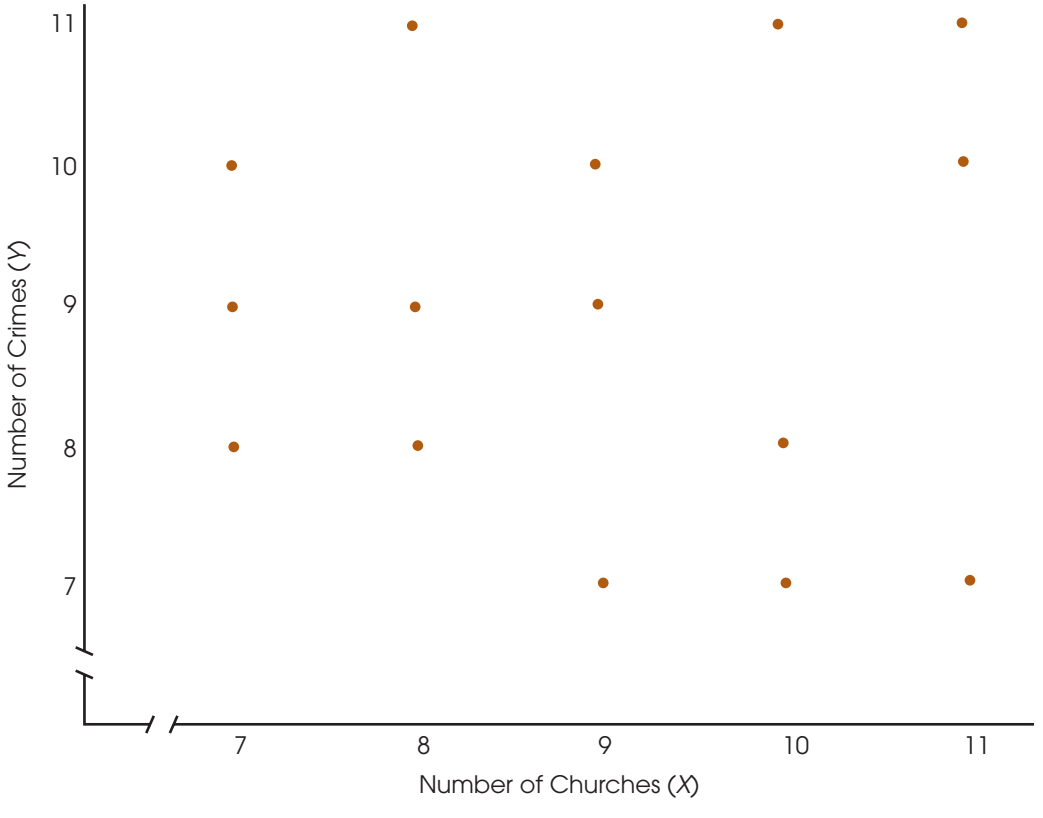


FIGURE 15.13

The relationship between the number of churches and the number of crimes for the same 15 cities shown in Figure 15.12 after the populations have been equalized.

habits and their sexual behaviors. The results showed a clear relationship between television viewing and behaviors. Specifically, the more sexual content the adolescents watched on television, the more likely they were to engage in sexual behaviors. One concern for the researchers was that the observed relationship may be influenced by the age of the participants. For example, as the adolescents mature from age 12 to age 17, they increasingly watch television programs with sexual content and they increase their own sexual behaviors. Although the viewing of sexual content on television and the participants' sexual behaviors are increasing together, the observed relationship may simply be the result of age differences. To address this problem, the researcher used a partial correlation technique to eliminate or hold constant the age variable. The results clearly showed that a relationship still exists between television sexual content and sexual behavior even after the influence of the participants' ages was accounted for.

Testing the significance of a partial correlation The statistical significance of a partial correlation is determined using the same procedure as is used to evaluate a regular Pearson correlation. Specifically, the partial correlation is compared with the critical values listed in Table B6. For a partial correlation, however, you must use $df = n - 3$ instead of the $n - 2$ value that is used for the Pearson correlation. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population.

LEARNING CHECK

1. Sales figures show a positive relationship between temperature and ice cream consumption; as temperature increases, ice cream consumption also increases. Other research shows a positive relationship between temperature and crime rate (Cohn & Rotton, 2000). When the temperature increases, both ice cream consumption and crime rates tend to increase. As a result, there is a positive correlation between ice cream consumption and crime rate. However, what do you think is the true relationship between ice cream consumption and crime rate? Specifically, what value would you predict for the partial correlation between the two variables if temperature were held constant?

ANSWER

1. There should be no systematic relationship between ice cream consumption and crime rate. The partial correlation should be near zero.

15.5 ALTERNATIVES TO THE PEARSON CORRELATION

The Pearson correlation measures the degree of linear relationship between two variables when the data (X and Y values) consist of numerical scores from an interval or ratio scale of measurement. However, other correlations have been developed for nonlinear relationships and for other types of data. In this section we examine three additional correlations: the Spearman correlation, the point-biserial correlation, and the phi-coefficient. As you will see, all three can be viewed as special applications of the Pearson correlation.

THE SPEARMAN CORRELATION

When the Pearson correlation formula is used with data from an ordinal scale (ranks), the result is called the *Spearman correlation*. The Spearman correlation is used in two situations.

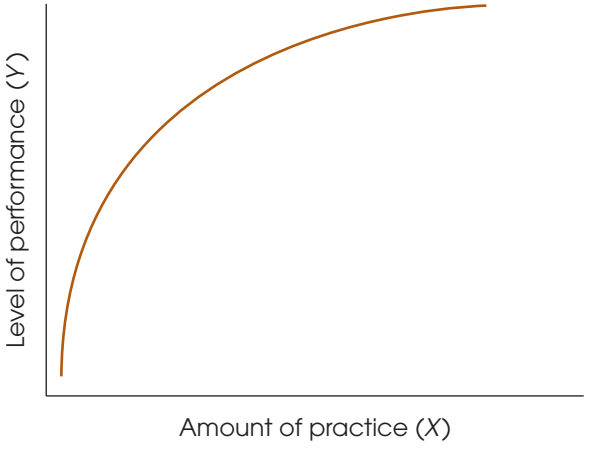
First, the Spearman correlation is used to measure the relationship between *X* and *Y* when both variables are measured on ordinal scales. Recall from Chapter 1 that an ordinal scale typically involves ranking individuals rather than obtaining numerical scores. Rank-order data are fairly common because they are often easier to obtain than interval or ratio scale data. For example, a teacher may feel confident about rank-ordering students' leadership abilities but would find it difficult to measure leadership on some other scale.

In addition to measuring relationships for ordinal data, the Spearman correlation can be used as a valuable alternative to the Pearson correlation, even when the original raw scores are on an interval or a ratio scale. As we have noted, the Pearson correlation measures the degree of *linear relationship* between two variables—that is, how well the data points fit on a straight line. However, a researcher often expects the data to show a consistently one-directional relationship but not necessarily a linear relationship. For example, Figure 15.14 shows the typical relationship between practice and performance. For nearly any skill, increasing amounts of practice tend to be associated with improvements in performance (the more you practice, the better you get). However, it is not a straight-line relationship. When you are first learning a new skill, practice produces large improvements in performance. After you have been performing a skill for several years, however, additional practice produces only minor changes in performance. Although there is a consistent relationship between the amount of practice and the quality of performance, it clearly is not linear. If the Pearson correlation were computed for these data, it would not produce a correlation of 1.00 because the data do not fit perfectly on a straight line. In a situation like this, the Spearman correlation can be used to measure the consistency of the relationship, independent of its form.

The reason that the Spearman correlation measures consistency, rather than form, comes from a simple observation: When two variables are consistently related, their ranks are linearly related. For example, a perfectly consistent positive relationship means that every time the *X* variable increases, the *Y* variable also increases. Thus, the smallest value of *X* is paired with the smallest value of *Y,* the second-smallest value of *X* is paired with the second smallest value of *Y,* and so on. Every time the rank for *X* goes up by 1 point, the rank for *Y* also goes up by 1 point. As a result, the ranks fit perfectly on a straight line. This phenomenon is demonstrated in the following example.

FIGURE 15.14

Hypothetical data showing the relationship between practice and performance. Although this relationship is not linear, there is a consistent positive relationship. An increase in performance tends to accompany an increase in practice.



EXAMPLE 15.9

Table 15.3 presents X and Y scores for a sample of $n = 4$ people. Note that the data show a perfectly consistent relationship. Each increase in X is accompanied by an increase in Y. However the relationship is not linear, as can be seen in the graph of the data in Figure 15.15(a).

Next, we convert the scores to ranks. The lowest X is assigned a rank of 1, the next lowest a rank of 2, and so on. The Y scores are then ranked in the same way. The ranks are listed in Table 15.3 and shown in Figure 15.15(b). Note that the perfect consistency for the scores produces a perfect linear relationship for the ranks.

The preceding example demonstrates that a consistent relationship among scores produces a linear relationship when the scores are converted to ranks. Thus, if you want to measure the consistency of a relationship for a set of scores, you can simply convert the scores to ranks and then use the Pearson correlation formula to measure the linear relationship for the ranked data. The degree of linear relationship for the ranks provides a measure of the degree of consistency for the original scores.

TABLE 15.3

Scores and ranks for Example 15.9.

Person	X	Y	X–Rank	Y–Rank
A	4	9	3	3
B	2	2	1	1
C	10	10	4	4
D	3	8	2	2

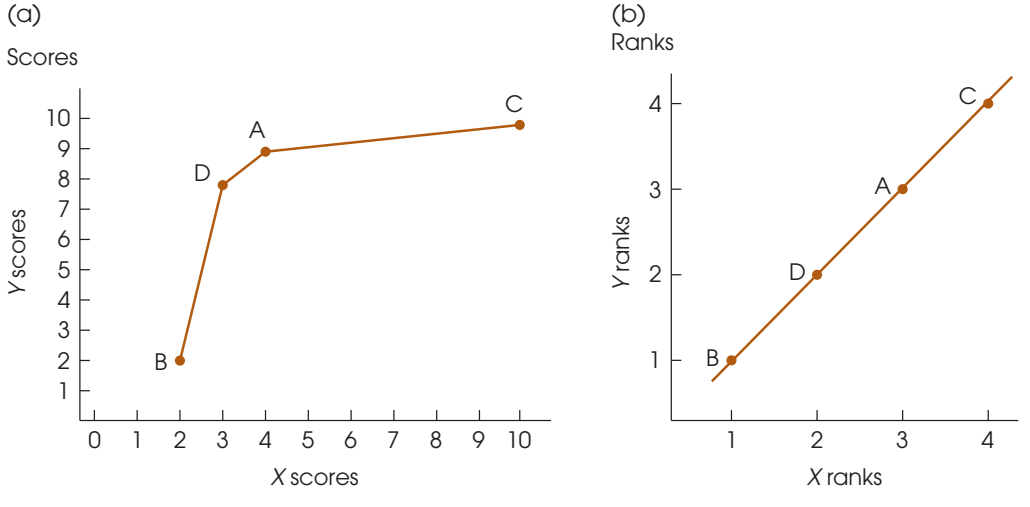


FIGURE 15.15

Scatter plots showing (a) the scores and (b) the ranks for the data in Example 15.9. Notice that there is a consistent, positive relationship between the X and Y scores, although it is not a linear relationship. Also notice that the scatter plot of the ranks shows a perfect linear relationship.

To summarize, the Spearman correlation measures the relationship between two variables when both are measured on ordinal scales (ranks). There are two general situations in which the Spearman correlation is used:

1. Spearman is used when the original data are ordinal; that is, when the X and Y values are ranks. In this case, you simply apply the Pearson correlation formula to the set of ranks.
2. Spearman is used when a researcher wants to measure the consistency of a relationship between X and Y, independent of the specific form of the relationship. In this case, the original scores are first converted to ranks, then the Pearson correlation formula is used with the ranks. Because the Pearson formula measures the degree to which the ranks fit on a straight line, it also measures the degree of consistency in the relationship for the original scores. Incidentally, when there is a consistently one-directional relationship between two variables, the relationship is said to be *monotonic*. Thus, the Spearman correlation measures the degree of monotonic relationship between two variables.

The word *monotonic* describes a sequence that is consistently increasing (or decreasing). Like the word *monotonous,* it means constant and unchanging.

In either case, the Spearman correlation is identified by the symbol r_S to differentiate it from the Pearson correlation. The complete process of computing the Spearman correlation, including ranking scores, is demonstrated in Example 15.10.

EXAMPLE 15.10

The following data show a nearly perfect monotonic relationship between X and Y. When X increases, Y tends to decrease, and there is only one reversal in this general trend. To compute the Spearman correlation, we first rank the X and Y values, and we then compute the Pearson correlation for the ranks.

We have listed the X values in order so that the trend is easier to recognize.

Original Data

X	Y
3	12
4	10
10	11
11	9
12	2

Ranks

X	Y	XY
1	5	5
2	3	6
3	4	12
4	2	8
5	1	5
		$36 = \Sigma XY$

To compute the correlation, we need SS for X, SS for Y, and SP. Remember that all of these values are computed with the ranks, not the original scores. The X ranks are simply the integers 1, 2, 3, 4, and 5. These values have $\Sigma X = 15$ and $\Sigma X^2 = 55$. The SS for the X ranks is

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 55 - \frac{(15)^2}{5} = 10$$

Note that the ranks for Y are identical to the ranks for X; that is, they are the integers 1, 2, 3, 4, and 5. Therefore, the SS for Y is identical to the SS for X:

$$SS_Y = 10$$

To compute the *SP* value, we need ΣX, ΣY, and ΣXY for the ranks. The *XY* values are listed in the table with the ranks, and we already have found that both the *X*s and the *Y*s have a sum of 15. Using these values, we obtain

$$SP = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 36 - \frac{(15)(15)}{5} = -9$$

Finally, the Spearman correlation simply uses the Pearson formula for the ranks.

$$r_s = \frac{SP}{\sqrt{(SS_X)(SS_Y)}} = \frac{-9}{\sqrt{10(10)}} = -0.9$$

The Spearman correlation indicates that the data show a consistent (nearly perfect) negative trend.

RANKING TIED SCORES

When you are converting scores into ranks for the Spearman correlation, you may encounter two (or more) identical scores. Whenever two scores have exactly the same value, their ranks should also be the same. This is accomplished by the following procedure:

1. List the scores in order from smallest to largest. Include tied values in the list.
2. Assign a rank (first, second, etc.) to each position in the ordered list.
3. When two (or more) scores are tied, compute the mean of their ranked positions, and assign this mean value as the final rank for each score.

The process of finding ranks for tied scores is demonstrated here. These scores have been listed in order from smallest to largest.

Scores	Rank Position	Final Rank	
3	1	1.5	Mean of 1 and 2
3	2	1.5	
5	3	3	
6	4	5	Mean of 4, 5, and 6
6	5	5	
6	6	5	
12	7	7	

Note that this example has seven scores and uses all seven ranks. For $X = 12$, the largest score, the appropriate rank is 7. It cannot be given a rank of 6 because that rank has been used for the tied scores.

SPECIAL FORMULA FOR THE SPEARMAN CORRELATION

After the original *X* values and *Y* values have been ranked, the calculations necessary for *SS* and *SP* can be greatly simplified. First, you should note that the *X* ranks and the *Y* ranks are really just a set of integers: 1, 2, 3, 4, ... , *n*. To compute the mean for these

integers, you can locate the midpoint of the series by $M = (n + 1)/2$. Similarly, the *SS* for this series of integers can be computed by

$$SS = \frac{n\left(n^2 - 1\right)}{12} \quad \left(\text{Try it out.}\right)$$

Also, because the *X* ranks and the *Y* ranks are the same values, the *SS* for *X* is identical to the *SS* for *Y*.

Because calculations with ranks can be simplified and because the Spearman correlation uses ranked data, these simplifications can be incorporated into the final calculations for the Spearman correlation. Instead of using the Pearson formula after ranking the data, you can put the ranks directly into a simplified formula:

$$r_S = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} \tag{15.7}$$

Caution: In this formula, you compute the value of the fraction and then subtract from 1. The 1 is not part of the fraction.

where *D* is the difference between the *X* rank and the *Y* rank for each individual. This special formula produces the same result that would be obtained from the Pearson formula. However, note that this special formula can be used only after the scores have been converted to ranks and only when there are no ties among the ranks. If there are relatively few tied ranks, the formula still may be used, but it loses accuracy as the number of ties increases. The application of this formula is demonstrated in the following example.

EXAMPLE 15.11

To demonstrate the special formula for the Spearman correlation, we use the same data that were presented in Example 15.10. The ranks for these data are shown again here:

Ranks		Difference	
X	*Y*	*D*	D^2
1	5	4	16
2	3	1	1
3	4	1	1
4	2	−2	4
5	1	−4	16
			$38 = \Sigma D^2$

Using the special formula for the Spearman correlation, we obtain

$$r_S = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} = 1 - \frac{6(38)}{5(25 - 1)} = 1 - \frac{228}{120} = 1 - 1.90 = -0.90$$

This is exactly the same answer that we obtained in Example 15.10, using the Pearson formula on the ranks.

TESTING THE SIGNIFICANCE OF THE SPEARMAN CORRELATION

Testing a hypothesis for the Spearman correlation is similar to the procedure used for the Pearson r. The basic question is whether a correlation exists in the population. The sample correlation could be the result of chance, or perhaps it reflects an actual relationship between the variables in the population. For the Pearson correlation, the Greek letter rho (ρ) was used for the population correlation. For the Spearman, ρ_S is used for the population parameter. Note that this symbol is consistent with the sample statistic, r_S. The null hypothesis states that there is no correlation (no monotonic relationship) between the variables for the population, or, in symbols:

$$H_0: \quad \rho_S = 0 \qquad \text{(The population correlation is zero.)}$$

The alternative hypothesis predicts that a nonzero correlation exists in the population, which can be stated in symbols as

$$H_1: \quad \rho_S \neq 0 \qquad \text{(There is a real correlation.)}$$

To determine whether the Spearman correlation is statistically significant (that is, H_0 should be rejected), consult Table B.7. This table is similar to the one used to determine the significance of Pearson's r (Table B.6); however, the first column is sample size (n) rather than degrees of freedom. To use the table, line up the sample size in the first column with one of the alpha levels listed across the top. The values in the body of the table identify the magnitude of the Spearman correlation that is necessary to be significant. The table is built on the concept that a sample correlation should be representative of the corresponding population value. In particular, if the population correlation is $\rho_S = 0$ (as specified in H_0), then the sample correlation should be near zero. For each sample size and alpha level, the table identifies the minimum sample correlation that is significantly different from zero. The following example demonstrates the use of the table.

EXAMPLE 15.12

An industrial psychologist selects a sample of $n = 15$ employees. These employees are ranked in order of work productivity by their manager. They also are ranked by a peer. The Spearman correlation computed for these data revealed a correlation of $r_S = .60$. Using Table B.7 with $n = 15$ and $\alpha = .05$, a correlation of at least ± 0.45 is needed to reject H_0. The observed correlation for the sample easily surpasses this critical value. The correlation between manager and peer ratings is statistically significant.

LEARNING CHECK

1. Describe what is measured by a Spearman correlation, and explain how this correlation is different from the Pearson correlation.
2. If the following scores are converted into ranks, what rank will be assigned to the individuals who have scores of $X = 7$?

   Scores: 1, 1, 1, 3, 6, 7, 7, 8, 10
3. Rank the following scores and compute the Spearman correlation:

X	Y
2	7
12	38
9	6
10	19

ANSWERS

1. The Spearman correlation measures the consistency of the direction of the relationship between two variables. The Spearman correlation does not depend on the form of the relationship, whereas the Pearson correlation measures how well the data fit a linear form.
2. Both scores get a rank of 6.5 (the average of 6 and 7).
3. $r_S = 0.80$

THE POINT-BISERIAL CORRELATION AND MEASURING EFFECT SIZE WITH r^2

In Chapters 9, 10, and 11 we introduced r^2 as a measure of effect size that often accompanies a hypothesis test using the t statistic. The r^2 used to measure effect size and the r used to measure a correlation are directly related, and we now have an opportunity to demonstrate the relationship. Specifically, we compare the independent-measures t test (Chapter 10) and a special version of the Pearson correlation known as the *point-biserial correlation.*

The point-biserial correlation is used to measure the relationship between two variables in situations in which one variable consists of regular, numerical scores, but the second variable has only two values. A variable with only two values is called a *dichotomous variable* or a *binomial variable*. Some examples of dichotomous variables are

1. Male versus female
2. College graduate versus not a college graduate
3. First-born child versus later-born child
4. Success versus failure on a particular task
5. Older than 30 years versus younger than 30 years

It is customary to use the numerical values 0 and 1, but any two different numbers would work equally well and would not affect the value of the correlation.

To compute the point-biserial correlation, the dichotomous variable is first converted to numerical values by assigning a value of zero (0) to one category and a value of one (1) to the other category. Then the regular Pearson correlation formula is used with the converted data.

To demonstrate the point-biserial correlation and its association with the r^2 measure of effect size, we use the data from Example 10.1 (p. 326). The original example compared high school grades for two groups of students: one group who regularly watched Sesame Street as 5-year-old children and one who did not watch the program. The data from the independent-measures study are presented on the left side of Table 15.4. Notice that the data consist of two separate samples and the independent-measures t was used to determine whether there was a significant mean difference between the two populations represented by the samples.

On the right-hand side of Table 15.4, we have reorganized the data into a form that is suitable for a point-biserial correlation. Specifically, we used each student's high school grade as the X value and we have created a new variable, Y, to represent the group, or condition, for each student. In this case, we have used $Y = 1$ for students who watched Sesame Street and $Y = 0$ for students who did not watch the program.

When the data in Table 15.4 were originally presented in Chapter 10, we conducted an independent-measures t hypothesis test and obtained $t = 4.00$ with $df = 18$. We measured the size of the treatment effect by calculating r^2, the percentage of variance accounted for, and obtained $r^2 = 0.47$.

Calculating the point-biserial correlation for these data also produces a value for r. Specifically, the X scores produce $SS = 680$; the Y values produce $SS = 5.00$, and the

TABLE 15.4

The same data are organized in two different formats. On the left-hand side, the data appear as two separate samples appropriate for an independent-measures *t* hypothesis test. On the right-hand side, the same data are shown as a single sample, with two scores for each individual: the original high school grade and a dichotomous score (*Y*) that identifies the group in which the participant is located (Seasame Street = 1 and No-Sesame Street = 0). The data on the right are appropriate for a point-biserial correlation.

Data for the Independent-Measures *t* test. Two separate samples, each with $n = 10$ scores.

Average High School Grade			
Watched Seasame Street		Did Not Watch Sesame Street	
86	99	90	79
87	97	89	83
91	94	82	86
97	89	83	81
98	92	85	92
$n = 10$		$n = 10$	
$M = 93$		$M = 85$	
$SS = 200$		$SS = 160$	

Data for the Point-Biserial Correlation. Two scores, *X* and *Y* for each of the $n = 20$ participants.

Participant	Grade *X*	Group *Y*
A	86	1
B	87	1
C	91	1
D	97	1
E	98	1
F	99	1
G	97	1
H	94	1
I	89	1
J	92	1
K	90	0
L	89	0
M	82	0
N	83	0
O	85	0
P	79	0
Q	83	0
R	86	0
S	81	0
T	92	0

sum of the products of the *X* and *Y* deviations produces $SP = 40$. The point-biserial correlation is

$$r = \frac{SP}{\sqrt{(SS_X)(SS_Y)}} = \frac{40}{\sqrt{(680)(5)}} = \frac{40}{58.31} = 0.686$$

Notice that squaring the value of the point-biserial correlation produces $r^2 = (0.686)^2 = 0.47$, which is exactly the value of r^2 we obtained measuring effect size.

In some respects, the point-biserial correlation and the independent-measures hypothesis test are evaluating the same thing. Specifically, both are examining the relationship between the TV-viewing habits of 5-year-old children and their future academic performance in high school.

1. The correlation is measuring the *strength* of the relationship between the two variables. A large correlation (near 1.00 or –1.00) would indicate that there is a consistent, predictable relationship between high school grades and watching Sesame Street as a 5-year-old child. In particular, the value of r^2 measures how much of the variability in grades can be predicted by knowing whether the participants watched Sesame Street.
2. The *t* test evaluates the *significance* of the relationship. The hypothesis test determines whether the mean difference in grades between the two groups is greater than can be reasonably explained by chance alone.

As we noted in Chapter 10 (pp. 332–333), the outcome of the hypothesis test and the value of r^2 are often reported together. The t value measures statistical significance and r^2 measures the effect size. Also, as we noted in Chapter 10, the values for t and r^2 are directly related. In fact, either can be calculated from the other by the equations

$$r^2 = \frac{t^2}{t^2 + df} \quad \text{and} \quad t^2 = \frac{r^2}{\left(1 - r^2\right)/df}$$

where df is the degrees of freedom for the t statistic.

However, you should note that r^2 is determined entirely by the size of the correlation, whereas t is influenced by the size of the correlation and the size of the sample. For example, a correlation of $r = 0.30$ produces $r^2 = 0.09$ (9%) no matter how large the sample may be. On the other hand, a point-biserial correlation of $r = 0.30$ for a total sample of 10 people ($n = 5$ in each group) produces a nonsignificant value of $t = 0.889$. If the sample is increased to 50 people ($n = 25$ in each group), the same correlation produces a significant t value of $t = 2.18$. Although t and r are related, they are measuring different things.

POINT-BISERIAL CORRELATION, PARTIAL CORRELATION, AND EFFECT SIZE FOR THE REPEATED-MEASURES *t* TEST

In the previous section we demonstrated that the point-biserial correlation produces an r value that is directly related to the r^2 value used to measure effect size for the independent-measures t test. With one modification, this same process can be duplicated for the repeated-measures t test. The modification involves using a partial correlation (see pp. 531–535) to control for individual differences.

You should recall from Chapters 11 and 13 that one of the major distinctions between independent-measures and repeated-measures designs is that the repeated-measures designs eliminate the influence of individual differences. When computing a point-biserial correlation for repeated-measures data, we can use a partial correlation to eliminate individual differences once again.

The left-hand side of Table 15.5 shows data from a repeated-measures study comparing two treatments with a sample of $n = 4$ participants. Note that we have added a column of P values, or participant totals, showing the sum of the two scores for each participant. For example, participant A has scores of 3 and 5, which add to $P = 8$. The P values provide an indication of the individual differences. Participant A, for example, has consistently smaller scores and a smaller P value than all of the other participants. These data produce $t = 2.00$ with $df = 3$, which results in $r^2 = 4/(4 + 3) = 0.5714$ as the measure of effect size.

TABLE 15.5

The data on the left represent scores from a repeated-measures study comparing two treatments with a sample of $n = 4$ participants. The data on the right are the same scores in a format compatible with the point-biserial correlation. The P values in each set of data show the sum of the two scores for each participant and provide a measure of individual differences.

	Treatment		
Participant	I	II	*P*
A	3	5	8
B	4	14	18
C	5	7	12
D	4	6	10

Score (*X*)	Treatment (*Y*)	*P*
3	0	8
4	0	18
5	0	12
4	0	10
5	1	8
14	1	18
7	1	12
6	1	10

On the right-hand side of Table 15.5 we have reorganized the data into a format compatible with the point-biserial correlation. The individual scores, or X values are listed in the first column. The second column, or Y values, are numerical codes corresponding to the two treatment conditions: Treatment I = 0 and Treatment II = 1. The third column contains the P value for each individual, which measures the individual differences between participants. For these data, the partial correlation between X and Y, controlling for the P values, is

$$r_{XY\text{-}P} = 0.756$$

Note that this is a slightly modified point-biserial correlation. The modification is that we used a partial correlation to control the individual differences. However, squaring this correlation produces $r^2 = (0.756)^2 = 0.5715$, which is identical, within rounding error, to the r^2 value that measures effect size for the repeated-measures t test.

THE PHI-COEFFICIENT

When both variables (X and Y) measured for each individual are dichotomous, the correlation between the two variables is called the *phi-coefficient.* To compute phi (ϕ), you follow a two-step procedure:

1. Convert each of the dichotomous variables to numerical values by assigning a 0 to one category and a 1 to the other category for each of the variables.
2. Use the regular Pearson formula with the converted scores.

This process is demonstrated in the following example.

EXAMPLE 15.13

A researcher is interested in examining the relationship between birth-order position and personality. A random sample of $n = 8$ individuals is obtained, and each individual is classified in terms of birth-order position as first-born or only child versus later-born. Then each individual's personality is classified as either introvert or extrovert.

The original measurements are then converted to numerical values by the following assignments:

Birth Order	Personality
1st or only child = 0	Introvert = 0
Later-born child = 1	Extrovert = 1

The original data and the converted scores are as follows:

Original Data		Converted Scores	
Birth Order (X)	Personality (Y)	Birth Order (X)	Personality (Y)
1st	Introvert	0	0
3rd	Extrovert	1	1
Only	Extrovert	0	1
2nd	Extrovert	1	1
4th	Extrovert	1	1
2nd	Introvert	1	0
Only	Introvert	0	0
3rd	Extrovert	1	1

The Pearson correlation formula is then used with the converted data to compute the phi-coefficient.

Because the assignment of numerical values is arbitrary (either category could be designated 0 or 1), the sign of the resulting correlation is meaningless. As with most correlations, the *strength* of the relationship is best described by the value of r^2, the coefficient of determination, which measures how much of the variability in one variable is predicted or determined by the association with the second variable.

We also should note that although the phi-coefficient can be used to assess the relationship between two dichotomous variables, the more common statistical procedure is a chi-square statistic, which is examined in Chapter 17.

LEARNING CHECK

1. Define a *dichotomous* variable.

2. The following data represent job-related stress scores for a sample of $n = 8$ individuals. These people also are classified by salary level.

a. Convert the data into a form suitable for the point-biserial correlation.

b. Compute the point-biserial correlation for these data.

Salary More than $40,000	Salary Less than $40,000
8	4
6	2
5	1
3	3

3. A researcher would like to know whether there is a relationship between gender and manual dexterity for 3-year-old children. A sample of $n = 10$ boys and $n = 10$ girls is obtained and each child is given a manual-dexterity test. Five of the girls failed the test and only two of the boys failed. Describe how these data could be coded into a form suitable for computing a phi-coefficient to measure the strength of the relationship.

ANSWERS

1. A dichotomous variable has only two possible values.

2. **a.** Salary level is a dichotomous variable and can be coded as $Y = 1$ for individuals with salary more than $40,000 and $Y = 0$ for salary less than $40,000. The stress scores produce $SS_X = 36$, the salary codes produce $SS_Y = 2$, and $SP = 6$.

b. The point-biserial correlation is 0.71.

3. Gender could be coded with male = 0 and female = 1. Manual dexterity could be coded with failure = 0 and success = 1. Eight boys would have scores of 0 and 1 and two would have scores of 0 and 0. Five girls would have scores of 1 and 1 and five would have scores of 1 and 0.

SUMMARY

1. A correlation measures the relationship between two variables, X and Y. The relationship is described by three characteristics:
   - **a.** *Direction*. A relationship can be either positive or negative. A positive relationship means that X and Y vary in the same direction. A negative relationship means that X and Y vary in opposite directions. The sign of the correlation (+ or –) specifies the direction.
   - **b.** *Form*. The most common form for a relationship is a straight line, which is measured by the Pearson correlation. Other correlations measure the consistency or strength of the relationship, independent of any specific form.
   - **c.** *Strength or consistency*. The numerical value of the correlation measures the strength or consistency of the relationship. A correlation of 1.00 indicates a perfectly consistent relationship and 0.00 indicates no relationship at all. For the Pearson correlation, $r = 1.00$ (or –1.00) means that the data points fit perfectly on a straight line.
2. The most commonly used correlation is the Pearson correlation, which measures the degree of linear relationship. The Pearson correlation is identified by the letter r and is computed by

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

   In this formula, SP is the sum of products of deviations and can be calculated with either a definitional formula or a computational formula:

$$\text{definitional formula: } SP = \Sigma(X - M_X)(Y - M_Y)$$

$$\text{computational formula: } SP = \Sigma XY - \frac{\Sigma X \Sigma Y}{n}$$

3. A correlation between two variables should not be interpreted as implying a causal relationship. Simply because X and Y are related does not mean that X causes Y or that Y causes X.
4. To evaluate the strength of a relationship, you square the value of the correlation. The resulting value, r^2, is called the *coefficient of determination* because it measures the portion of the variability in one variable that can be predicted using the relationship with the second variable.
5. A partial correlation measures the linear relationship between two variables by eliminating the influence of a third variable by holding it constant.
6. The Spearman correlation (r_S) measures the consistency of direction in the relationship between X and Y—that is, the degree to which the relationship is one-directional, or monotonic. The Spearman correlation is computed by a two-stage process:
   - **a.** Rank the X scores and the Y scores separately.
   - **b.** Compute the Pearson correlation using the ranks.
7. The point-biserial correlation is used to measure the strength of the relationship when one of the two variables is dichotomous. The dichotomous variable is coded using values of 0 and 1, and the regular Pearson formula is applied. Squaring the point-biserial correlation produces the same r^2 value that is obtained to measure effect size for the independent-measures t test. When both variables, X and Y, are dichotomous, the phi-coefficient can be used to measure the strength of the relationship. Both variables are coded 0 and 1, and the Pearson formula is used to compute the correlation.

KEY TERMS

correlation (510)
positive correlation (512)
negative correlation (512)
perfect correlation (513)
Pearson correlation (514)
sum of products (*SP*) (515)
restricted range (522)
coefficient of determination (524)
regression toward the mean (526)
correlation matrix (530)
partial correlation (531)
Spearman correlation (535)
point-biserial correlation (542)
phi-coefficient (545)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 15 on the book companion website. The website also provides access to two workshops entitled *Correlation* and *Bivariate Scatter Plots,* which include information on regression.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Pearson, Spearman, point-biserial, and partial correlations.** *Note:* We focus on the Pearson correlation and then describe how slight modifications to this procedure can be made to compute the Spearman, point-biserial, and partial correlations. Separate instructions for the **phi-coefficient** are presented at the end of this section.

Data Entry

The data are entered into two columns in the data editor, one for the *X* values (VAR00001) and one for the *Y* values (VAR00002), with the two scores for each individual in the same row.

Data Analysis

1. Click **Analyze** on the tool bar, select **Correlate,** and click on **Bivariate.**
2. One by one, move the labels for the two data columns into the **Variables** box. (Hightlight each label and click the arrow to move it into the box.)

3. The **Pearson** box should be checked but, at this point, you can switch to the Spearman correlation by clicking the appropriate box.
4. Click **OK.**

SPSS Output

We used SPSS to compute the correlation for the data in Example 15.3 and the output is shown in Figure 15.16. The program produces a correlation matrix showing all the possible correlations, including the correlation of *X* with *X* and the correlation of *Y* with *Y* (both are perfect correlations). You want the correlation of *X* and *Y*, which is contained in the upper right corner (or the lower left). The output includes the significance level (*p* value or alpha level) for the correlation.

To compute a partial correlation, click **Analyze** on the tool bar, select **Correlate,** and click on **Partial.** Move the column labels for the two variables to be correlated into the **Variables** box and move the column label for the variable to be held constant into the **Controlling for** box and click **OK.**

To compute the **Spearman** correlation, enter either the *X* and *Y* ranks or the *X* and *Y* scores into the first two columns. Then follow the same Data Analysis instructions that were presented for the Pearson correlation. At step 3 in the instructions, click on the **Spearman** box before the final OK. (*Note:* If you enter *X* and *Y* scores into the data editor, SPSS converts the scores to ranks before computing the Spearman correlation.)

To compute the **point-biserial** correlation, enter the scores (*X* values) in the first column and enter the numerical values (usually 0 and 1) for the dichotomous variable in the second column. Then, follow the same Data Analysis instructions that were presented for the Pearson correlation.

The **phi-coefficient** can also be computed by entering the complete string of 0s and 1s into two columns of the SPSS data editor, then following the same Data Analysis instructions that were presented for the Pearson correlation. However, this can be tedious, especially with a large set of scores. The following is an alternative procedure for computing the phi-coefficient with large data sets.

FIGURE 15.16
The SPSS output for the correlation in Example 15.3.

Correlations

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	.875
	Sig. (2-tailed)		.052
	N	5	5
VAR00002	Pearson Correlation	.875	1
	Sig. (2-tailed)	.052	
	N	5	5

Data Entry

1. Enter the values, 0, 0, 1, 1 (in order) into the first column of the SPSS data editor.
2. Enter the values 0, 1, 0, 1 (in order) into the second column.
3. Count the number of individuals in the sample who are classified with $X = 0$ and $Y = 0$. Enter this frequency in the top box in the third column of the data editor. Then, count how many have $X = 0$ and $Y = 1$ and enter the frequency in the second box of the third column. Continue with the number who have $X = 1$ and $Y = 0$, and finally the number who have $X = 1$ and $Y = 1$. You should end up with 4 values in column three.
4. Click **Data** on the Tool Bar at the top of the SPSS Data Editor page and select **Weight Cases** at the bottom of the list.
5. Click the circle labeled **Weight cases by,** and then highlight the label for the column containing your frequencies (VAR00003) on the left and move it into the **Frequency Variable** box by clicking on the arrow.
6. Click **OK**.
7. Click **Analyze** on the tool bar, select **Correlate,** and click on **Bivariate.**
8. One by one, move the labels for the two data columns containing the 0s and 1s (probably VAR00001 and VAR00002) into the **Variables** box. (Highlight each label and click the arrow to move it into the box.)
9. Verify that the **Pearson** box is checked.
10. Click **OK.**

SPSS Output

The program produces the same correlation matrix that was described for the Pearson correlation. Again, you want the correlation between *X* and *Y*, which is in the upper right corner (or lower left). Remember, with the phi-coefficient, the sign of the correlation is meaningless.

FOCUS ON PROBLEM SOLVING

1. A correlation always has a value from +1.00 to –1.00. If you obtain a correlation outside this range, then you have made a computational error.
2. When interpreting a correlation, do not confuse the sign (+ or –) with its numerical value. The sign and the numerical value must be considered separately. Remember that the sign indicates the direction of the relationship between *X* and *Y*. On the other hand, the numerical value reflects the strength of the relationship or how well the points approximate a linear (straight-line) relationship. Therefore, a correlation of –0.90 is as strong as a correlation of +0.90. The signs tell us that the first correlation is an inverse relationship.
3. Before you begin to calculate a correlation, sketch a scatter plot of the data and make an estimate of the correlation. (Is it positive or negative? Is it near 1 or near 0?) After computing the correlation, compare your final answer with your original estimate.
4. The definitional formula for the sum of products (*SP*) should be used only when you have a small set (*n*) of scores and the means for *X* and *Y* are both whole numbers. Otherwise, the computational formula produces quicker, easier, and more accurate results.

5. For computing a correlation, n is the number of individuals (and therefore the number of *pairs* of X and Y values).

DEMONSTRATION 15.1

CORRELATION

Calculate the Pearson correlation for the following data:

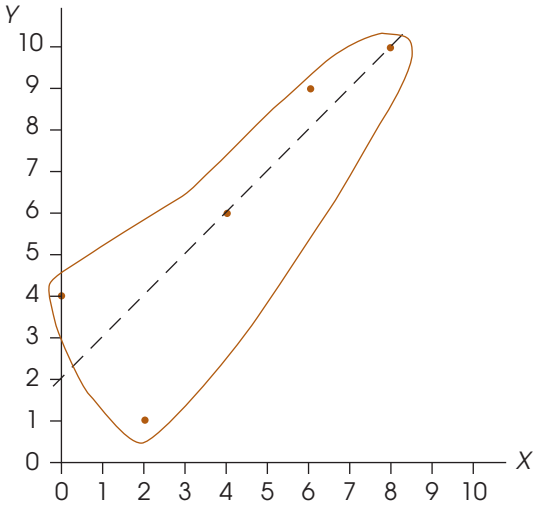
Person	X	Y	
A	0	4	$M_X = 4$ with $SS_X = 40$
B	2	1	$M_Y = 6$ with $SS_Y = 54$
C	8	10	$SP = 40$
D	6	9	
E	4	6	

STEP 1 **Sketch a scatter plot.** We have constructed a scatter plot for the data (Figure 15.17) and placed an envelope around the data points to make a preliminary estimate of the correlation. Note that the envelope is narrow and elongated. This indicates that the correlation is large—perhaps 0.80 to 0.90. Also, the correlation is positive because increases in X are generally accompanied by increases in Y.

STEP 2 **Compute the Pearson correlation.** For these data, the Pearson correlation is

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{40}{\sqrt{40(54)}} = \frac{40}{\sqrt{2160}} = \frac{40}{46.48} = 0.861$$

FIGURE 15.17
The scatter plot for the data of Demonstration 15.1. An envelope is drawn around the points to estimate the magnitude of the correlation. A line is drawn through the middle of the envelope.



In step 1, our preliminary estimate for the correlation was between +0.80 and +0.90. The calculated correlation is consistent with this estimate.

STEP 3 **Evaluate the significance of the correlation.** The null hypothesis states that, for the population, there is no linear relationship between X and Y, and that the value obtained for the sample correlation is simply the result of sampling error. Specifically, H_0 says that the population correlation is zero ($\rho = 0$). With $n = 5$ pairs of X and Y values the test has $df = 3$. Table B.6 lists a critical value of 0.878 for a two-tailed test with $\alpha = .05$. Because our correlation is smaller than this value, we fail to reject the null hypothesis and conclude that the correlation is not significant.

PROBLEMS

1. What information is provided by the sign (+ or –) of the Pearson correlation?

2. What information is provided by the numerical value of the Pearson correlation?

3. Calculate *SP* (the sum of products of deviations) for the following scores. *Note:* Both means are whole numbers, so the definitional formula works well

X	Y
0	2
1	4
4	5
3	3
7	6

4. Calculate *SP* (the sum of products of deviations) for the following scores. *Note:* Both means are decimal values, so the computational formula works well.

X	Y
0	2
0	1
1	0
2	1
1	2
0	3

5. For the following scores,

X	Y
7	6
9	6
6	3
12	5
9	6
5	4

a. Sketch a scatter plot showing the six data points.
b. Just looking at the scatter plot, estimate the value of the Pearson correlation.
c. Compute the Pearson correlation.

6. For the following scores,

X	Y
1	3
3	5
2	1
2	3

a. Sketch a scatter plot and estimate the Pearson correlation.
b. Compute the Pearson correlation.

7. For the following scores,

X	Y
1	7
4	2
1	3
1	6
2	0
0	6
2	3
1	5

a. Sketch a scatter plot and estimate the Pearson correlation.
b. Compute the Pearson correlation.

8. For the following scores,

X	Y
1	6
4	1
1	4
1	3
3	1

a. Sketch a scatter plot and estimate the value of the Pearson correlation.
b. Compute the Pearson correlation.

9. With a small sample, a single point can have a large effect on the magnitude of the correlation. To create the following data, we started with the scores from problem 8 and changed the first X value from $X = 1$ to $X = 6$.

X	Y
6	6
4	1
1	4
1	3
3	1

a. Sketch a scatter plot and estimate the value of the Pearson correlation.
b. Compute the Pearson correlation.

10. For the following set of scores,

X	Y
6	4
3	1
5	0
6	7
4	2
6	4

a. Compute the Pearson correlation.
b. Add 2 points to each X value and compute the correlation for the modified scores. How does adding a constant to every score affect the value of the correlation?
c. Multiply each of the original X values by 2 and compute the correlation for the modified scores. How does multiplying each score by a constant affect the value of the correlation?

11. Correlation studies are often used to help determine whether certain characteristics are controlled more by genetic influences or by environmental influences. These studies often examine adopted children and compare their behaviors with the behaviors of their birth parents and their adoptive parents. One study examined how much time individuals spend watching TV (Plomin, Corley, DeFries, & Fulker, 1990). The following data are similar to the results obtained in the study.

Amount of Time Spent Watching TV

Adopted Children	Birth Parents	Adoptive Parents
2	0	1
3	3	4
6	4	2
1	1	0
3	1	0
0	2	3
5	3	2
2	1	3
5	3	3

a. Compute the correlation between the children and their birth parents.
b. Compute the correlation between the children and their adoptive parents.
c. Based on the two correlations, does TV watching appear to be inherited from the birth parents or is it learned from the adoptive parents?

12. Judge and Cable (2010) report the results of a study demonstrating a negative relationship between weight and income for a group of women professionals. Following are data similar to those obtained in the study. To simplify the weight variable, the women are classified into five categories that measure actual weight relative to height, from 1 = thinnest to 5 = heaviest. Income figures are annual income (in thousands), rounded to the nearest $1,000.
a. Calculate the Pearson correlation for these data.
b. Is the correlation statistically significant? Use a two-tailed test with $\alpha = .05$.

Weight (X)	Income (Y)
1	125
2	78
4	49
3	63
5	35
2	84
5	38
3	51
1	93
4	44

13. The researchers cited in the previous problem also examined the weight/salary relationship for men and found a positive relationship, suggesting that we have very different standards for men than for women

(Judge & Cable, 2010). The following are data similar to those obtained for working men. Again, weight relative to height is coded in five categories from 1 = thinnest to 5 = heaviest. Income is recorded as thousands earned annually.

a. Calculate the Pearson correlation for these data.

b. Is the correlation statistically significant? Use a two-tailed test with $\alpha = .05$.

Weight (*X*)	Income (*Y*)
4	156
3	88
5	49
2	73
1	45
3	92
1	53
5	148

14. Identifying individuals with a high risk of Alzheimer's disease usually involves a long series of cognitive tests. However, researchers have developed a 7-Minute Screen, which is a quick and easy way to accomplish the same goal. The question is whether the 7-Minute Screen is as effective as the complete series of tests. To address this question, Ijuin et al. (2008) administered both tests to a group of patients and compared the results. The following data represent results similar to those obtained in the study.

Patient	7-Minute Screen	Cognitive Series
A	3	11
B	8	19
C	10	22
D	8	20
E	4	14
F	7	13
G	4	9
H	5	20
I	14	25

a. Compute the Pearson correlation to measure the degree of relationship between the two test scores.

b. Is the correlation statistically significant? Use a two-tailed test with $\alpha = .01$.

c. What percentage of variance for the cognitive scores is predicted from the 7-Minute Screen scores? (Compute the value of r^2.)

15. Assuming a two-tailed test with $\alpha = .05$, how large a correlation is needed to be statistically significant for each of the following samples?

a. A sample of $n = 8$

b. A sample of $n = 18$

c. A sample of $n = 28$

16. As we have noted in previous chapters, even a very small effect can be significant if the sample is large enough. For each of the following, determine how large a sample is necessary for the correlation to be significant. Assume a two-tailed test with $\alpha = .05$. (*Note:* The table does not list all the possible *df* values. Use the sample size corresponding to the appropriate *df* value that is listed in the table.)

a. A correlation of $r = 0.30$.

b. A correlation of $r = 0.25$.

c. A correlation of $r = 0.20$.

17. A researcher measures three variables, *X*, *Y*, and *Z*, for each individual in a sample of $n = 25$. The Pearson correlations for this sample are $r_{XY} = 0.8$, $r_{XZ} = 0.6$, and $r_{YZ} = 0.7$.

a. Find the partial correlation between *X* and *Y*, holding *Z* constant.

b. Find the partial correlation between *X* and *Z*, holding *Y* constant. (*Hint:* Simply switch the labels for the variables *Y* and *Z* to correspond with the labels in the equation.)

18. A researcher records the annual number of serious crimes and the amount spent on crime prevention for several small towns, medium cities, and large cities across the country. The resulting data show a strong positive correlation between the number of serious crimes and the amount spent on crime prevention. However, the researcher suspects that the positive correlation is actually caused by population; as population increases, both the amount spent on crime prevention and the number of crimes also increases. If population is controlled, there probably should be a negative correlation between the amount spent on crime prevention and the number of serious crimes. The following data show the pattern of results obtained. Note that the municipalities are coded in three categories. Use a partial correlation, holding population constant, to measure the true relationship between crime rate and the amount spent on prevention.

Number of Crimes	Amount for Prevention	Population Size
3	6	1
4	7	1
6	3	1
7	4	1
8	11	2
9	12	2
11	8	2
12	9	2
13	16	3
14	17	3
16	13	3
17	14	3

19. A common concern for students (and teachers) is the assignment of grades for essays or term papers. Because there are no absolute right or wrong answers, these grades must be based on a judgment of quality. To demonstrate that these judgments actually are reliable, an English instructor asked a colleague to rank-order a set of term papers. The ranks and the instructor's grades for these papers are as follows:

Rank	Grade
1	A
2	B
3	A
4	B
5	B
6	C
7	D
8	C
9	C
10	D
11	F

a. Compute the Spearman correlation for these data. (*Note:* You must convert the letter grades to ranks, using tied ranks to represent tied grades.)

b. Is the Spearman correlation statistically significant? Use a two-tailed test with $\alpha = .05$.

20. It appears that there is a significant relationship between cognitive ability and social status, at least for birds. Boogert, Reader, and Laland (2006) measured social status and individual learning ability for a group of starlings. The following data represent results similar to those obtained in the study. Because social status is an ordinal variable consisting of five ordered categories, the Spearman correlation is appropriate for these data. Convert the social status categories and the learning scores to ranks, and compute the Spearman correlation.

Subject	Social Status	Learning Score
A	1	3
B	3	10
C	2	7
D	3	11
E	5	19
F	4	17
G	5	17
H	2	4
I	4	12
J	2	3

21. Problem 12 presented data showing a negative relationship between weight and income for a sample of working women. However, weight was coded in five categories, which could be viewed as an ordinal scale rather than an interval or ratio scale. If so, a Spearman correlation is more appropriate than a Pearson correlation.

a. Convert the weights and the incomes into ranks and compute the Spearman correlation for the scores in problem 12.

b. Is the Spearman correlation large enough to be significant?

22. Problem 22 in Chapter 10 presented data showing that mature soccer players, who have a history of hitting soccer balls with their heads, had significantly lower cognitive scores than mature swimmers, who do not suffer repeated blows to the head. The independent-measures t test produced $t = 2.11$ with $df = 11$ and a value of $r^2 = 0.288$ (28.8%).

a. Convert the data from this problem into a form suitable for the point-biserial correlation (use 1 for the swimmers and 0 for the soccer players), and then compute the correlation.

b. Square the value of the point-biserial correlation to verify that you obtain the same r^2 value that was computed in Chapter 10.

23. Problem 14 in Chapter 10 described a study by Rozin, Bauer, and Cantanese (2003) comparing attitudes toward eating for male and female college students. The results showed that females are much more concerned about weight gain and other negative aspects of eating than are males. The following data represent the results from one measure of concern about weight gain.

Males	Females
22	54
44	57
39	32
27	53
35	49
19	41
	35
	36
	48

Convert the data into a form suitable for the point-biserial correlation and compute the correlation.

24. Studies have shown that people with high intelligence are generally more likely to volunteer as participants in research, but not for research that involves unusual experiences such as hypnosis. To examine this phenomenon, a researcher administers a questionnaire to a sample of college students. The survey asks for the student's grade point average (as a measure of intelligence) and whether the student would like to take part in a future study in which participants would be hypnotized. The results showed that 7 of the 10 lower-intelligence people were willing to participant but only 2 of the 10 higher-intelligence people were willing.

a. Convert the data to a form suitable for computing the phi-coefficient. (Code the two intelligence categories as 0 and 1 for the *X* variable, and code the willingness to participate as 0 and 1 for the *Y* variable.)

b. Compute the phi-coefficient for the data.

CHAPTER

16

Introduction to Regression

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Sum of squares (*SS*) (Chapter 4)
  - Computational formula
  - Definitional formula
- *z*-scores (Chapter 5)
- Analysis of variance (Chapter 12)
  - *MS* values and *F*-ratios
- Pearson correlation (Chapter 15)
  - Sum of products (*SP*)

Preview

In Chapter 15, we noted that one common application of correlations is for purposes of prediction. Whenever there is a consistent relationship between two variables, it is possible to use the value of one variable to predict the value of another. Managers at the electric company, for example, can use the weather forecast to predict power demands for upcoming days. If exceptionally hot summer weather is forecast, they can anticipate an exceptionally high demand for electricity. In the field of psychology, a known relationship between certain personality characteristics and eating disorders can allow clinicians to predict that individuals who show specific characteristics are more likely to develop disorders. A common prediction that is especially relevant for college students (and potential college students) is based on the relationship between scores on aptitude tests (such as the SAT) and future grade point averages in college. Each year, SAT scores from thousands of high school students are used to help college admissions officers decide who should be admitted and who should not.

The Problem: The correlations introduced in Chapter 15 allow researchers to measure and describe relationships, and the hypothesis tests allow researchers to evaluate the significance of correlations. However, we now want to go one step further and actually use a correlation to make predictions.

The Solution: In this chapter we introduce some of the statistical techniques that are used to make predictions based on correlations. Whenever there is a linear relationship (Pearson correlation) between two variables, it is possible to compute an equation that provides a precise, mathematical description of the relationship. With the equation, it is possible to plug in the known value for one variable (for example, your SAT score), and then calculate a predicted value for the second variable (for example, your college grade point average). The general statistical process of finding and using a prediction equation is known as *regression*.

Beyond finding a prediction equation, however, it is reasonable to ask how good its predictions are. For example, I can make predictions about the outcome of a coin toss by simply guessing. However, my predictions are correct only about 50% of the time. In statistical terms, my predictions are not significantly better than chance. In the same way, it is appropriate to challenge the significance of any prediction equation. In this chapter we introduce the techniques that are used to find prediction equations, as well as the techniques that are used to determine whether their predictions are statistically significant. Incidentally, although there is some controversy about the practice of using SAT scores to predict college performance, there is a great deal of research showing that SAT scores really are valid and significant predictors (Camera & Echternacht, 2000; Geiser & Studley, 2002).

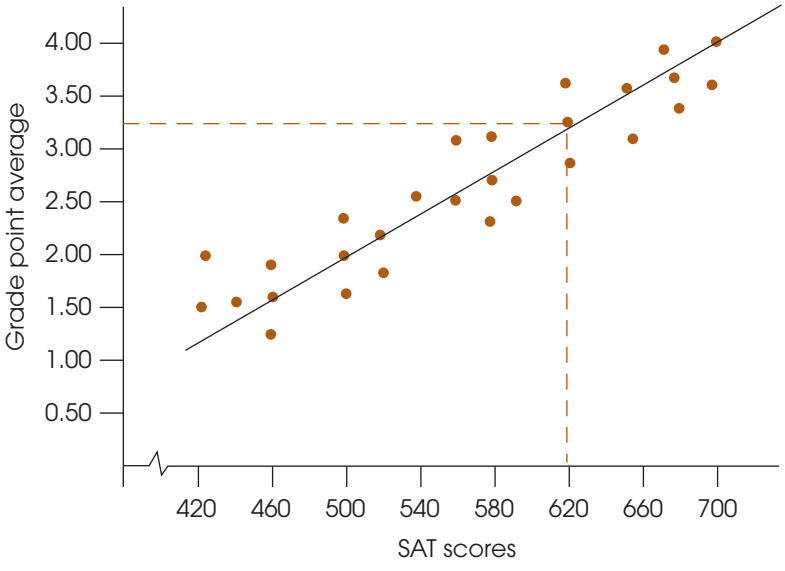
16.1 INTRODUCTION TO LINEAR EQUATIONS AND REGRESSION

In the previous chapter, we introduced the Pearson correlation as a technique for describing and measuring the linear relationship between two variables. Figure 16.1 presents hypothetical data showing the relationship between SAT scores and college grade point average (GPA). Note that the figure shows a good, but not perfect, positive relationship. Also note that we have drawn a line through the middle of the data points. This line serves several purposes:

1. The line makes the relationship between SAT scores and GPA easier to see.
2. The line identifies the center, or *central tendency,* of the relationship, just as the mean describes central tendency for a set of scores. Thus, the line provides a simplified description of the relationship. For example, if the data points were removed, the straight line would still give a general picture of the relationship between SAT scores and GPA.
3. Finally, the line can be used for prediction. The line establishes a precise, one-to-one relationship between each X value (SAT score) and a corresponding Y value (GPA). For example, an SAT score of 620 corresponds to a GPA of 3.25 (see Figure 16.1). Thus, the college admissions officers could use the straight-line

FIGURE 16.1

Hypothetical data showing the relationship between SAT scores and GPA with a regression line drawn through the data points. The regression line defines a precise, one-to-one relationship between each *X* value (SAT score) and its corresponding *Y* value (GPA).



relationship to predict that a student entering college with an SAT score of 620 should achieve a college GPA of approximately 3.25.

Our goal in this section is to develop a procedure that identifies and defines the straight line that provides the best fit for any specific set of data. This straight line does not have to be drawn on a graph; it can be presented in a simple equation. Thus, our goal is to find the equation for the line that best describes the relationship for a set of *X* and *Y* data.

LINEAR EQUATIONS

In general, a *linear relationship* between two variables *X* and *Y* can be expressed by the equation

$$Y = bX + a \tag{16.1}$$

where *a* and *b* are fixed constants.

For example, a local video store charges a membership fee of $5 per month, which allows you to rent videos and games for $2 each. With this information, the total cost for 1 month can be computed using a *linear equation* that describes the relationship between the total cost (*Y*) and the number of videos and games rented (*X*).

$$Y = 2X + 5$$

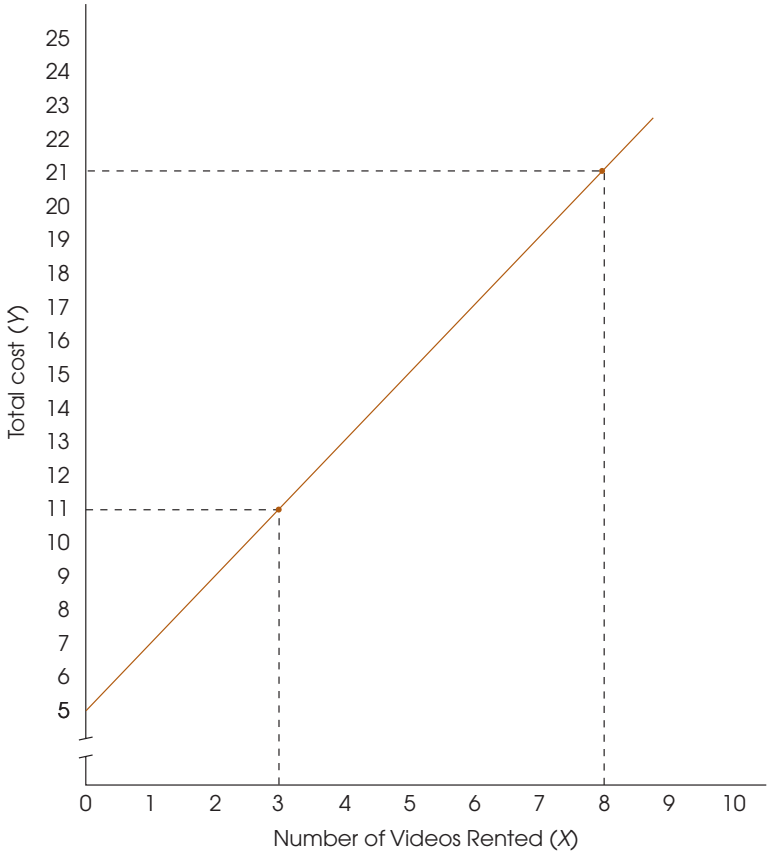
Note that a positive slope means that *Y* increases when *X* is increased, and a negative slope indicates that *Y* decreases when *X* is increased.

In the general linear equation, the value of *b* is called the *slope*. The slope determines how much the *Y* variable changes when *X* is increased by 1 point. For the video store example, the slope is $b = 2$ and indicates that your total cost increases by $2 for each video you rent. The value of *a* in the general equation is called the Y-*intercept* because it determines the value of *Y* when $X = 0$. (On a graph, the *a* value identifies the point where the line intercepts the *Y*-axis.) For the video store example, $a = 5$; there is a $5 membership charge even if you never rent a video.

Figure 16.2 shows the general relationship between the monthly cost and number of videos for the video store example. Notice that the relationship results in a straight

FIGURE 16.2

The relationship between total cost and number of videos rented each month. The video store charges a $5 monthly membership fee and $2 for each video or game rented. The relationship is described by a linear equation $Y = 2X + 5$ where Y is the total cost and X is the number of videos.



line. To obtain this graph, we picked any two values of X and then used the equation to compute the corresponding values for Y. For example,

when $X = 3$:	when $X = 8$:
$Y = bX + a$	$Y = bX + a$
$= \$2(3) + \$5$	$= \$2(8) + \$5$
$= \$6 + \$5$	$= \$16 + \$5$
$= \$11$	$= \$21$

When drawing a graph of a linear equation, it is wise to compute and plot at least three points to be certain that you have not made a mistake.

Next, these two points are plotted on the graph: one point at $X = 3$ and $Y = 11$, the other point at $X = 8$ and $Y = 21$. Because two points completely determine a straight line, we simply drew the line so that it passed through these two points.

LEARNING CHECK

1. A local gym charges a $25 monthly membership fee plus $2 per hour for aerobics classes. What is the linear equation that describes the relationship between the total monthly cost (Y) and the number of class hours each month (X)?
2. For the following linear equation, what happens to the value of Y each time X is increased by 1 point?

$$Y = -3X + 7$$

3. Use the linear equation $Y = 2X - 7$ to determine the value of Y for each of the following values of X: 1, 3, 5, 10.

4. If the slope constant (b) in a linear equation is positive, then a graph of the equation is a line tilted from lower left to upper right. (True or false?)

ANSWERS

1. $Y = 2X + 25$

2. The slope is -3, so Y decreases by 3 points each time X increases by 1 point.

3.

X	Y
1	-5
3	-1
5	3
10	13

4. True. A positive slope indicates that Y increases (goes up in the graph) when X increases (goes to the right in the graph).

REGRESSION

Because a straight line can be extremely useful for describing a relationship between two variables, a statistical technique has been developed that provides a standardized method for determining the best-fitting straight line for any set of data. The statistical procedure is *regression,* and the resulting straight line is called the *regression line.*

DEFINITION

The statistical technique for finding the best-fitting straight line for a set of data is called **regression,** and the resulting straight line is called the **regression line.**

The goal for regression is to find the best-fitting straight line for a set of data. To accomplish this goal, however, it is first necessary to define precisely what is meant by "best fit." For any particular set of data, it is possible to draw lots of different straight lines that all appear to pass through the center of the data points. Each of these lines can be defined by a linear equation of the form $Y = bX + a$ where b and a are constants that determine the slope and Y-intercept of the line, respectively. Each individual line has its own unique values for b and a. The problem is to find the specific line that provides the best fit to the actual data points.

THE LEAST-SQUARES SOLUTION

To determine how well a line fits the data points, the first step is to define mathematically the distance between the line and each data point. For every X value in the data, the linear equation determines a Y value on the line. This value is the predicted Y and is called $\hat{Y}$ ("Y hat"). The distance between this predicted value and the actual Y value in the data is determined by

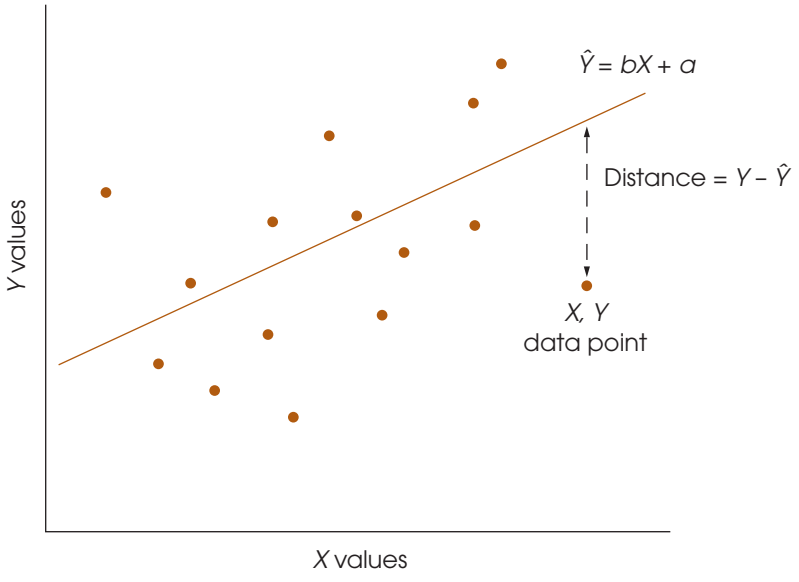
$$\text{distance} = Y - \hat{Y}$$

Note that we simply are measuring the vertical distance between the actual data point (Y) and the predicted point on the line. This distance measures the error between the line and the actual data (Figure 16.3).

Because some of these distances are positive and some are negative, the next step is to square each distance to obtain a uniformly positive measure of error. Finally, to

FIGURE 16.3

The distance between the actual data point (Y) and the predicted point on the line ($\hat{Y}$) is defined as $Y - \hat{Y}$. The goal of regression is to find the equation for the line that minimizes these distances.



determine the total error between the line and the data, we add the squared errors for all of the data points. The result is a measure of overall squared error between the line and the data:

$$\text{total squared error} = \Sigma(Y - \hat{Y})^2$$

Now we can define the *best-fitting* line as the one that has the smallest total squared error. For obvious reasons, the resulting line is commonly called the *least-squared-error solution*. In symbols, we are looking for a linear equation of the form

$$\hat{Y} = bX + a$$

For each value of X in the data, this equation determines the point on the line ($\hat{Y}$) *that gives the best prediction of Y*. The problem is to find the specific values for a and b that make this the best-fitting line.

The calculations that are needed to find this equation require calculus and some sophisticated algebra, so we do not present the details of the solution. The results, however, are relatively straightforward, and the solutions for b and a are as follows:

$$b = \frac{SP}{SS_X} \tag{16.2}$$

where SP is the sum of products and SS_X is the sum of squares for the X scores.

A commonly used alternative formula for the slope is based on the standard deviations for X and Y. The alternative formula is

$$b = r\frac{s_Y}{s_X} \tag{16.3}$$

where s_Y is the standard deviation for the Y scores, s_X is the standard deviation for the X scores, and r is the Pearson correlation for X and Y. The value of the constant a in the equation is determined by

$$a = M_Y - bM_X \tag{16.4}$$

Note that these formulas determine the linear equation that provides the best prediction of Y values. This equation is called the *regression equation for* Y.

DEFINITION

The **regression equation for *Y*** is the linear equation

$$\hat{Y} = bX + a \tag{16.5}$$

where the constant b is determined by Equation 16.2, or 16.3 and the constant a is determined by Equation 16.4. This equation results in the least squared error between the data points and the line.

EXAMPLE 16.1

The scores in the following table are used to demonstrate the calculation and use of the regression equation for predicting Y.

X	Y	$X - M_X$	$Y - M_Y$	$(X - M_X)^2$	$(Y - M_X)^2$	$(X - M_x)(Y - M_Y)$
2	3	−2	−5	4	25	10
6	11	2	3	4	9	6
0	6	−4	−2	16	4	8
4	6	0	−2	0	4	0
7	12	3	4	9	16	12
5	7	1	−1	1	1	−1
5	10	1	2	1	4	2
3	9	−1	1	1	1	−1
				$SS_X = 36$	$SS_Y = 64$	$SP = 36$

For these data, $\Sigma X = 32$, so $M_X = 4$. Also, $\Sigma Y = 64$, so $M_Y = 8$. These values have been used to compute the deviation scores for each X and Y value. The final three columns show the squared deviations for X and for Y, and the products of the deviation scores.

Our goal is to find the values for b and a in the regression equation. Using Equations 16.2 and 16.4, the solutions for b and a are

$$b = \frac{SP}{SS_X} = \frac{36}{36} = 1.00$$

$$a = M_Y - bM_X = 8 - 1(4) = 4.00$$

The resulting equation is

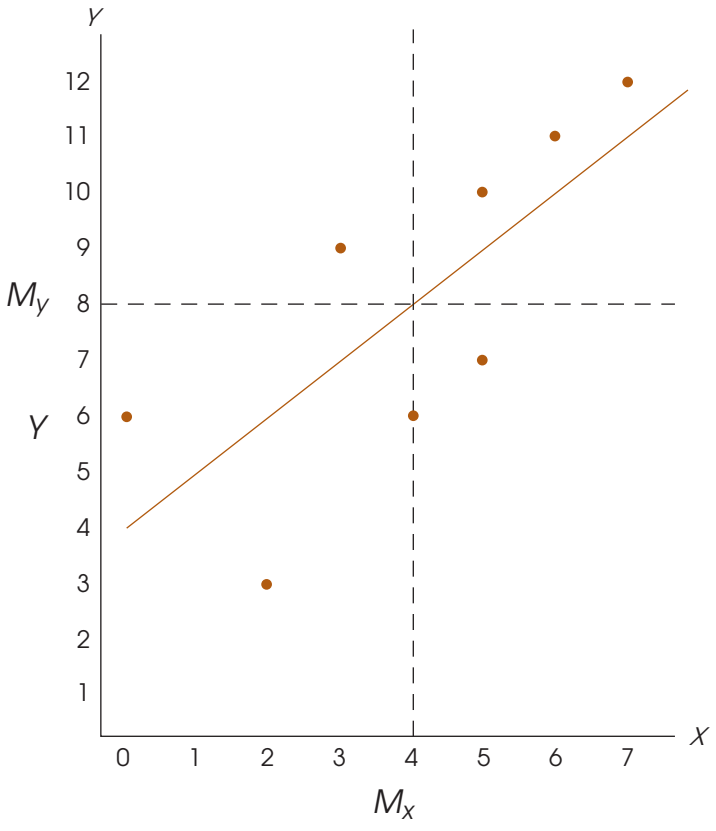
$$\hat{Y} = X + 4$$

The original data and the regression line are shown in Figure 16.4.

The regression line shown in Figure 16.4 demonstrates some simple and very predictable facts about regression. First, the calculation of the Y-intercept (Equation 16.4) ensures that the regression line passes through the point defined by the mean for X and the mean for Y. That is, the point identified by the coordinates M_X, M_Y will always be on the line. We have included the two means in Figure 16.4 to show that

FIGURE 16.4

The X and Y data points and the regression line for the $n = 8$ pairs of scores in Example 16.1.



the point they define is on the regression line. Second, the sign of the correlation (+ or −) is the same as the sign of the slope of the regression line. Specifically, if the correlation is positive, then the slope is also positive and the regression line slopes up to the right. On the other hand, if the correlation is negative, then the slope is negative and the line slopes down to the right. A correlation of zero means that the slope is also zero and the regression equation produces a horizontal line that passes through the data at a level equal to the mean for the Y values. Note that the regression line in Figure 16.4 has a positive slope. One consequence of this fact is that all of the points on the line that are above the mean for X are also above the mean for Y. Similarly, all of the points below the mean for X are also below the mean for Y. Thus, every individual with a positive deviation for X is predicted to have a positive deviation for Y, and everyone with a negative deviation for X is predicted to have a negative deviation for Y.

USING THE REGRESSION EQUATION FOR PREDICTION

As we noted at the beginning of this section, one common use of regression equations is for prediction. For any given value of X, we can use the equation to compute a predicted value for Y. For the equation from Example 16.1, an individual with a score of $X = 1$ would be predicted to have a Y score of

$$\hat{Y} = X + 4 = 1 + 4 = 5$$

Although regression equations can be used for prediction, a few cautions should be considered whenever you are interpreting the predicted values:

1. The predicted value is not perfect (unless $r = +1.00$ or -1.00). If you examine Figure 16.4, it should be clear that the data points do not fit perfectly on the line. In general, there is some error between the predicted Y values (on the line) and the actual data. Although the amount of error varies from point to point, on average the errors are directly related to the magnitude of the correlation. With a correlation near 1.00 (or -1.00), the data points generally are clustered close to the line and the error is small. As the correlation gets nearer to zero, the points move away from the line and the magnitude of the error increases.
2. The regression equation should not be used to make predictions for X values that fall outside of the range of values covered by the original data. For Example 16.1, the X values ranged from $X = 0$ to $X = 7$, and the regression equation was calculated as the best-fitting line within this range. Because you have no information about the X-Y relationship outside this range, the equation should not be used to predict Y for any X value lower than 0 or greater than 7.

STANDARDIZED FORM OF THE REGRESSION EQUATIONS

So far we have presented the regression equation in terms of the original values, or raw scores, for X and Y. Occasionally, however, researchers standardize the scores by transforming the X and Y values into z-scores before finding the regression equation. The resulting equation is often called the standardized form of the regression equation and is greatly simplified compared to the raw-score version. The simplification comes from the fact that z-scores have standardized characteristics. Specifically, the mean for a set of z-scores is always zero and the standard deviation is always 1. As a result, the standardized form of the regression equation becomes

$$\hat{z}_Y = (\text{beta})z_X \tag{16.6}$$

First notice that we are now using the z-score for each X value (z_X) to predict the z-score for the corresponding Y value (z_Y). Also, note that the slope constant that was identified as b in the raw-score formula is now identified as beta. Because both sets of z-scores have a mean of zero, the constant a disappears from the regression equation. Finally, when one variable, X, is being used to predict a second variable, Y, the value of beta is equal to the Pearson correlation for X and Y. Thus, the standardized form of the regression equation can also be written as

$$\hat{z}_Y = rz_X \tag{16.7}$$

Because the process of transforming all of the original scores into z-scores can be tedious, researchers usually compute the raw-score version of the regression equation (Equation 16.5) instead of the standardized form. However, most computer programs report the value of beta as part of the output from linear regression, and you should understand what this value represents.

LEARNING CHECK

1. Sketch a scatter plot for the following data—that is, a graph showing the X, Y data points:

X	Y
1	4
3	9
5	8

a. Find the regression equation for predicting Y from X. Draw this line on your graph. Does it look like the best-fitting line?

b. Use the regression equation to find the predicted Y value corresponding to each X in the data.

ANSWERS

1. a. $SS_X = 8$, $SP = 8$, $b = 1$, $a = 4$. The equation is $\hat{Y} = X + 4$.

b. The predicted Y values are 5, 7, and 9.

THE STANDARD ERROR OF ESTIMATE

It is possible to determine a regression equation for any set of data by simply using the formulas already presented. The linear equation you obtain is then used to generate predicted Y values for any known value of X. However, it should be clear that the accuracy of this prediction depends on how well the points on the line correspond to the actual data points—that is, the amount of error between the predicted values, $\hat{Y}$, and the actual scores, Y values. Figure 16.5 shows two different sets of data that have exactly the same regression equation. In one case, there is a perfect correlation ($r = +1$) between X and Y, so the linear equation fits the data perfectly. For the second set of data, the predicted Y values on the line only approximate the real data points.

A regression equation, by itself, allows you to make predictions, but it does not provide any information about the accuracy of the predictions. To measure the precision of the regression, it is customary to compute a *standard error of estimate.*

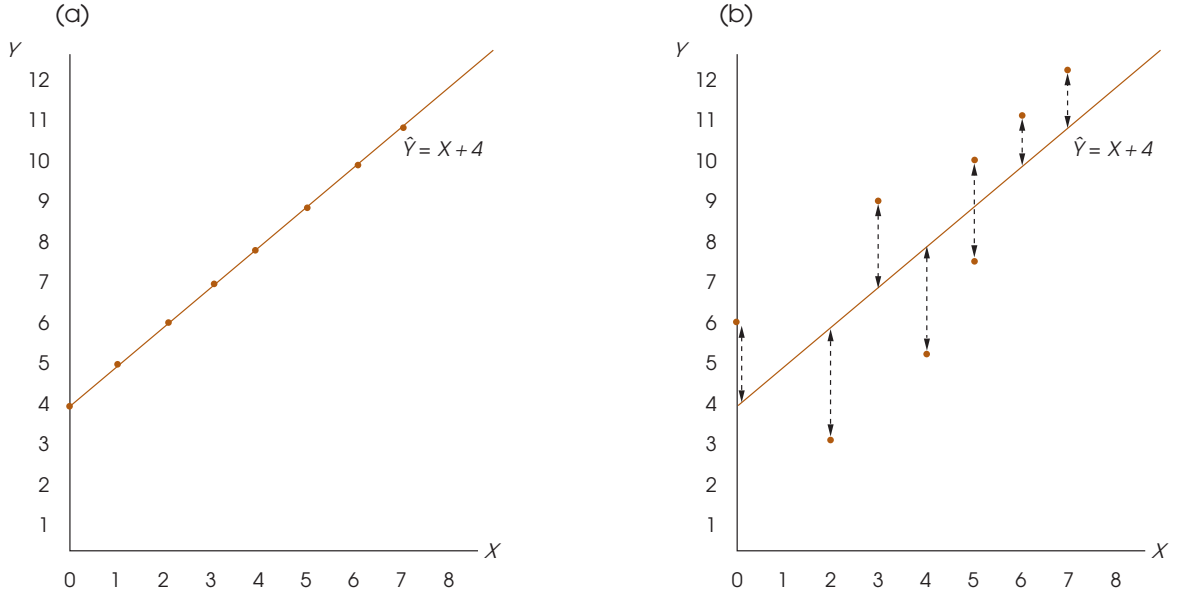


FIGURE 16.5

(a) A scatter plot showing data points that perfectly fit the regression line defined by the equation $\hat{Y} = X + 4$. Note that the correlation is $r = +1.00$. (b) A scatter plot for the data from Example 16.1. Notice that there is error between the actual data points and the predicted Y values of the regression line.

DEFINITION

The **standard error of estimate** gives a measure of the standard distance between the predicted Y values on the regression line and the actual Y values in the data.

Conceptually, the standard error of estimate is very much like a standard deviation: Both provide a measure of standard distance. Also, the calculation of the standard error of estimate is very similar to the calculation of standard deviation.

To calculate the standard error of estimate, we first find the sum of squared deviations (SS). Each deviation measures the distance between the actual Y value (from the data) and the predicted Y value (from the regression line). This sum of squares is commonly called SS_{residual} because it is based on the remaining distance between the actual Y scores and the predicted values.

$$SS_{\text{residual}} = \Sigma(Y - \hat{Y})^2 \tag{16.8}$$

The obtained SS value is then divided by its degrees of freedom to obtain a measure of variance. This procedure should be very familiar:

$$\text{Variance} = \frac{SS}{df}$$

The degrees of freedom for the standard error of estimate are $df = n - 2$. The reason for having $n - 2$ degrees of freedom, rather than the customary $n - 1$, is that we now are measuring deviations from a line rather than deviations from a mean. To find the equation for the regression line, you must know the means for both the X and the Y scores. Specifying these two means places two restrictions on the variability of the data, with the result that the scores have only $n - 2$ degrees of freedom. (*Note*: the $df = n - 2$ for SS_{residual} is the same $df = n - 2$ that we encountered when testing the significance of the Pearson correlation on page 529.)

Recall that variance measures the average squared distance.

The final step in the calculation of the standard error of estimate is to take the square root of the variance to obtain a measure of standard distance. The final equation is

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{df}} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} \tag{16.9}$$

The following example demonstrates the calculation of this standard error.

EXAMPLE 16.2

The same data that were used in Example 16.1 are used here to demonstrate the calculation of the standard error of estimate. These data have the regression equation

$$\hat{Y} = X + 4$$

Using this regression equation, we have computed the predicted Y value, the residual, and the squared residual for each individual, using the data from Example 16.1.

Data		Predicted Y Values	Residual	Squared Residual
X	Y	$\hat{Y} = X + 4$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
2	3	6	−3	9
6	11	10	1	1
0	6	4	2	4
4	6	8	−2	4
5	7	9	−2	4

(continued)

Data		Predicted Y Values	Residual	Squared Residual
X	Y	$\hat{Y} = X + 4$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
7	12	11	1	1
5	10	9	1	1
3	9	7	2	4
			0	$SS_{residual} = 28$

First note that the sum of the residuals is equal to zero. In other words, the sum of the distances above the line is equal to the sum of the distances below the line. This is true for any set of data and provides a way to check the accuracy of your calculations. The squared residuals are listed in the final column. For these data, the sum of the squared residuals is $SS_{residual} = 28$. With $n = 8$, the data have $df = n - 2 = 6$, so the standard error of estimate is

$$\text{standard error of estimate} = \sqrt{\frac{SS_{residual}}{df}} = \sqrt{\frac{28}{6}} = 2.16$$

Remember: The standard error of estimate provides a measure of how accurately the regression equation predicts the Y values. In this case, the standard distance between the actual data points and the regression line is measured by standard error of estimate = 2.16.

RELATIONSHIP BETWEEN THE STANDARD ERROR AND THE CORRELATION

It should be clear from Example 16.2 that the standard error of estimate is directly related to the magnitude of the correlation between X and Y. If the correlation is near 1.00 (or −1.00), then the data points are clustered close to the line, and the standard error of estimate is small. As the correlation gets nearer to zero, the data points become more widely scattered, the line provides less accurate predictions, and the standard error of estimate grows larger.

Earlier (p. 524), we observed that squaring the correlation provides a measure of the accuracy of prediction. The squared correlation, r^2, is called the coefficient of determination because it determines what proportion of the variability in Y is predicted by the relationship with X. Because r^2 measures the predicted portion of the variability in the Y scores, we can use the expression $(1 - r^2)$ to measure the unpredicted portion. Thus,

$$\text{predicted variability} = SS_{regression} = r^2 SS_Y \tag{16.10}$$

$$\text{unpredicted variability} = SS_{residual} = (1 - r^2)SS_Y \tag{16.11}$$

For example, if $r = 0.80$, then the *predicted variability* is $r^2 = 0.64$ (or 64%) of the total variability for the Y scores and the remaining 36% $(1 - r^2)$ is the *unpredicted variability*. Note that when $r = 1.00$, the prediction is perfect and there are no residuals. As the correlation approaches zero, the data points move farther off the line and the residuals grow larger. Using Equation 16.11 to compute $SS_{residual}$, the standard error of estimate can be computed as

$$\text{standard error of estimate} = \sqrt{\frac{SS_{residual}}{df}} = \sqrt{\frac{(1 - r^2)SS_Y}{n - 2}} \tag{16.12}$$

Because it is usually much easier to compute the Pearson correlation than to compute the individual $(Y - \hat{Y})^2$ values, Equation 16.11 is usually the easiest way to compute $SS_{residual}$, and Equation 16.12 is usually the easiest way to compute the standard error of estimate for a regression equation. The following example demonstrates this new formula.

EXAMPLE 16.3

We use the same data used in Examples 16.1 and 16.2, which produced $SS_X = 36$, $SS_Y = 64$, and $SP = 36$. For these data, the Pearson correlation is

$$r = \frac{36}{\sqrt{36(64)}} = \frac{36}{48} = 0.75$$

With $SS_Y = 64$ and a correlation of $r = 0.75$, the predicted variability from the regression equation is

$$SS_{regression} = r^2 SS_Y = (0.75^2)(64) = 0.5625(64) = 36.00$$

Similarly, the unpredicted variability is

$$SS_{residual} = (1 - r^2)SS_Y = (1 - 0.75^2)(64) = 0.4375(64) = 28.00$$

Notice that the new formula for $SS_{residual}$ produces exactly the same value that we obtained by adding the squared residuals in Example 16.2. Also note that this new formula is generally much easier to use because it requires only the correlation value (r) and the SS for Y. The primary point of this example, however, is that $SS_{residual}$ and the standard error of estimate are closely related to the value of the correlation. With a large correlation (near +1.00 or −1.00), the data points are close to the regression line, and the standard error of estimate is small. As a correlation gets smaller (near zero), the data points move away from the regression line, and the standard error of estimate gets larger.

Because it is possible to have the same regression equation for several different sets of data, it is also important to consider r^2 and the standard error of estimate. The regression equation simply describes the best-fitting line and is used for making predictions. However, r^2 and the standard error of estimate indicate how accurate these predictions are.

LEARNING CHECK

1. Describe what is measured by the standard error of estimate for a regression equation.
2. As the numerical value of a correlation increases, what happens to the standard error of estimate?
3. A sample of $n = 6$ pairs of X and Y scores produces a correlation of $r = 0.80$ and $SS_Y = 100$. What is the standard error of estimate for the regression equation?

ANSWERS

1. The standard error of estimate measures the average, or standard, distance between the predicted Y values on the regression line and the actual Y values in the data.
2. A larger correlation means that the data points are clustered closer to the line, which means the standard error of estimate is smaller.
3. The standard error of estimate $= \sqrt{36/4} = 3$.

16.2 ANALYSIS OF REGRESSION: TESTING THE SIGNIFICANCE OF THE REGRESSION EQUATION

As we noted in Chapter 15, a sample correlation is expected to be representative of its population correlation. For example, if the population correlation is zero, then the sample correlation is expected to be near zero. Note that we do not expect the sample correlation to be exactly equal to zero. This is the general concept of *sampling error* that was introduced in Chapter 1 (p. 8). The principle of sampling error is that there is typically some discrepancy or error between the value obtained for a sample statistic and the corresponding population parameter. Thus, when there is no relationship whatsoever in the population, a correlation of $\rho = 0$, you are still likely to obtain a nonzero value for the sample correlation. In this situation, however, the sample correlation is caused by chance and a hypothesis test usually demonstrates that the correlation is not significant.

Whenever you obtain a nonzero value for a sample correlation, you also obtain real, numerical values for the regression equation. However, if there is no real relationship in the population, both the sample correlation and the regression equation are meaningless—they are simply the result of sampling error and should not be viewed as an indication of any relationship between X and Y. In the same way that we tested the significance of a Pearson correlation, we can test the significance of the regression equation. In fact, when a single variable, X, is being used to predict a single variable, Y, the two tests are equivalent. In each case, the purpose for the test is to determine whether the sample correlation represents a real relationship or is simply the result of sampling error. For both tests, the null hypothesis states that there is no relationship between the two variables in the population. A more specific null hypothesis for testing the significance of a regression equation is that the equation does not account for a significant proportion of the variance in the Y scores. An alternative version of H_0 states that the values of b or beta that are computed for the regression equation do not represent any real relationship between X and Y but rather are simply the result of chance or sampling error. In other words, the true population value of b or beta is zero.

The process of testing the significance of a regression equation is called *analysis of regression* and is very similar to the analysis of variance (ANOVA) presented in Chapter 12. As with ANOVA, the regression analysis uses an F-ratio to determine whether the variance predicted by the regression equation is significantly greater than would be expected if there were no relationship between X and Y. The F-ratio is a ratio of two variances, or mean square (MS) values, and each variance is obtained by dividing an SS value by its corresponding degrees of freedom. The numerator of the F-ratio is $MS_{\text{regression}}$, which is the variance in the Y scores that is predicted by the regression equation. This variance measures the systematic changes in Y that occur when the value of X increases or decreases. The denominator is MS_{residual}, which is the unpredicted variance in the Y scores. This variance measures the changes in Y that are independent of changes in X. The two MS value are defined as

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{regression}}} \text{ with } df = 1 \quad \text{and} \quad MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}} \text{ with } df = n - 2$$

The F-ratio is

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} \quad \text{with } df = 1, n - 2 \tag{16.13}$$

The complete analysis of *SS* and degrees of freedom is diagrammed in Figure 16.6. The analysis of regression procedure is demonstrated in the following example, using the same data that we used in Examples 16.1, 16.2, and 16.3.

EXAMPLE 16.4

The data consist of $n = 8$ pairs of scores with a correlation of $r = 0.75$ and $SS_Y = 64$. The null hypothesis either states that there is no relationship between X and Y in the population, or that the regression equation does not account for a significant portion of the variance for the Y scores.

The F-ratio for the analysis of regression has $df = 2, n - 2$. For these data, $df = 1, 6$. With $\alpha = .05$, the critical value is 5.99.

As noted in the previous section, the *SS* for the Y scores can be separated into two components: the predicted portion corresponding to r^2 and the unpredicted, or residual, portion corresponding to $(1-r^2)$. With $r = 0.75$, we obtain $r^2 = 0.5625$ and

$$\text{predicted variability} = SS_{\text{regression}} = 0.5625(64) = 36$$

$$\text{unpredicted variability} = SS_{\text{residual}} = (1 - 0.5625)(64) = 0.4375(64) = 28$$

Using these *SS* values and the corresponding *df* values, we calculate a variance, or *MS*, for each component. For these data the *MS* values are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{residual}}} = \frac{36}{1} = 36$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}} = \frac{28}{6} = 4.67$$

Finally, the F-ratio for evaluating the significance of the regression equation is

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{36.00}{4.67} = 7.71$$

The F-ratio is in the critical region, so we reject the null hypothesis and conclude that the regression equation does account for a significant portion of the variance for the Y scores. The complete analysis of regression is summarized in Table 16.1, which is a common format for computer printouts of regression analysis.

FIGURE 16.6

The partitioning of *SS* and *df* for analysis of regression. The variability in the original Y scores (both SS_Y and df_Y) is partitioned into two components: (1) the variability that is explained by the regression equation, and (2) the residual variability.

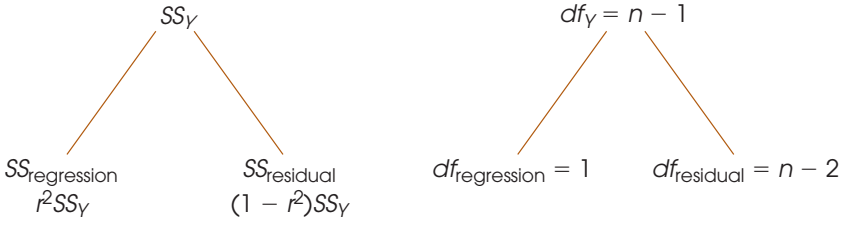


TABLE 16.1

A summary table showing the results of the analysis of regression in Example 16.4.

Source	*SS*	*df*	*MS*	*F*
Regression	36	1	36.60	7.71
Residual	28	6	4.67	
Total	64	7		

SIGNIFICANCE OF REGRESSION AND SIGNIFICANCE OF THE CORRELATION

As noted earlier, in situation with a single X variable and a single Y variable, testing the significance of the regression equation is equivalent to testing the significance of the Pearson correlation. Therefore, whenever the correlation between two variables is significant, you can conclude that the regression equation is also significant. Similarly, if a correlation is not significant, then the regression equation is also not significant. For the data in Example 16.3, we concluded that the regression equation is significant. This conclusion is perfectly consistent with the corresponding test for the significance of the Pearson correlation. For these data, the Pearson correlation is $r = 0.75$ with $n = 8$. Checking Table B.6 in Appendix B, you should find a critical value of 0.707. Our correlation exceeds this criterion, so we conclude that the correlation is also significant. In fact, the critical values listed in Table B.6 were developed using the F-ratio (Equation 16.13) from analysis of regression.

LEARNING CHECK

1. A set of $n = 18$ pairs of scores produces a Pearson correlation of $r = 0.60$ with $SS_Y = 100$. Find $SS_{\text{regression}}$ and SS_{residual} and compute the F-ratio to evaluate the significance of the regression equation of predicting Y.

ANSWER

1. $SS_{\text{regression}} = 36$ with df $= 1$. $SS_{\text{residual}} = 64$ with $df = 16$. $F = 9.00$. With $df = 1, 16$, the F-ratio is significant with either $\alpha = .05$ or $\alpha = .01$.

16.3 INTRODUCTION TO MULTIPLE REGRESSION WITH TWO PREDICTOR VARIABLES

Thus far, we have looked at regression in situations in which one variable is being used to predict a second variable. For example, IQ scores can be used to predict academic performance for a group of college students. However, a variable such as academic performance is usually related to a variety of other factors. For example, college GPA is probably related to motivation, self-esteem, SAT score, rank in high school graduating class, parents' highest level of education, and many other variables. In this case, it is possible to combine several predictor variables to obtain a more accurate prediction. For example, IQ predicts some of academic performance, but you can probably get a better prediction if you use IQ and SAT scores together. The process of using several predictor variables to help obtain more accurate predictions is called *multiple regression.*

Although it is possible to combine a large number of predictor variables in a single multiple-regression equation, we limit our discussion to the two-predictor case. There are two practical reasons for this limitation.

1. Multiple regression, even limited to two predictors, can be relatively complex. Although we present equations for the two-predictor case, the calculations are

usually performed by a computer, so there is not much point in developing a set of complex equations when people are going to use a computer instead.

2. Usually, different predictor variables are related to each other, which means that they are often measuring and predicting the same thing. Because the variables may overlap with each other, adding another predictor variable to a regression equation does not always add to the accuracy of prediction. This situation is shown in Figure 16.7. In the figure, IQ overlaps with academic performance, which means that part of academic performance can be predicted by IQ. In this example, IQ overlaps (predicts) 40% of the variance in academic performance (combine sections a and b in the figure). The figure also shows that SAT scores overlap with academic performance, which means that part of academic performance can be predicted by knowing SAT scores. Specifically, SAT scores overlap, or predict, 30% of the variance (combine sections b and c). Thus, using both IQ and SAT scores to predict academic performance should produce better predictions than would be obtained from IQ alone. However, there is also a lot of overlap between SAT scores and IQ. In particular, much of the prediction from SAT scores overlaps with the prediction from IQ (section b). As a result, adding SAT scores as a second predictor only adds a small amount to the variance already predicted by IQ (section c). Because variables tend to overlap in this way, adding new variables beyond the first one or two predictors often does not add significantly to the quality of the prediction.

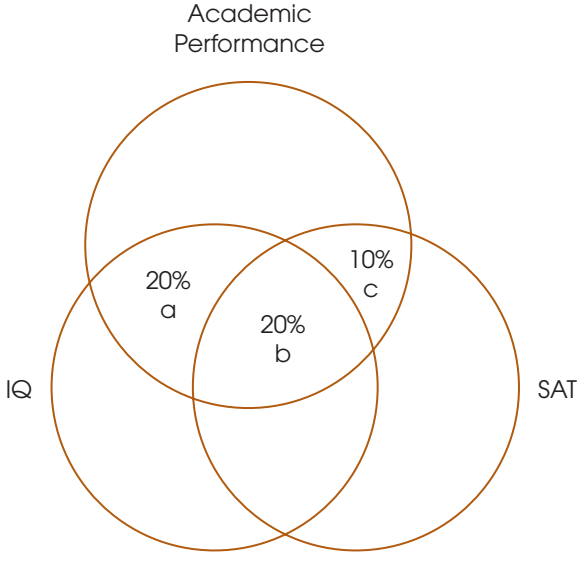
REGRESSION EQUATIONS WITH TWO PREDICTORS

We identify the two predictor variables as X_1 and X_2. The variable we are trying to predict is identified as Y. Using this notation, the general form of the multiple regression equation with two predictors is

$$\hat{Y} = b_1X_1 + b_2X_2 + a \tag{16.14}$$

FIGURE 16.7

Predicting the variance in academic performance from IQ and SAT scores. The overlap between IQ and academic performance indicates that 40% of the variance in academic performance can be predicted from IQ scores. Similarly, 30% of the variance in academic performance can be predicted from SAT scores. However, IQ and SAT also overlap, so that SAT scores contribute an additional predication of only 10% beyond what is already predicted by IQ.



If all three variables, X_1, X_2, and Y, have been standardized by transformation into z-scores, then the standardized form of the multiple regression equation predicts the z-score for each Y value. The standardized form is

$$\hat{z}_Y = (\text{beta}_1)z_{X1} + (\text{beta}_2)z_{X2} \tag{16.15}$$

Researchers rarely transform raw X and Y scores into z-scores before finding a regression equation, however, the beta values are meaningful and are usually reported by computer programs conducting multiple regression. We return to the discussion of beta values later in this section.

The goal of the multiple-regression equation is to produce the most accurate estimated values for Y. As with the single-predictor regression, this goal is accomplished with a least-squared solution. First, we define "error" as the difference between the predicted Y value from the regression equation and the actual Y value for each individual. Each error is then squared to produce uniformly positive values, and then we add the squared errors. Finally, we calculate values for b_1, b_2, and a that produce the smallest possible sum of squared errors. The derivation of the final values is beyond the scope of this text, but the final equations are as follows:

$$b_1 = \frac{(SP_{X1Y})(SS_{X2}) - (SP_{X1X2})(SP_{X2Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} \tag{16.16}$$

$$b_2 = \frac{(SP_{X2Y})(SS_{X1}) - (SP_{X1X2})(SP_{X1Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} \tag{16.17}$$

$$a = M_Y - b_1M_{X1} - b_2M_{X2} \tag{16.18}$$

In these equations, you should recognize the following SS and SP values:

SS_{X1} is the sum of squared deviations for X_1

SS_{X2} is the sum of squared deviations for X_2

SP_{X1Y} is the sum of products of deviations for X_1 and Y

SP_{X2Y} is the sum of products of deviations for X_2 and Y

SP_{X1X2} is the sum of products of deviations for X_1 and X_2

Note: More detailed information about the calculation of SS is presented in Chapter 4 (pp. 111–112) and information concerning SP is in Chapter 15 (pp. 515–516). The following example demonstrates multiple regression with two predictor variables.

EXAMPLE 16.5

We use the data in Table 16.2 to demonstrate multiple regression. Note that each individual has a Y score and two X scores that are used as predictor variables. Also note that we have already computed the SS values for Y and for both of the X scores, as well as all of the SP values. These values are used to compute the coefficients, b_1 and b_2, and the constant, a, for the regression equation.

$$\hat{Y} = b_1X_1 + b_2X_2 + a$$

TABLE 16.2

Hypothetical data consisting of three scores for each person. Two of the scores, X_1 and X_2, are used to predict the Y score for each individual.

Person	Y	X_1	X_2	
A	11	4	10	$SP_{X1Y} = 54$
B	5	5	6	$SP_{X2Y} = 47$
C	7	3	7	$SP_{X1X2} = 42$
D	3	2	4	
E	4	1	3	
F	12	7	5	
G	10	8	8	
H	4	2	4	
I	8	7	10	
J	6	1	3	
	$M_Y = 7$	$M_{X1} = 4$	$M_{X2} = 6$	
	$SS_Y = 90$	$SS_{X1} = 62$	$SS_{X2} = 64$	

$$b_1 = \frac{(SP_{X1Y})(SS_{X2}) - (SP_{X1X2})(SP_{X2Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(54)(64) - (42)(47)}{(62)(64) - (42)^2} = 0.672$$

$$b_2 = \frac{(SP_{X2Y})(SS_{X1}) - (SP_{X1X2})(SP_{X1Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(47)(62) - (42)(54)}{(62)(64) - (42)^2} = 0.293$$

$$a = M_Y - b_1M_{X1} - b_2M_{X2} = 7 - 0.672(4) - 0.293(6) = 7 - 2.688 - 1.758 = 2.554$$

Thus, the final regression equation is,

$$\hat{Y} = 0.672X_1 + 0.293X_2 + 2.554$$

Example 16.5 also demonstrates that multiple regression can be a tedious process. As a result, multiple regression is usually conducted on a computer. To demonstrate this process, we used the SPSS computer program to perform a multiple regression on the data in Table 16.2 and the output from the program is shown in Figure 16.8. At this time, focus on the Coefficients Table at the bottom of the printout. The values in the first column of Unstandardized Coefficients include the constant, b_1 and b_2 for the regression equation. We discuss other portions of the SPSS output later in this chapter.

LEARNING CHECK

1. A researcher computes a multiple-regression equation for predicting annual income for 40-year-old men based on their level of education (X_1 = number of years after high school) and their social skills (X_2 = score from a self-report questionnaire). The regression equation is $\hat{Y} = 8.3X_1 + 2.1X_2 + 3.5$ and predicts income in thousands of dollars. Two individuals are selected from the sample. One has $X_1 = 0$ and $X_2 = 16$; the other has $X_1 = 3$ and $X_2 = 12$. Compute the predicted income for each.

ANSWER

1. The first man has a predicted income of $\hat{Y} = 37.1$ thousand dollars and the second has $\hat{Y} = 53.6$ thousand dollars.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.746[a]	.557	.430	2.38788

a. Predictors: (Constant), VAR00003, VAR00002

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	50.086	2	25.043	4.392	.058[a]
	Residual	39.914	7	5.702		
	Total	90.000	9			

a. Predictors: (Constant), VAR00003, VAR00002
b. Dependent Variable: VAR00001

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.552	1.944		1.313	.231
	VAR00002	.672	.407	.558	1.652	.142
	VAR00003	.293	.401	.247	.732	.488

a. Dependent Variable: VAR00001

FIGURE 16.8
The SPSS output for the multiple regression in Example 16.5.

R^2 AND RESIDUAL VARIANCE

In the same way that we computed an r^2 value to measure the percentage of variance accounted for with the single-predictor regression, it is possible to compute a corresponding percentage for multiple regression. For a multiple-regression equation, this percentage is identified by the symbol R^2. The value of R^2 describes the proportion of the total variability of the Y scores that is accounted for by the regression equation. In symbols,

$$R^2 = \frac{SS_{\text{regression}}}{SS_Y} \quad \text{or} \quad SS_{\text{regression}} = R^2 SS_Y$$

For a regression with two predictor variables, R^2 can be computed directly from the regression equation as follows:

$$R^2 = \frac{b_1 SP_{X1Y} + b_2 SP_{X2Y}}{SS_Y} \tag{16.19}$$

For the data in Table 16.2, we obtain a value of

$$R^2 = \frac{0.672(54) + 0.293(47)}{90} = \frac{50.059}{90} = 0.5562 \text{ (or } 55.62\%)$$

In the computer printout in Figure 16.8, the value of R^2 is reported in the Model Summary table at the top.

Thus, 55.6% of the variance for the Y scores can be predicted by the regression equation. For the data in Table 16.2, SSY = 90, so the predicted portion of the variability is

$$SS_{\text{regression}} = R^2 SS_Y = 0.5562(90) = 50.06$$

The unpredicted, or residual, variance is determined by $1 - R^2$. For the data in Table 16.2, this is

$$SS_{\text{residual}} = (1 - R^2)SS_Y = 0.4438(90) = 39.94$$

COMPUTING R^2 AND $1 - R^2$ FROM THE RESIDUALS

The value of R^2 and $1 - R^2$ can also be obtained by computing the residual, or difference between the predicted Y and the actual Y for each individual, then computing the sum of the squared residuals. The resulting value is SS_{residual} and measures the unpredicted portion of the variability of Y, which is equal to $(1 - R^2)SS_Y$. For the data in Table 16.2, we first use the multiple-regression equation to compute the predicted value of Y for each individual. The process of finding and squaring each residual is shown in Table 16.3.

Note that the sum of the squared residuals, the unpredicted portion of SS_Y, is 39.960. This value corresponds to 44.4% of the variability for the Y scores:

$$\frac{SS_{\text{residual}}}{SS_Y} = \frac{39.96}{90} = 0.444 \text{ (or } 44.4\%)$$

Because the unpredicted portion of the variability is $1 - R^2 = 44.4\%$, we conclude that the predicted portion is $R^2 = 55.6\%$. Note that this answer is within rounding error of $R^2 = 55.62\%$ that we obtained from equation 16.19.

TABLE 16.3

The predicted Y values and the residuals for the data in Table 16.2. The predicted Y values were obtained using the values of X_1 and X_2 in the multiple-regression equation for each individual.

Actual Y	Predicted Y ($\hat{Y}$)	Residual ($Y - \hat{Y}$)	Squared Residual $(Y - \hat{Y})^2$
11	8.17	2.83	8.010
5	7.67	−2.67	7.129
7	6.62	0.38	0.144
3	5.07	−2.07	4.285
4	4.10	−0.10	0.010
12	8.72	3.28	10.758
10	10.27	−0.27	0.073
4	5.07	−1.07	1.145
8	10.19	−2.19	4.796
6	4.10	1.90	3.610
			39.960 = $SS_{\text{residuals}}$

THE STANDARD ERROR OF ESTIMATE

On page 567, we defined the standard error of estimate for a linear regression equation as the standard distance between the regression line and the actual data points. In more general terms, the standard error of estimate can be defined as the standard distance between the predicted Y values (from the regression equation) and the actual Y values (in the data). The more general definition applies equally well to both linear and multiple regression.

To find the standard error of estimate for either linear regression or multiple regression, we begin with SS_{residual}. For linear regression with one predictor, $SS_{\text{residual}} = (1 - r^2)SS_Y$ and has $df = n - 2$. For multiple regression with two predictors, $SS_{\text{residual}} = (1 - R^2)SS_Y$ and has $df = n - 3$. In each case, we can use the SS and df values to compute a variance or MS_{residual}.

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df}$$

The variance, or MS value, is a measure of the average squared distance between the actual Y values and the predicted Y values. By simply taking the square root, we obtain a measure of standard deviation or standard distance. This standard distance for the residuals is the standard error of estimate. Thus, for both linear regression and multiple regression

$$\text{the standard error of estimate} = \sqrt{MS_{\text{residual}}}$$

In the computer printout in Figure 16.8, the standard error of estimate is reported in the Model Summary table at the top.

For either linear or multiple regression, you do not expect the predictions from the regression equation to be perfect. In general, there is some discrepancy between the predicted values of Y and the actual values. The standard error of estimate provides a measure of how much discrepancy, on average, there is between the $\hat{Y}$ values and the actual Y values.

TESTING THE SIGNIFICANCE OF THE MULTIPLE REGRESSION EQUATION: ANALYSIS OF REGRESSION

Just as we did with the single-predictor equation, we can evaluate the significance of a multiple-regression equation by computing an F-ratio to determine whether the equation predicts a significant portion of the variance for the Y scores. The total variability of the Y scores is partitioned into two components, $SS_{\text{regression}}$ and SS_{residual}. With two predictor variables, $SS_{\text{regression}}$ has $df = 2$, and SS_{residual} has $df = n - 3$. Therefore, the two MS values for the F-ratio are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{2} \tag{16.20}$$

and

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{n - 3} \tag{16.21}$$

Because of rounding error, the value we obtain for SS_{residual} is slightly different from the value in Table 16.3.

The data for the $n = 10$ people in Table 16.2 have produced $R^2 = 0.5562$ (or 55.62%) and $SS_Y = 90$. Thus,

$$SS_{\text{regression}} = R^2SS_Y = 0.556(90) = 50.06$$

$$SS_{\text{residual}} = (1 - R^2)SS_Y = 0.4438(90) = 39.94$$

Therefore, $MS_{\text{regession}} = \dfrac{50.06}{2} = 25.03$ and $MS_{\text{residual}} = \dfrac{39.94}{7} = 5.71$

and $$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{25.03}{5.71} = 4.38$$

With $df = 2, 7$, this F-ratio is not significant with $\alpha = .05$, so we cannot conclude that the regression equation accounts for a significant portion of the variance for the Y scores.

The analysis of regression is summarized in the following table, which is a common component of the output from most computer versions of multiple regression. In the computer printout in Figure 16.8, this summary table is reported in the ANOVA table in the center.

Source	*SS*	*df*	*MS*	*F*
Regression	50.06	2	25.03	4.38
Residual	39.94	7	5.71	
Total	90.00	9		

LEARNING CHECK

1. Data from a sample of $n = 15$ individuals are used to compute a multiple-regression equation with two predictor variables. The equation has $R^2 = 0.20$ and $SS_Y = 150$.

a. Find $SS_{residual}$ and compute the standard error of estimate for the regression equation.

b. Find $SS_{regression}$ and compute the F-ratio to evaluate the significance of the regression equation.

ANSWER

1. a. $SS_{residual} = 120$. The standard error of estimate is $\sqrt{10} = 3.16$.

b. $SS_{regression} = 30$ with $df = 2$. $SS_{residual} = 120$ with $df = 12$. $F = 1.50$. With $df = 2, 12$, the F-ratio is not significant.

16.4 EVALUATING THE CONTRIBUTION OF EACH PREDICTOR VARIABLE

In addition to evaluating the overall significance of the multiple-regression equation, researchers are often interested in the relative contribution of each of the two predictor variables. Is one of the predictors responsible for more of the prediction than the other? Unfortunately, the b values in the regression equation are influenced by a variety of other factors and do not address this issue. If b_1 is larger than b_2, it does not necessarily mean that X_1 is a better predictor than X_2. However, in the standardized form of the regression equation, the relative size of the beta values is an indication of the relative contribution of the two variables. For the data in Table 16.3, the standardized regression equation is

$$\hat{z}_Y = (\text{beta}_1)z_{X1} + (\text{beta}_2)z_{X2}$$
$$= 0.558z_{X1} + 0.247z_{X2}$$

For the SPSS printout in Figure 16.8, the beta values are shown in the Coefficients table.

In this case, the larger beta value for the X_1 predictor indicates that X_1 predicts more of the variance than does X_2. The signs of the beta values are also meaningful. In this example, both betas are positive, indicating the both X_1 and X_2 are positively related to Y.

Beyond judging the relative contribution for each of the predictor variables, it also is possible to evaluate the significance of each contribution. For example, does variable X_2 make a significant contribution beyond what is already predicted by variable X_1? The null hypothesis states that the multiple-regression equation (using X_2 in addition to X_1) is not any better than the simple regression equation using X_1 as a single predictor variable. An alternative view of the null hypothesis is that the b_2 (or beta$_2$) value in the equation is not significantly different from zero. To test this hypothesis, we first determine how much more variance is predicted using X_1 and X_2 together than is predicted using X_1 alone.

Earlier we found that the multiple regression equation with both X_1 and X_2 predicted $R^2 = 55.62\%$ of the variance for the Y scores. To determine how much is predicted by X_1 alone, we begin with the correlation between X_1 and Y, which is

$$r = \frac{SP_{X1Y}}{\sqrt{(SS_{X1})(SS_Y)}} = \frac{54}{\sqrt{(62)(90)}} = \frac{54}{74.70} = 0.7229$$

Squaring the correlation produces $r^2 = (0.7229)^2 = 0.5226$ or 52.26%. This means that the relationship with X_1 predicts 52.26% of the variance for the Y scores. Therefore, the additional contribution made by adding X_2 to the regression equation can be computed as

$$(\%\text{ with both } X_1 \text{ and } X_2) - (\%\text{ with } X_1 \text{ alone})$$

$$= 55.62\% - 52.26\%$$

$$= 3.36\%$$

Because $SS_Y = 90$, the additional variability from adding X_2 as a predictor amounts to

$$SS_{\text{additional}} = 3.36\% \text{ of } 90 = 0.0336(90) = 3.024$$

This SS value has $df = 1$, and can be used to compute an F-ratio evaluating the significance of the contribution of X_2. First,

$$MS_{\text{additional}} = \frac{SS_{\text{additional}}}{1} = \frac{3.024}{1} = 3.024$$

This MS value is evaluated by computing an F-ratio with the MS_{residual} value from the multiple regression as the denominator. (*Note:* This is the same denominator that was used in the F-ratio to evaluate the significance of the multiple-regression equation.) For these data, we obtain

$$F = \frac{MS_{\text{additional}}}{MS_{\text{residual}}} = \frac{3.024}{5.71} = 0.5296$$

With $df = 1, 7$, this F-ratio is not significant. Therefore, we conclude that adding X_2 to the regression equation does not significantly improve the prediction compared to using X_1 as a single predictor. The computer printout shown in Figure 16.8 reports a t statistic instead of an F-ratio to evaluate the contribution for each predictor variable. Each t value is simply the square root of the F-ratio and is reported in the right-hand side of the Coefficients table. Variable X_2, for example, is reported as VAR00003 in the table and has $t = 0.732$, which is within rounding error of the F-ratio we obtained; $\sqrt{F} = \sqrt{0.5296} = 0.728$.

MULTIPLE REGRESSION AND PARTIAL CORRELATIONS

In Chapter 15 we introduced *partial correlation* as a technique for measuring the relationship between two variables while eliminating the influence of a third variable. At that time, we noted that partial correlations serve two general purposes:

1. A partial correlation can demonstrate that an apparent relationship between two variables is actually caused by a third variable. Thus, there is no direct relationship between the original two variables.
2. Partial correlation can demonstrate that there is a relationship between two variables even after a third variable is controlled. Thus, there really is a relationship between the original two variables that is not being caused by a third variable.

Multiple regression provides an alternative procedure for accomplishing both of these goals. Specifically, the regression analysis evaluates the contribution of each predictor variable after the influence of the other predictor has been considered. Thus, you can determine whether each predictor variable contributes to the relationship by itself or simply duplicates the contribution already made by another variable.

SUMMARY

1. When there is a general linear relationship between two variables, X and Y, it is possible to construct a linear equation that allows you to predict the Y value corresponding to any known value of X.

$$\text{predicted } Y \text{ value} = \hat{Y} = bX + a$$

The technique for determining this equation is called regression. By using a *least-squares* method to minimize the error between the predicted Y values and the actual Y values, the best-fitting line is achieved when the linear equation has

$$b = \frac{SP}{SS_X} = r\frac{s_Y}{s_X} \qquad \text{and} \qquad a = M_Y - bM_X$$

2. The linear equation generated by regression (called the regression equation) can be used to compute a predicted Y value for any value of X. However, the prediction is not perfect, so for each Y value, there is a predicted portion and an unpredicted, or residual, portion. Overall, the predicted portion of the Y score variability is measured by r^2, and the residual portion is measured by $1 - r^2$.

$$\text{predicted variability} = SS_{\text{regression}} = r^2 SS_Y$$

$$\text{unpredicted variability} = SS_{\text{residual}} = (1 - r^2)SS_Y$$

3. The residual variability can be used to compute the standard error of estimate, which provides a measure of the standard distance (or error) between the predicted Y values on the line and the actual data points. The standard error of estimate is computed by

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{n-2}} = \sqrt{MS_{\text{residual}}}$$

4. It is also possible to compute an F-ratio to evaluate the significance of the regression equation. The process is called analysis of regression and determines whether the equation predicts a significant portion of the variance for the Y scores. First a variance, or MS, value is computed for the predicted variability and the residual variability,

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{regression}}} \qquad MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}}$$

where $df_{\text{regression}} = 1$ and $df_{\text{residual}} = n - 2$. Next, an F-ratio is computed to evaluate the significance of the regression equation.

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} \qquad \text{with } df = 1, n - 2$$

5. Multiple regression involves finding a regression equation with more than one predictor variable. With two predictors (X_1 and X_2), the equation becomes

$$\hat{Y} = b_1X_1 + b_2X_2 + a$$

with the values for b_1, b_2, and a computed using equations 16.16, 16.17, and 16.18.

6. For multiple regression, the value of R^2 describes the proportion of the total variability of the Y scores that is accounted for by the regression equation. With two predictor variables,

$$R^2 = \frac{b_1 SP_{X1Y} + b_2 SP_{X2Y}}{SS_Y}$$

Predicted variability = $SS_{\text{regression}} = R^2 SS_Y$.
Unpredicted variability = $SS_{\text{residual}} = (1 - R^2) SS_Y$.

7. The residual variability for the multiple-regression equation can be used to compute a standard error of estimate, which provides a measure of the standard distance (or error) between the predicted Y values from the equation and the actual data points. For multiple regression with two predictors, the standard error of estimate is computed by

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{n - 3}}$$
$$= \sqrt{MS_{\text{residual}}}$$

8. Evaluating the significance of the two-predictor multiple-regression equation involves computing an F-ratio that divides the $MS_{\text{regression}}$ (with $df = 2$) by the MS_{residual} (with $df = n - 3$). A significant F-ratio indicates that the regression equation accounts for a significant portion of the variance for the Y scores.

9. An F-ratio can also be used to determine whether a second predictor variable (X_2) significantly improves the prediction beyond what was already predicted by X_1. The numerator of the F-ratio measures the additional SS that is predicted by adding X_2 as a second predictor.

$$SS_{\text{additional}} = SS_{\text{regression with } X1 \text{ and } X2} - SS_{\text{regression with } X1 \text{ alone}}$$

This SS value has $df = 1$. The denominator of the F-ratio is the MS_{residual} from the two-predictor regression equation.

KEY TERMS

linear relationship (559)
linear equation (559)
slope (559)
Y-intercept (559)
regression (561)
regression line (561)
least-squared-error solution (562)
regression equation for *Y* (563)
standard error of estimate (567)
predicted variability ($SS_{\text{regression}}$) (568)
unpredicted variability (SS_{residual}) (568)
analysis of regression (570)
multiple regression (572)
partial correlation (581)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 16 on the book companion website. The website also provides access to a workshop entitled *Correlation* that includes information on regression.

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform the **Linear Regression** and **Multiple Regression** presented in this chapter.

Data Entry

With one predictor variable (X), you enter the X values in one column and the Y values in a second column of the SPSS data editor. With two predictors (X_1 and X_2), enter the X_1 values in one column, X_2 in a second column, and Y in a third column.

Data Analysis

1. Click **Analyze** on the tool bar, select **Regression,** and click on **Linear.**
2. In the left-hand box, highlight the column label for the Y values, then click the arrow to move the column label into the **Dependent Variable** box.
3. For one predictor variable, highlight the column label for the X values and click the arrow to move it into the **Independent Variable(s)** box. For two predictor variables, highlight the X_1 and X_2 column labels, one at a time, and click the arrow to move them into the **Independent Variable(s)** box.
4. Click **OK.**

SPSS Output

We used SPSS to perform multiple regression for the data in Example 16.4 and the output is shown in Figure 16.8 (p. 576). The Model Summary table presents the values for R, R^2, and the standard error of estimate. (*Note:* For a single predictor, R is simply the Pearson correlation between X and Y.) The ANOVA table presents the analysis of regression evaluating the significance of the regression equation, including the F-ratio and the level of significance (the p value or alpha level for the test). The **Coefficients** table summarizes both the unstandardized and the standardized coefficients for the regression equation. For one predictor, the table shows the values for the constant (a) and the coefficient (b). For two predictors, the table shows the constant (a) and the two coefficients (b_1 and b_2). The standardized coefficients are the beta values. For one predictor, beta is simply the Pearson correlation between X and Y. Finally, the table uses a t statistic to evaluate the significance of each predictor variable. For one predictor variable, this is identical to the significance of the regression equation and you should find that t is equal to the square root of the F-ratio from the analysis of regression. For two predictor variables, the t values measure the significance of the contribution of each variable beyond what is already predicted by the other variable.

FOCUS ON PROBLEM SOLVING

1. A basic understanding of the Pearson correlation, including the calculation of *SP* and *SS* values, is critical for understanding and computing regression equations.
2. You can calculate $SS_{residual}$ directly by finding the residual (the difference between the actual *Y* and the predicted *Y* for each individual), squaring the residuals, and adding the squared values. However, it usually is much easier to compute r^2 (or R^2) and then find $SS_{residual} = (1 - r^2)SS_Y$.
3. The *F*-ratio for analysis of regression is usually calculated using the actual $SS_{regression}$ and $SS_{residual}$. However, you can simply use r^2 (or R^2) in place of $SS_{regression}$ and you can use $1 - r^2$ or $(1 - R^2)$ in place of $SS_{residual}$. *Note:* You must still use the correct *df* value for the numerator and the denominator.

DEMONSTRATION 16.1

LINEAR REGRESSION

The following data are used to demonstrate the process of linear regression. The scores and summary statistics are as follows:

Person	*X*	*Y*
A	0	4
B	2	1
C	8	10
D	6	9
E	4	6

$M_X = 4$ with $SS_X = 40$
$M_Y = 6$ with $SS_Y = 54$
$SP = 40$

These data produce a Pearson correlation of $r = 0.861$.

STEP 1 **Compute the values for the regression equation.** The general form of the regression equation is

$$\hat{Y} = bX + a \qquad \text{where } b = \frac{SP}{SS_X} \qquad \text{and} \qquad a = M_Y - bM_X$$

$$\text{For these data, } b = \frac{40}{40} = 1.00 \qquad \text{and} \qquad a = 6 - 1(4) = +2.00$$

Thus, the regression equation is $\hat{Y} = (1)X + 2.00$ or simply, $\hat{Y} = X + 2$.

STEP 2 **Evaluate the significance of the regression equation.** The null hypothesis states that the regression equation does not predict a significant portion of the variance for the *Y* scores. To conduct the test, the total variability for the *Y* scores, $SS_Y = 54$, is partitioned into the portion predicted by the regression equation and the residual portion.

$$SS_{regression} = r^2(SS_Y) = 0.741(54) = 40.01 \text{ with } df = 1$$

$$SS_{residual} = (1 - r^2)(SS_Y) = 0.259(54) = 13.99 \text{ with } df = n - 2 = 3$$

The two *MS* values (variances) for the *F*-ratio are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df} = \frac{40.01}{1} = 40.01$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df} = \frac{13.99}{3} = 4.66$$

And the *F*-ratio is

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{40.01}{4.66} = 8.59$$

With $df = 1, 3$ and $\alpha = .05$, the critical value for the *F*-ratio is 10.13. Therefore, we fail to reject the null hypothesis and conclude that the regression equation does not predict a significant portion of the variance for the *Y* scores.

DEMONSTRATION 16.2

MULTIPLE REGRESSION

The following data are used to demonstrate the process of multiple regression. Note that there are two predictor variables, X_1 and X_2, that are used to compute a predicted *Y* score for each individual.

Person	X_1	X_2	Y
A	0	5	2
B	3	1	4
C	5	2	7
D	6	0	9
E	8	4	5
F	2	6	3
	$M_{X1} = 4$	$M_{X2} = 3$	$M_Y = 5$
	$SS_{X1} = 42$	$SS_{X2} = 28$	$SS_Y = 34$
	$SP_{X1Y} = 27$	$SP_{X2Y} = -24$	$SP_{X1X2} = -15$

STEP 1 **Compute the values for the multiple regression equation.** The general form of the multiple-regression equation is

$$\hat{Y} = b_1X_1 + b_2X_2 + a$$

The values for the multiple regression equation are

$$b_1 = \frac{(SP_{X1Y})(SS_{X2}) - (SP_{X1X2})(SP_{X2Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(27)(28) - (-15)(-24)}{(42)(28) - (-15)^2} = 0.416$$

$$b_2 = \frac{(SP_{X2Y})(SS_{X1}) - (SP_{X1X2})(SP_{X1Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(-24)(42) - (-15)(27)}{(42)(28) - (-15)^2} = -0.634$$

$$a = M_Y - b_1 M_{X1} - b_2 M_{X2} = 5 - 0.416(4) - -0.634(3) = 5 - 1.664 + 1.902 = 5.238$$

The multiple-regression equation is

$$\hat{Y} = 0.416X_1 - 0.634X_2 + 5.238$$

STEP 2 **Evaluate the significance of the regression equation.** The null hypothesis states that the regression equation does not predict a significant portion of the variance for the Y scores. To conduct the test, the total variability for the Y scores, $SS_Y = 34$, is partitioned into the portion predicted by the regression equation and the residual portion. To find each portion, we must first compute the value of R^2.

$$R^2 = \frac{b_1 SP_{X1Y} + b_2 SP_{X2Y}}{SS_Y}$$

$$\frac{(0.416)(27) + (-0.634)(-24)}{34} = 0.778 \text{ (or } 77.8\%)$$

Then, the two components for the F-ratio are

$$SS_{\text{regression}} = R^2(SS_Y) = 0.778(34) = 26.45 \text{ with } df = 2$$

$$SS_{\text{residual}} = (1 - R^2)(SS_Y) = 0.222(34) = 7.55 \text{ with } df = n - 3 = 3$$

The two MS values (variances) and the F-ratio are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df} = \frac{26.45}{2} = 13.23$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df} = \frac{7.55}{3} = 2.52$$

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{13.23}{2.52} = 5.25$$

with $df = 2, 3$, the F-ratio is not significant.

PROBLEMS

1. Sketch a graph showing the line for the equation $Y = -2X + 4$. On the same graph, show the line for $Y = X - 4$.

2. The regression equation is intended to be the "best fitting" straight line for a set of data. What is the criterion for "best fitting"?

3. A set of $n = 20$ pairs of scores (X and Y values) has $SS_X = 16$, $SS_Y = 100$, and $SP = 32$. If the mean for the X values is $M_X = 6$ and the mean for the Y values is $M_Y = 20$.

a. Calculate the Pearson correlation for the scores.

b. Find the regression equation for predicting Y from the X values.

4. A set of $n = 25$ pairs of scores (X and Y values) produces a regression equation of $\hat{Y} = 3X - 2$. Find the predicted Y value for each of the following X scores: 0, 1, 3, -2.

5. Briefly explain what is measured by the standard error of estimate.

6. In general, how is the magnitude of the standard error of estimate related to the value of the correlation?

7. For the following set of data, find the linear regression equation for predicting Y from X:

X	Y
7	6
9	6
6	3
12	5
9	6
5	4

8. For the following data:
- **a.** Find the regression equation for predicting Y from X.
- **b.** Calculate the Pearson correlation for these data. Use r^2 and SS_Y to compute $SS_{residual}$ and the standard error of estimate for the equation.

X	Y
1	2
4	7
3	5
2	1
5	14
3	7

9. Does the regression equation from problem 8 account for a significant portion of the variance in the Y scores? Use $\alpha = .05$ to evaluate the F-ratio.

10. For the following scores,

X	Y
3	6
6	1
3	4
3	3
5	1

- **a.** Find the regression equation for predicting Y from X.
- **b.** Calculate the predicted Y value for each X.

11. Problem 12 in Chapter 15 examined the relationship between weight and income for a sample of $n = 10$ women. Weights were classified in five categories and had a mean of $M = 3$ with $SS = 20$. Income, measured in thousands, had a mean score of $M = 66$ with $SS = 7430$, and $SP = -359$.
- **a.** Find the regression equation for predicting income from weight. (Identify the income scores as X values and the weight scores as Y values.)
- **b.** What percentage of the variance in the income is accounted for by the regression equation? (Compute the correlation, r, then find r^2.)
- **c.** Does the regression equation account for a significant portion of the variance in income? Use $\alpha = .05$ to evaluate the F-ratio.

12. A professor obtains SAT scores and freshman grade point averages (GPAs) for a group of $n = 15$ college students. The SAT scores have a mean of $M = 580$ with $SS = 22{,}400$, and the GPAs have a mean of 3.10 with $SS = 1.26$, and $SP = 84$.
- **a.** Find the regression equation for predicting GPA from SAT scores.
- **b.** What percentage of the variance in GPAs is accounted for by the regression equation? (Compute the correlation, r, then find r^2.)
- **c.** Does the regression equation account for a significant portion of the variance in GPA? Use $\alpha = .05$ to evaluate the F-ratio.

13. Problem 14 in Chapter 15 described a study examining the effectiveness of a 7-Minute Screen test for Alzheimer's disease. The study evaluated the relationship between scores from the 7-Minute Screen and scores for the same patients from a set of cognitive exams that are typically used to test for Alzheimer's disease. For a sample of $n = 9$ patients, the scores for the 7-Minute Screen averaged $M = 7$ with $SS = 92$. The cognitive test scores averaged $M = 17$ with $SS = 236$. For these data, $SP = 127$.
- **a.** Find the regression equation for predicting the cognitive scores from the 7-Minute Screen score.
- **b.** What percentage of variance in the cognitive scores is accounted for by the regression equation?
- **c.** Does the regression equation account for a significant portion of the variance in the cognitive scores? Use $\alpha = .05$ to evaluate the F-ratio.

14. There appears to be some evidence suggesting that earlier retirement may lead to memory decline (Rohwedder & Willis, 2010). The researchers gave a memory test to men and women aged 60 to 64 years in several countries that have different retirement ages. For each country, the researchers recorded the average memory score and the percentage of individuals in the 60 to 64 age range who were retired. Note that a higher percentage retired indicates a younger retirement age for that country. The following data are similar to the results from the study. Use the data to find the regression equation for predicting memory scores from the percentage of people aged 60 to 64 who are retired.

Country	% Retired (X)	Memory Score (Y)
Sweden	39	9.3
U.S.A.	48	10.9
England	59	10.7
Germany	70	9.1
Spain	74	6.4
Netherlands	78	9.1
Italy	81	7.2
France	87	7.9
Belgium	88	8.5
Austria	91	9.0

15. The regression equation is computed for a set of $n = 18$ pairs of X and Y values with a correlation of $r = +80$ and $SS_Y = 100$.
- **a.** Find the standard error of estimate for the regression equation.
- **b.** How big would the standard error be if the sample size were $n = 38$?

16. **a.** One set of 20 pairs of scores, X and Y values, produces a correlation of $r = 0.70$. If $SS_Y = 150$, find the standard error of estimate for the regression line.
- **b.** A second set of 20 pairs of X and Y values produces of correlation of $r = 0.30$. If $SS_Y = 150$, find the standard error of estimate for the regression line.

17. **a.** A researcher computes the regression equation for a sample of $n = 25$ pairs of scores, X and Y values. If an analysis of regression is used to test the significance of the equation, what are the df values for the F-ratio?
- **b.** A researcher evaluating the significance of a regression equation obtains an F-ratio with $df = 1, 18$. How many pairs of scores, X and Y values, are in the sample?

18. For the following data:
- **a.** Find the regression equation for predicting Y from X.
- **b.** Use the regression equation to find a predicted Y for each X.
- **c.** Find the difference between the actual Y value and the predicted Y value for each individual, square the differences, and add the squared values to obtain $SS_{residual}$.
- **d.** Calculate the Pearson correlation for these data. Use r^2 and SS_Y to compute $SS_{residual}$ with Equation 16.11. You should obtain the same value as in part c.

X	Y
7	16
5	2
6	1
3	2
4	9

19. A multiple-regression equation with two predictor variables produces $R^2 = .22$.
- **a.** If $SS_Y = 20$ for a sample of $n = 18$ individuals, does the equation predict a significant portion of the variance for the Y scores? Test with $\alpha = .05$.
- **b.** If $SS_Y = 20$ for a sample of $n = 8$ individuals, does the equation predict a significant portion of the variance for the Y scores? Test with $\alpha = .05$.

20. A researcher obtained the following multiple-regression equation using two predictor variables: $\hat{Y} = 0.5X_1 + 4.5X_2 + 9.6$. Given that $SS_Y = 210$, the SP value for X_1 and Y is 40, and the SP value for X_2 and Y is 9, find R^2, the percentage of variance accounted for by the equation.

21. In Chapter 15 (p. 531), we presented an example showing the general relationship among the number of churches, the number of serious crimes, and the population for a set of cities. At that time, we used a partial correlation to evaluate the relationship between churches and crime while controlling population. It is possible to use multiple regression to accomplish essentially the same purpose. For the following data,

Number of Churches (X_1)	Population (X_2)	Number of Crimes (Y)
1	1	4
2	1	1
3	1	2
4	1	3
5	1	5
7	2	8
8	2	11
9	2	9
10	2	7
11	2	10
13	3	15
14	3	14
15	3	16
16	3	17
17	3	13

- **a.** Find the multiple regression equation for predicting the number of crimes using the number of churches and population as predictor variables.
- **b.** Find the value of R^2 for the regression equation.
- **c.** The correlation between the number of crimes and population is $r = 0.961$, which means that $r^2 = .924$ (92.4%) is the proportion of variance in the number of crimes that is predicted by population size. Does adding the number of churches as a second variable in the multiple

regression equation add a significant amount to the prediction? Test with $\alpha = .05$.

22. Problem 11 in Chapter 15 examined the TV-viewing habits of adopted children in relation to their biological parents and their adoptive parents. The data are reproduced as follows. If both the biological and adoptive parents are used to predict the viewing habits of the children in a multiple-regression equation, what percentage of the variance in the children's scores would be accounted for? That is, compute R^2.

Amount of Time Spent Watching TV		
Adopted Children Y	Birth Parents X_1	Adoptive Parents X_2
2	0	1
3	3	4
6	4	2
1	1	0
3	1	0
0	2	3
5	3	2
2	1	3
5	3	3
$SS_Y = 32$	$SS_{X1} = 14$	$SS_{X2} = 16$

$SP_{X1X2} = 8$
$SP_{X1Y} = 15$
$SP_{X2Y} = 3$

23. For the data in problem 22, the correlation between the children's scores and the biological parents' scores is $r = 0.709$. Does adding the adoptive parents' scores as a second predictor significantly improve the ability to predict the children's scores? Use $\alpha = .05$ to evaluate the F-ratio.

24. For the following data, find the multiple-regression equation for predicting Y from X_1 and X_2.

X_1	X_2	Y
1	3	1
2	4	2
3	5	6
6	9	8
4	8	3
2	7	4
$M = 3$	$M = 6$	$M = 4$
$SS_{X1} = 16$	$SS_{X2} = 28$	$SS_Y = 34$

$SP_{X1X2} = 18$
$SP_{X1Y} = 19$
$SP_{X2Y} = 21$

25. A researcher evaluates the significance of a multiple-regression equation and obtains an F-ratio with $df = 2, 36$. How many participants were in the sample?

This page intentionally left blank

CHAPTER

17

The Chi-Square Statistic: Tests for Goodness of Fit and Independence

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Proportions (math review, Appendix A)
- Frequency distributions (Chapter 2)

Preview

Loftus and Palmer (1974) conducted a classic experiment demonstrating how language can influence eyewitness memory. A sample of 150 students watched a film of an automobile accident. After watching the film, the students were separated into three groups. One group was asked, "About how fast were the cars going when they smashed into each other?" Another group received the same question except that the verb was changed to "hit" instead of "smashed into." A third group served as a control and was not asked any question about the speed of the two cars. A week later, the participants returned and were asked if they remembered seeing any broken glass in the accident. (There was no broken glass in the film.) Notice that the researchers are manipulating the form of the initial question and then measuring a yes/no response to a follow-up question 1 week later. Table 17.1 shows the structure of this design represented by a matrix with the independent variable (different groups) determining the rows of the matrix and the two categories for the dependent variable (yes/no) determining the columns. The number in each cell of the matrix is the frequency count showing how many participants are classified in that category. For example, of the 50 students who heard the word *smashed,* there were 16 (32%) who claimed to remember seeing broken glass even though there was none in the film. By comparison, only 7 of the 50 students (14%) who heard the word *hit* said they recalled seeing broken glass. The researchers would like to use these data to support the argument that a witness's "memory" can be influenced by the language used during questioning. If the two cars *smashed* into each other, then there must have been some broken glass.

The Problem: Although the Loftus and Palmer study involves an independent variable (the form of the question) and a dependent variable (memory for broken glass), you should realize that this study is different from any experiment we have considered in the past. Specifically, the Loftus and Palmer study does not produce a numerical score for each participant. Instead, each participant is simply classified into one of two categories (yes or no). The data consist of *frequencies* or *proportions* describing how many individuals are in each category. You should also note that Loftus and Palmer want to use a hypothesis test to evaluate the data. The null hypothesis would state that the form of the question has no effect on the memory of the witness. The hypothesis test would determine whether the sample data provide enough evidence to reject this null hypothesis.

Because there are no numerical scores, it is impossible to compute a mean or a variance for the sample data. Therefore, it is impossible to use any of the familiar hypothesis tests (such as a *t* test or analysis of variance [ANOVA]) to determine whether there is a significant difference between the treatment conditions. What is needed is a new hypothesis testing procedure that can be used with non-numerical data.

The Solution: In this chapter we introduce two hypothesis tests based on the *chi-square* statistic. Unlike earlier tests that require numerical scores (*X* values), the chi-square tests use sample frequencies and proportions to test hypothesis about the corresponding population values.

TABLE 17.1

A frequency distribution table showing the number of participants who answered either yes or no when asked whether they recalled seeing any broken glass 1 week after witnessing an automobile accident. Immediately after the accident, one group was asked how fast the cars were going when they smashed into each other. A second group was asked how fast the cars were going when they hit each other. A third group served as a control and was not asked about the speed of the cars.

Verb Used to Ask about the Speed of the Cars	Response to the Question: Did You See Any Broken Glass? Yes	No
Smashed into	16	34
Hit	7	43
Control (Not Asked)	6	44

17.1 PARAMETRIC AND NONPARAMETRIC STATISTICAL TESTS

All of the statistical tests that we have examined thus far are designed to test hypotheses about specific population parameters. For example, we used *t* tests to assess hypotheses about a population mean (μ) or mean difference ($\mu_1 - \mu_2$). In addition, these tests typically make assumptions about other population parameters. Recall that, for analysis of variance (ANOVA), the population distributions are assumed to be normal and homogeneity of variance is required. Because these tests all concern parameters and require assumptions about parameters, they are called *parametric tests*.

Another general characteristic of parametric tests is that they require a numerical score for each individual in the sample. The scores then are added, squared, averaged, and otherwise manipulated using basic arithmetic. In terms of measurement scales, parametric tests require data from an interval or a ratio scale (see Chapter 1).

Often, researchers are confronted with experimental situations that do not conform to the requirements of parametric tests. In these situations, it may not be appropriate to use a parametric test. Remember that when the assumptions of a test are violated, the test may lead to an erroneous interpretation of the data. Fortunately, there are several hypothesis-testing techniques that provide alternatives to parametric tests. These alternatives are called *nonparametric tests*.

In this chapter, we introduce two commonly used examples of nonparametric tests. Both tests are based on a statistic known as chi-square and both tests use sample data to evaluate hypotheses about the proportions or relationships that exist within populations. Note that the two chi-square tests, like most nonparametric tests, do not state hypotheses in terms of a specific parameter and they make few (if any) assumptions about the population distribution. For the latter reason, nonparametric tests sometimes are called *distribution-free tests*.

One of the most obvious differences between parametric and nonparametric tests is the type of data they use. All of the parametric tests that we have examined so far require numerical scores. For nonparametric tests, on the other hand, the participants are usually just classified into categories such as Democrat and Republican, or High, Medium, and Low IQ. Note that these classifications involve measurement on nominal or ordinal scales, and they do not produce numerical values that can be used to calculate means and variances. Instead, the data for many nonparametric tests are simply frequencies—for example, the number of Democrats and the number of Republicans in a sample of $n = 100$ registered voters.

Occasionally, you have a choice between using a parametric and a nonparametric test. Changing to a nonparametric test usually involves transforming the data from numerical scores to nonnumerical categories. For example, you could start with numerical scores measuring self-esteem and create three categories consisting of high, medium, and low self-esteem. In most situations, the parametric test is preferred because it is more likely to detect a real difference or a real relationship. However, there are situations for which transforming scores into categories might be a better choice.

1. Occasionally, it is simpler to obtain category measurements. For example, it is easier to classify students as high, medium, or low in leadership ability than to obtain a numerical score measuring each student's ability.
2. The original scores may violate some of the basic assumptions that underlie certain statistical procedures. For example, the *t* tests and ANOVA assume that the data come from normal distributions. Also, the independent-measures

tests assume that the different populations all have the same variance (the homogeneity-of-variance assumption). If a researcher suspects that the data do not satisfy these assumptions, it may be safer to transform the scores into categories and use a nonparametric test to evaluate the data.

3. The original scores may have unusually high variance. Variance is a major component of the standard error in the denominator of t statistics and the error term in the denominator of F-ratios. Thus, large variance can greatly reduce the likelihood that these parametric tests will find significant differences. Converting the scores to categories essentially eliminates the variance. For example, all individuals fit into three categories (high, medium, and low), no matter how variable the original scores are.
4. Occasionally, an experiment produces an undetermined, or infinite, score. For example, a rat may show no sign of solving a particular maze after hundreds of trials. This animal has an infinite, or undetermined, score. Although there is no absolute number that can be assigned, you can say that this rat is in the highest category, and then classify the other scores according to their numerical values.

17.2 THE CHI-SQUARE TEST FOR GOODNESS OF FIT

Parameters such as the mean and the standard deviation are the most common way to describe a population, but there are situations in which a researcher has questions about the proportions or relative frequencies for a distribution. For example,

How does the number of women lawyers compare with the number of men in the profession?

Of the two leading brands of cola, which is preferred by most Americans?

In the past 10 years, has there been a significant change in the proportion of college students who declare a business major?

The name of the test comes from the Greek letter χ (chi, pronounced "kye"), which is used to identify the test statistic.

Note that each of the preceding examples asks a question about proportions in the population. In particular, we are not measuring a numerical score for each individual. Instead, the individuals are simply classified into categories and we want to know what proportion of the population is in each category. The *chi-square test for goodness of fit* is specifically designed to answer this type of question. In general terms, this chi-square test uses the proportions obtained for sample data to test hypotheses about the corresponding proportions in the population.

DEFINITION

The **chi-square test for goodness of fit** uses sample data to test hypotheses about the shape or proportions of a population distribution. The test determines how well the obtained sample proportions fit the population proportions specified by the null hypothesis.

Recall from Chapter 2 that a frequency distribution is defined as a tabulation of the number of individuals located in each category of the scale of measurement. In a frequency distribution graph, the categories that make up the scale of measurement are listed on the X-axis. In a frequency distribution table, the categories are listed in the first column. With chi-square tests, however, it is customary to present the scale of measurement as a series of boxes, with each box corresponding to a separate category

on the scale. The frequency corresponding to each category is simply presented as a number written inside the box. Figure 17.1 shows how a distribution of eye colors for a set of $n = 40$ students can be presented as a graph, a table, or a series of boxes. The scale of measurement for this example consists of four categories of eye color (blue, brown, green, other).

THE NULL HYPOTHESIS FOR THE GOODNESS-OF-FIT TEST

For the chi-square test of goodness of fit, the null hypothesis specifies the proportion (or percentage) of the population in each category. For example, a hypothesis might state that 50% of all lawyers are men and 50% are women. The simplest way of presenting this hypothesis is to put the hypothesized proportions in the series of boxes representing the scale of measurement:

	Men	Women
H_0:	50%	50%

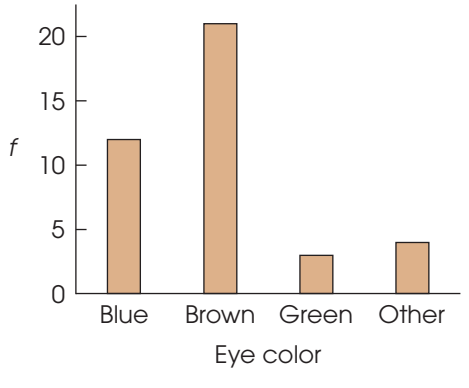
Although it is conceivable that a researcher could choose any proportions for the null hypothesis, there usually is some well-defined rationale for stating a null hypothesis. Generally H_0 falls into one of the following categories:

1. **No Preference, Equal Proportions.** The null hypothesis often states that there is no preference among the different categories. In this case, H_0 states that the population is divided equally among the categories. For example, a hypothesis stating that there is no preference among the three leading brands of soft drinks would specify a population distribution as follows:

	Brand X	Brand Y	Brand Z
H_0:	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(Preferences in the population are equally divided among the three soft drinks.)

The no-preference hypothesis is used in situations in which a researcher wants to determine whether there are any preferences among the categories, or whether the proportions differ from one category to another.



Eye color (X)	f
Blue	12
Brown	21
Green	3
Other	4

Blue	Brown	Green	Other
12	21	3	4

FIGURE 17.1

Distribution of eye colors for a sample of $n = 40$ individuals. The same frequency distribution is shown as a bar graph, as a table, and with the frequencies written in a series of boxes.

Because the null hypothesis for the goodness-of-fit test specifies an exact distribution for the population, the alternative hypothesis (H_1) simply states that the population distribution has a different shape from that specified in H_0. If the null hypothesis states that the population is equally divided among three categories, then the alternative hypothesis says that the population is not divided equally.

2. **No Difference from a Known Population.** The null hypothesis can state that the proportions for one population are not different from the proportions that are known to exist for another population. For example, suppose it is known that 28% of the licensed drivers in the state are younger than 30 years old and 72% are 30 or older. A researcher might wonder whether this same proportion holds for the distribution of speeding tickets. The null hypothesis would state that tickets are handed out equally across the population of drivers, so there is no difference between the age distribution for drivers and the age distribution for speeding tickets. Specifically, the null hypothesis would be

H_0:

Tickets Given to Drivers Younger Than 30	Tickets Given to Drivers 30 or Older
28%	72%

(Proportions for the population of tickets are not different from proportions for drivers.)

The no-difference hypothesis is used when a specific population distribution is already known. For example, you may have a known distribution from an earlier time, and the question is whether there has been any change in the proportions. Or, you may have a known distribution for one population (drivers) and the question is whether a second population (speeding tickets) has the same proportions.

Again, the alternative hypothesis (H_1) simply states that the population proportions are not equal to the values specified by the null hypothesis. For this example, H_1 would state that the number of speeding tickets is disproportionately high for one age group and disproportionately low for the other.

THE DATA FOR THE GOODNESS-OF-FIT TEST

The data for a chi-square test are remarkably simple. There is no need to calculate a sample mean or SS; you just select a sample of n individuals and count how many are in each category. The resulting values are called observed frequencies. The symbol for observed frequency is f_o. For example, the following data represent observed frequencies for a sample of 40 college students. The students were classified into three categories based on the number of times they reported exercising each week.

No Exercise	1 Time a Week	More Than Once a Week
15	19	6

$n = 40$

Notice that each individual in the sample is classified into one and only one of the categories. Thus, the frequencies in this example represent three completely separate groups of students: 15 who do not exercise regularly, 19 who average once a week, and 6 who exercise more than once a week. Also note that the observed frequencies add up to the total sample size: $\sum f_o = n$. Finally, you should realize that we are not assigning individuals to categories. Instead, we are simply measuring individuals to determine the category in which they belong.

DEFINITION

The **observed frequency** is the number of individuals from the sample who are classified in a particular category. Each individual is counted in one and only one category.

EXPECTED FREQUENCIES

The general goal of the chi-square test for goodness of fit is to compare the data (the observed frequencies) with the null hypothesis. The problem is to determine how well the data fit the distribution specified in H_0—hence the name *goodness of fit*.

The first step in the chi-square test is to construct a hypothetical sample that represents how the sample distribution would look if it were in perfect agreement with the proportions stated in the null hypothesis. Suppose, for example, the null hypothesis states that the population is distributed in three categories with the following proportions:

	Category A	Category B	Category C
H_0:	25%	50%	25%

(The population is distributed across the three categories with 25% in category A, 50% in category B, and 25% in category C.)

If this hypothesis is correct, how would you expect a random sample of $n = 40$ individuals to be distributed among the three categories? It should be clear that your best strategy is to predict that 25% of the sample would be in category A, 50% would be in category B, and 25% would be in category C. To find the exact frequency expected for each category, multiply the sample size (n) by the proportion (or percentage) from the null hypothesis. For this example, you would expect

$$25\% \text{ of } 40 = 0.25(40) = 10 \text{ individuals in category A}$$

$$50\% \text{ of } 40 = 0.50(40) = 20 \text{ individuals in category B}$$

$$25\% \text{ of } 40 = 0.25(40) = 10 \text{ individuals in category C}$$

The frequency values predicted from the null hypothesis are called *expected frequencies*. The symbol for expected frequency is f_e, and the expected frequency for each category is computed by

$$\text{expected frequency} = f_e = pn \tag{17.1}$$

where p is the proportion stated in the null hypothesis and n is the sample size.

DEFINITION

The **expected frequency** for each category is the frequency value that is predicted from the proportions in the null hypothesis and the sample size (n). The expected frequencies define an ideal, *hypothetical* sample distribution that would be obtained if the sample proportions were in perfect agreement with the proportions specified in the null hypothesis.

Note that the no-preference null hypothesis always produces equal f_e values for all categories because the proportions (p) are the same for all categories. On the other hand, the no-difference null hypothesis typically does not produce equal values for the expected frequencies because the hypothesized proportions typically vary from one category to another. You also should note that the expected frequencies are calculated, hypothetical values and the numbers that you obtain may be decimals or fractions. The observed frequencies, on the other hand, always represent real individuals and always are whole numbers.

THE CHI-SQUARE STATISTIC

The general purpose of any hypothesis test is to determine whether the sample data support or refute a hypothesis about the population. In the chi-square test for goodness of fit, the sample is expressed as a set of observed frequencies (f_o values), and the null hypothesis is used to generate a set of expected frequencies (f_e values). The *chi-square statistic* simply measures how well the data (f_o) fit the hypothesis (f_e). The symbol for the chi-square statistic is χ^2. The formula for the chi-square statistic is

$$\text{chi-square} = \chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} \tag{17.2}$$

As the formula indicates, the value of chi-square is computed by the following steps:

1. Find the difference between f_o (the data) and f_e (the hypothesis) for each category.
2. Square the difference. This ensures that all values are positive.
3. Next, divide the squared difference by f_e.
4. Finally, add the values from all of the categories.

The first two steps determine the numerator of the chi-square statistic and should be easy to understand. Specifically, the numerator measures how much difference there is between the data (the f_o values) and the hypothesis (represented by the f_e values). The final step is also reasonable: we add the values to obtain the total discrepancy between the data and the hypothesis. Thus, a large value for chi-square indicates that the data do not fit the hypothesis, and leads us to reject the null hypothesis.

However, the third step, which determines the denominator of the chi-square statistic, is not so obvious. Why must we divide by f_e before we add the category values? The answer to this question is that the obtained discrepancy between f_o and f_e is viewed as *relatively* large or *relatively* small depending on the size of the expected frequency. This point is demonstrated in the following analogy.

Suppose that you were going to throw a party and you *expected* 1,000 people to show up. However, at the party you counted the number of guests and *observed* that 1,040 actually showed up. Forty more guests than expected are no major problem when all along you were planning for 1,000. There will still probably be enough beer and potato chips for everyone. On the other hand, suppose you had a party and you expected 10 people to attend but instead 50 actually showed up. Forty more guests in this case spell big trouble. How "significant" the discrepancy is depends in part on what you were originally expecting. With very large expected frequencies, allowances are made for more error between f_o and f_e. This is accomplished in the chi-square formula by dividing the squared discrepancy for each category, $(f_o - f_e)^2$, by its expected frequency.

THE CHI-SQUARE DISTRIBUTION AND DEGREES OF FREEDOM

It should be clear from the chi-square formula that the numerical value of chi-square is a measure of the discrepancy between the observed frequencies (data) and the expected frequencies (H_0). As usual, the sample data are not expected to provide a perfectly accurate representation of the population. In this case, the proportions or observed frequencies in the sample are not expected to be exactly equal to the proportions in the population. Thus, if there are small discrepancies between the f_o and f_e values, we obtain a small value for chi-square and we conclude that there is a good fit between the data and the hypothesis (fail to reject H_0). However, when there

are large discrepancies between f_o and f_e, we obtain a large value for chi-square and conclude that the data do not fit the hypothesis (reject H_0). To decide whether a particular chi-square value is "large" or "small," we must refer to a *chi-square distribution*. This distribution is the set of chi-square values for all of the possible random samples when H_0 is true. Much like other distributions that we have examined (t distribution, F distribution), the chi-square distribution is a theoretical distribution with well-defined characteristics. Some of these characteristics are easy to infer from the chi-square formula.

1. The formula for chi-square involves adding squared values, so you can never obtain a negative value. Thus, all chi-square values are zero or larger.
2. When H_0 is true, you expect the data (f_o values) to be close to the hypothesis (f_e values). Thus, we expect chi-square values to be small when H_0 is true.

These two factors suggest that the typical chi-square distribution is positively skewed (Figure 17.2). Note that small values, near zero, are expected when H_0 is true and large values (in the right-hand tail) are very unlikely. Thus, unusually large values of chi-square form the critical region for the hypothesis test.

Although the typical chi-square distribution is positively skewed, there is one other factor that plays a role in the exact shape of the chi-square distribution—the number of categories. Recall that the chi-square formula requires that you add values from every category. The more categories you have, the more likely it is that you will obtain a large sum for the chi-square value. On average, chi-square is larger when you are adding values from 10 categories than when you are adding values from only 3 categories. As a result, there is a whole family of chi-square distributions, with the exact shape of each distribution determined by the number of categories used in the study. Technically, each specific chi-square distribution is identified by degrees of freedom (df) rather than the number of categories. For the goodness-of-fit test, the degrees of freedom are determined by

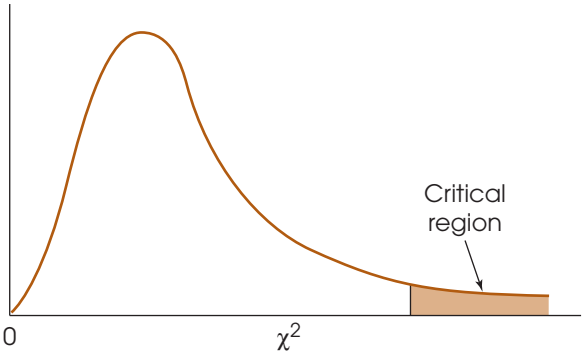
$$df = C - 1 \tag{17.3}$$

Caution: The *df* for a chi-square test is *not* related to sample size (n), as it is in most other tests.

where C is the number of categories. A brief discussion of this df formula is presented in Box 17.1. Figure 17.3 shows the general relationship between df and the shape of the chi-square distribution. Note that the chi-square values tend to get larger (shift to the right) as the number of categories and the degrees of freedom increase.

FIGURE 17.2

Chi-square distributions are positively skewed. The critical region is placed in the extreme tail, which reflects large chi-square values.



BOX 17.1 A CLOSER LOOK AT DEGREES OF FREEDOM

Degrees of freedom for the chi-square test literally measure the number of free choices that exist when you are determining the null hypothesis or the expected frequencies. For example, when you are classifying individuals into three categories, you have exactly two free choices in stating the null hypothesis. You may select any two proportions for the first two categories, but then the third proportion is determined. If you hypothesize 25% in the first category and 50% in the second category, then the third category must be 25% to account for 100% of the population.

Category A	Category B	Category C
10	20	?

In general, you are free to select proportions for all but one of the categories, but then the final proportion is determined by the fact that the entire set must total 100%. Thus, you have $C - 1$ free choices, where C is the number of categories: degrees of freedom, df, equal $C - 1$.

LOCATING THE CRITICAL REGION FOR A CHI-SQUARE TEST

Recall that a large value for the chi-square statistic indicates a big discrepancy between the data and the hypothesis, and suggests that we reject H_0. To determine whether a particular chi-square value is significantly large, you must consult the table entitled The Chi-Square Distribution (Appendix B). A portion of the chi-square table is shown in Table 17.2. The first column lists df values for the chi-square test, and the other column heads are proportions (alpha levels) in the extreme right-hand tail of the distribution. The numbers in the body of the table are the critical values of chi-square. The table shows, for example, that when the null hypothesis is true and $df = 3$, only 5% (.05) of the chi-square values are greater than 7.81, and only 1% (.01) are greater than 11.34. Thus, with $df = 3$, any chi-square value greater than 7.81 has a probability of $p < .05$, and any value greater than 11.34 has a probability of $p < .01$.

FIGURE 17.3

The shape of the chi-square distribution for different values of df. As the number of categories increases, the peak (mode) of the distribution has a larger chi-square value.

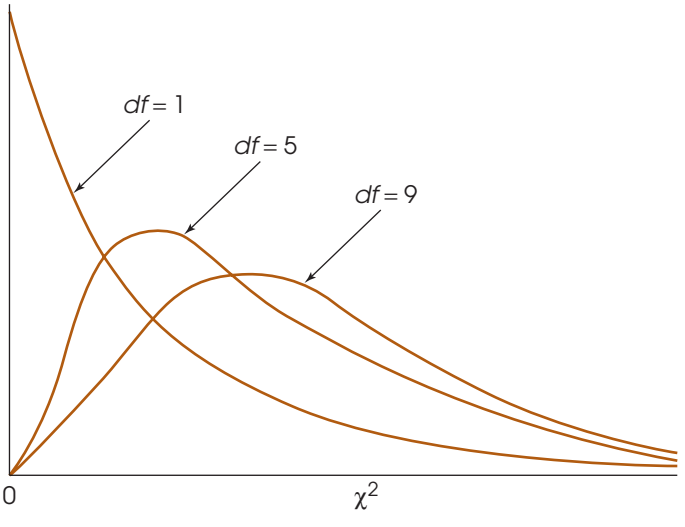


TABLE 17.2

A portion of the table of critical values for the chi-square distribution.

	Proportion in Critical Region				
df	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59

EXAMPLE OF THE CHI-SQUARE TEST FOR GOODNESS OF FIT

We use the same step-by-step process for testing hypotheses with chi-square as we used for other hypothesis tests. In general, the steps consist of stating the hypotheses, locating the critical region, computing the test statistic, and making a decision about H_0. The following example demonstrates the complete process of hypothesis testing with the goodness-of-fit test.

EXAMPLE 17.1

A psychologist examining art appreciation selected an abstract painting that had no obvious top or bottom. Hangers were placed on the painting so that it could be hung with any one of the four sides at the top. The painting was shown to a sample of $n = 50$ participants, and each was asked to hang the painting in the orientation that looked correct. The following data indicate how many people chose each of the four sides to be placed at the top:

Top Up (Correct)	Bottom Up	Left Side Up	Right Side Up
18	17	7	8

The question for the hypothesis test is whether there are any preferences among the four possible orientations. Are any of the orientations selected more (or less) often than would be expected simply by chance?

STEP 1 State the hypotheses and select an alpha level. The hypotheses can be stated as follows:

H_0: In the general population, there is no preference for any specific orientation. Thus, the four possible orientations are selected equally often, and the population distribution has the following proportions:

Top Up (Correct)	Bottom Up	Left Side Up	Right Side Up
25%	25%	25%	25%

H_1: In the general population, one or more of the orientations is preferred over the others.

We use $\alpha = .05$.

STEP 2 Locate the critical region. For this example, the value for degrees of freedom is

$$df = C - 1 = 4 - 1 = 3$$

For $df = 3$ and $\alpha = .05$, the table of critical values for chi-square indicates that the critical χ^2 has a value of 7.81. The critical region is sketched in Figure 17.4.

STEP 3 Calculate the chi-square statistic. The calculation of chi-square is actually a two-stage process. First, you must compute the expected frequencies from H_0 and then calculate the value of the chi-square statistic. For this example, the null hypothesis specifies that one-quarter of the population ($p = 25\%$) will be in each of the four categories. According to this hypothesis, we should expect one-quarter of the sample to be in each category. With a sample of $n = 50$ individuals, the expected frequency for each category is

Expected frequencies are computed and may be decimal values. Observed frequencies are always whole numbers.

$$f_e = pn = \frac{1}{4}(50) = 12.5$$

The observed frequencies and the expected frequencies are presented in Table 17.3. Using these values, the chi-square statistic may now be calculated.

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$
$$= \frac{(18-12.5)^2}{12.5} + \frac{(17-12.5)^2}{12.5} + \frac{(7-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5}$$
$$= \frac{30.25}{12.5} + \frac{20.25}{12.5} + \frac{30.25}{12.5} + \frac{20.25}{12.5}$$
$$= 2.42 + 1.62 + 2.42 + 1.62$$
$$= 8.08$$

STEP 4 State a decision and a conclusion. The obtained chi-square value is in the critical region. Therefore, H_0 is rejected, and the researcher may conclude that the four orientations are not equally likely to be preferred. Instead, there are significant differences among the four orientations, with some selected more often and others less often than would be expected by chance.

FIGURE 17.4

For Example 17.1, the critical region begins at a chi-square value of 7.81.

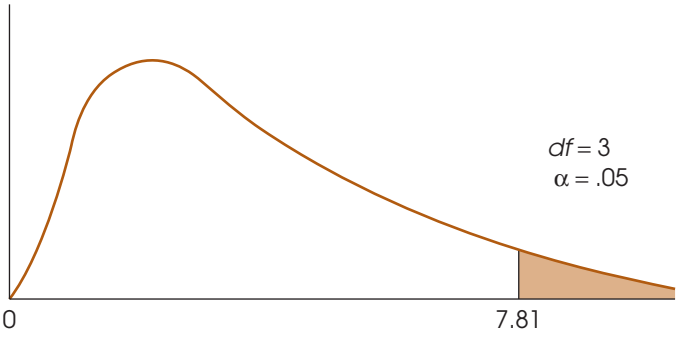


TABLE 17.3

The observed frequencies and the expected frequencies for the chi-square test in Example 17.1.

	Top Up (Correct)	Bottom Up	Left Side Up	Right Side Up
Observed Frequencies	18	17	7	8
Expected Frequencies	12.5	12.5	12.5	12.5

IN THE LITERATURE

REPORTING THE RESULTS FOR CHI-SQUARE

APA style specifies the format for reporting the chi-square statistic in scientific journals. For the results of Example 17.1, the report might state:

> The participants showed significant preferences among the four orientations for hanging the painting, $\chi^2(3, n = 50) = 8.08, p < .05$.

Note that the form of the report is similar to that of other statistical tests we have examined. Degrees of freedom are indicated in parentheses following the chi-square symbol. Also contained in the parentheses is the sample size (n). This additional information is important because the degrees of freedom value is based on the number of categories (C), not sample size. Next, the calculated value of chi-square is presented, followed by the probability that a Type I error has been committed. Because we obtained an extreme, very unlikely value for the chi-square statistic, the probability is reported as *less than* the alpha level. Additionally, the report may provide the observed frequencies (f_o) for each category. This information may be presented in a simple sentence or in a table.

GOODNESS OF FIT AND THE SINGLE-SAMPLE *t* TEST

We began this chapter with a general discussion of the difference between parametric tests and nonparametric tests. In this context, the chi-square test for goodness of fit is an example of a nonparametric test; that is, it makes no assumptions about the parameters of the population distribution, and it does not require data from an interval or ratio scale. In contrast, the single-sample t test introduced in Chapter 9 is an example of a parametric test: It assumes a normal population, it tests hypotheses about the population mean (a parameter), and it requires numerical scores that can be added, squared, divided, and so on.

Although the chi-square test and the single-sample t are clearly distinct, they are also very similar. In particular, both tests are intended to use the data from a single sample to test hypotheses about a single population.

The primary factor that determines whether you should use the chi-square test or the t test is the type of measurement that is obtained for each participant. If the sample data consist of numerical scores (from an interval or ratio scale), it is appropriate to compute a sample mean and use a t test to evaluate a hypothesis about the population mean. For example, a researcher could measure the IQ for each individual in a sample of registered voters. A t test could then be used to evaluate a hypothesis about the mean IQ for the entire population of registered voters. On the other hand, if the individuals in the sample are classified into nonnumerical categories (on a nominal or ordinal scale), then the researcher would use a chi-square test to evaluate a hypothesis about the population proportions. For example, a researcher could classify people according to gender by simply counting the number of males and females in a sample of registered voters. A chi-square test would then be appropriate to evaluate a hypothesis about the population proportions.

LEARNING CHECK

1. For a chi-square test, the observed frequencies are always whole numbers. (True or false?)
2. For a chi-square test, the expected frequencies are always whole numbers. (True or false?)
3. A researcher has developed three different designs for a computer keyboard. A sample of $n = 60$ participants is obtained, and each individual tests all three keyboards and identifies his or her favorite. The frequency distribution of preferences is as follows:

Design A	Design B	Design C
23	12	25

$n = 60$

   a. What is the *df* value for the chi-square statistic?
   b. Assuming that the null hypothesis states that there are no preferences among the three designs, find the expected frequencies for the chi-square test.

ANSWERS

1. True. Observed frequencies are obtained by counting people in the sample.
2. False. Expected frequencies are computed and may be fractions or decimal values.
3. a. $df = 2$
   b. According to the null hypothesis one-third of the population would prefer each design. The expected frequencies should show one-third of the sample preferring each design. The expected frequencies are all 20.

17.3 THE CHI-SQUARE TEST FOR INDEPENDENCE

The chi-square statistic may also be used to test whether there is a relationship between two variables. In this situation, each individual in the sample is measured or classified on two separate variables. For example, a group of students could be classified in terms of personality (introvert, extrovert) and in terms of color preference (red, yellow, green, or blue). Usually, the data from this classification are presented in the form of a matrix, where the rows correspond to the categories of one variable and the columns correspond to the categories of the second variable. Table 17.4 presents hypothetical data for a sample of $n = 200$ students who have been classified by personality and color preference. The number in each box, or cell, of the matrix indicates the frequency, or number of individuals in that particular group. In Table 17.4, for example, there are 10 students who were classified as introverted and who selected red as their preferred color. To obtain these data, the researcher first selects a random sample of $n = 200$ students. Each student is then given a personality test and is asked to select a preferred color from among the four choices. Note that the classification is based on the measurements for each student; the researcher does not assign students to categories. Also, note that the data consist of frequencies, not scores, from a sample. The goal is to use the frequencies from the sample to test a hypothesis about the population frequency distribution. Specifically, are these data sufficient to conclude that there is a significant relationship between personality and color preference in the population of students?

You should realize that the color preference study shown in Table 17.3 is an example of nonexperimental research (Chapter 1, page 17). The researcher did not manipulate any variable and the participants were not randomly assigned to groups or

TABLE 17.4
Color preferences according to personality types.

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

treatment conditions. However, similar data are often obtained from true experiments. A good example is the study described in the Preview, in which Loftus and Palmer (1974) demonstrate how eyewitness memory can be influenced by the kinds of questions that witnesses are asked. In the study, a sample of 150 students watched a film of an automobile accident. After watching the film, the students were separated into three groups and questioned about the accident. The researchers manipulated the type of question each group was asked. One group was asked to estimate the speed of the cars when they "smashed into each other." Another group estimated speed when the cars "hit each other." A third group served as a control and was not asked any question about the speed of the two cars. A week later, the participants returned and were asked if they remembered seeing any broken glass in the accident. (There was no broken glass in the film.) The researchers recorded the number of Yes and No responses for each group (see Table 17.1, page 592). As with the color preference data, the researchers would like to use the frequencies from the sample to test a hypothesis about the corresponding frequency distribution in the population. In this case, the researchers would like to know whether the sample data provide enough evidence to conclude that there is a significant relationship between eyewitnesses' memories and the questions they were asked.

The procedure for using sample frequencies to evaluate hypotheses concerning relationships between variables involves another test using the chi-square statistic. In this situation, however, the test is called the *chi-square test for independence.*

DEFINITION

The **chi-square test for independence** uses the frequency data from a sample to evaluate the relationship between two variables in the population. Each individual in the sample is classified on both of the two variables, creating a two-dimensional frequency-distribution matrix. The frequency distribution for the sample is then used to test hypotheses about the corresponding frequency distribution for the population.

THE NULL HYPOTHESIS FOR THE CHI-SQUARE TEST FOR INDEPENDENCE

The null hypothesis for the chi-square test for independence states that the two variables being measured are independent; that is, for each individual, the value obtained for one variable is not related to (or influenced by) the value for the second variable. This general hypothesis can be expressed in two different conceptual forms, each viewing the data and the test from slightly different perspectives. The data in Table 17.4 describing color preference and personality are used to present both versions of the null hypothesis.

H_0 version 1 For this version of H_0, the data are viewed as a single sample with each individual measured on two variables. The goal of the chi-square test is to evaluate the relationship between the two variables. For the example we are considering, the goal is to determine whether there is a consistent, predictable relationship between personality and color preference. That is, if I know your personality, will it help me to predict your color preference? The null hypothesis states that there is no relationship. The alternative hypothesis, H_1, states that there is a relationship between the two variables.

H_0: For the general population of students, there is no relationship between color preference and personality.

This version of H_0 demonstrates the similarity between the chi-square test for independence and a correlation. In each case, the data consist of two measurements (X and Y) for each individual, and the goal is to evaluate the relationship between the two variables. The correlation, however, requires numerical scores for X and Y. The chi-square test, on the other hand, simply uses frequencies for individuals classified into categories.

H_0 version 2 For this version of H_0, the data are viewed as two (or more) separate samples representing two (or more) populations or treatment conditions. The goal of the chi-square test is to determine whether there are significant differences between the populations. For the example we are considering, the data in Table 17.4 would be viewed as a sample of $n = 50$ introverts (top row) and a separate sample of $n = 150$ extroverts (bottom row). The chi-square test determines whether the distribution of color preferences for introverts is significantly different from the distribution of color preferences for extroverts. From this perspective, the null hypothesis is stated as follows:

H_0: In the population of students, the proportions in the distribution of color preferences for introverts are not different from the proportions in the distribution of color preferences for extroverts. The two distributions have the same shape (same proportions).

This version of H_0 demonstrates the similarity between the chi-square test and an independent-measures t test (or ANOVA). In each case, the data consist of two (or more) separate samples that are being used to test for differences between two (or more) populations. The t test (or ANOVA) requires numerical scores to compute means and mean differences. However, the chi-square test simply uses frequencies for individuals classified into categories. The null hypothesis for the chi-square test states that the populations have the same proportions (same shape). The alternative hypothesis, H_1, simply states that the populations have different proportions. For the example we are considering, H_1 states that the shape of the distribution of color preferences for introverts is different from the shape of the distribution of color preferences for extroverts.

Equivalence of H_0 version 1 and H_0 version 2 Although we have presented two different statements of the null hypothesis, these two versions are equivalent. The first version of H_0 states that color preference is not related to personality. If this hypothesis is correct, then the distribution of color preferences should not depend on personality. In other words, the distribution of color preferences should have the same proportions for introverts and for extroverts, which is the second version of H_0.

For example, if we found that 60% of the introverts preferred red, then H_0 would predict that we also should find that 60% of the extroverts prefer red. In this case, knowing that an individual prefers red does not help you predict his or her personality. Note that finding the *same proportions* indicates *no relationship*.

On the other hand, if the proportions were different, it would suggest that there is a relationship. For example, if red is preferred by 60% of the extroverts but only 10% of the introverts, then there is a clear, predictable relationship between personality and color preference. (If I know your personality, then I can predict your color preference.) Thus, finding *different proportions* means that there is *a relationship* between the two variables.

DEFINITION

Two variables are **independent** when there is no consistent, predictable relationship between them. In this case, the frequency distribution for one variable is not related to (or dependent on) the categories of the second variable. As a result, when two variables are independent, the frequency distribution for one variable has the same shape (same proportions) for all categories of the second variable.

Thus, stating that there is no relationship between two variables (version 1 of H_0) is equivalent to stating that the distributions have equal proportions (version 2 of H_0).

OBSERVED AND EXPECTED FREQUENCIES

The chi-square test for independence uses the same basic logic that was used for the goodness-of-fit test. First, a sample is selected and each individual is classified or categorized. Because the test for independence considers two variables, every individual is classified on both variables, and the resulting frequency distribution is presented as a two-dimensional matrix (see Table 17.4). As before, the frequencies in the sample distribution are called *observed frequencies* and are identified by the symbol f_o.

The next step is to find the expected frequencies, or f_e values, for this chi-square test. As before, the e*xpected frequencies* define an ideal hypothetical distribution that is in perfect agreement with the null hypothesis. Once the expected frequencies are obtained, we compute a chi-square statistic to determine how well the data (observed frequencies) fit the null hypothesis (expected frequencies).

Although you can use either version of the null hypothesis to find the expected frequencies, the logic of the process is much easier when you use H_0 stated in terms of equal proportions. For the example we are considering, the null hypothesis states

H_0: The frequency distribution of color preference has the same shape (same proportions) for both categories of personality.

To find the expected frequencies, we first determine the overall distribution of color preferences and then apply this distribution to both categories of personality. Table 17.5 shows an empty matrix corresponding to the data from Table 17.4. Notice that the empty matrix includes all of the row totals and column totals from the original sample data. The row totals and column totals are essential for computing the expected frequencies.

The column totals for the matrix describe the overall distribution of color preferences. For these data, 100 people selected red as their preferred color. Because the total sample consists of 200 people, the proportion selecting red is 100 out of 200, or 50%. The complete set of color preference proportions is as follows:

100 out of 200 = 50% prefer red

20 out of 200 = 10% prefer yellow

40 out of 200 = 20% prefer green

40 out of 200 = 20% prefer blue

The row totals in the matrix define the two samples of personality types. For example, the matrix in Table 17.5 shows a total of 50 introverts (the top row) and a sample of 150 extroverts (the bottom row). According to the null hypothesis, both personality groups should have the same proportions for color preferences. To find the expected frequencies, we simply apply the overall distribution of color preferences to

TABLE 17.5

An empty frequency distribution matrix showing only the row totals and column totals. (These numbers describe the basic characteristics of the sample from Table 17.4.)

	Red	Yellow	Green	Blue	
Introvert					50
Extrovert					150
	100	20	40	40	

each sample. Beginning with the sample of 50 introverts in the top row, we obtain expected frequencies of

50% prefer red: $f_e = 50\%$ of $50 = 0.50(50) = 25$
10% prefer yellow: $f_e = 10\%$ of $50 = 0.10(50) = 5$
20% prefer green: $f_e = 20\%$ of $50 = 0.20(50) = 10$
20% prefer blue: $f_e = 20\%$ of $50 = 0.20(50) = 10$

Using exactly the same proportions for the sample of $n = 150$ extroverts in the bottom row, we obtain expected frequencies of

50% prefer red: $f_e = 50\%$ of $150 = 0.50(50) = 75$
10% prefer yellow: $f_e = 10\%$ of $150 = 0.10(50) = 15$
20% prefer green: $f_e = 20\%$ of $150 = 0.20(50) = 30$
20% prefer blue: $f_e = 20\%$ of $150 = 0.20(50) = 30$

The complete set of expected frequencies is shown in Table 17.6. Notice that the row totals and the column totals for the expected frequencies are the same as those for the original data (the observed frequencies) in Table 17.3.

A simple formula for determining expected frequencies Although expected frequencies are derived directly from the null hypothesis and the sample characteristics, it is not necessary to go through extensive calculations to find f_e values. In fact, there is a simple formula that determines f_e for any cell in the frequency distribution matrix:

$$f_e = \frac{f_c f_r}{n} \tag{17.4}$$

where f_c is the frequency total for the column (column total), f_r is the frequency total for the row (row total), and n is the number of individuals in the entire sample. To demonstrate this formula, we compute the expected frequency for introverts selecting yellow in Table 17.6. First, note that this cell is located in the top row and second column in the table. The column total is $f_c = 20$, the row total is $f_r = 50$, and the sample size is $n = 200$. Using these values in formula 17.4, we obtain

$$f_e = \frac{f_c f_r}{n} = \frac{20(50)}{200} = 5$$

TABLE 17.6

Expected frequencies corresponding to the data in Table 17.4. (This is the distribution predicted by the null hypothesis.)

	Red	Yellow	Green	Blue	
Introvert	25	5	10	10	50
Extrovert	75	15	30	30	150
	100	20	40	40	

This is identical to the expected frequency we obtained using percentages from the overall distribution.

THE CHI-SQUARE STATISTIC AND DEGREES OF FREEDOM

The chi-square test for independence uses exactly the same chi-square formula as the test for goodness of fit:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

As before, the formula measures the discrepancy between the data (f_o values) and the hypothesis (f_e values). A large discrepancy produces a large value for chi-square and indicates that H_0 should be rejected. To determine whether a particular chi-square statistic is significantly large, you must first determine degrees of freedom (*df*) for the statistic and then consult the chi-square distribution in Appendix B. For the chi-square test of independence, degrees of freedom are based on the number of cells for which you can freely choose expected frequencies. Recall that the f_e values are partially determined by the sample size (n) and by the row totals and column totals from the original data. These various totals restrict your freedom in selecting expected frequencies. This point is illustrated in Table 17.7. Once three of the f_e values have been selected, all of the other f_e values in the table are also determined. For example, the bottom number in the first column must be 75 to produce a column total of 100. Similarly, the last number in the top row must be 10 to produce a row total of 50. In general, the row totals and the column totals restrict the final choices in each row and column. As a result, we may freely choose all but one f_e in each row and all but one f_e in each column. If R is the number of rows and C is the number of columns, and you remove the last column and the bottom row from the matrix, you are left with a smaller matrix that has $C - 1$ columns and $R - 1$ rows. The number of cells in the smaller matrix determines the *df* value. Thus, the total number of f_e values that you can freely choose is $(R - 1)(C - 1)$, and the degrees of freedom for the chi-square test of independence are given by the formula

$$df = (R - 1)(C - 1) \tag{17.5}$$

Also note that once you calculate the expected frequencies to fill the smaller matrix, the rest of the f_e values can be found by subtraction.

AN EXAMPLE OF THE CHI-SQUARE TEST FOR INDEPENDENCE

The following example demonstrates the complete hypothesis-testing procedure for the chi-square test for independence.

TABLE 17.7

Degrees of freedom and expected frequencies. (Once three values have been selected, all the remaining expected frequencies are determined by the row totals and the column totals. This example has only three free choices, so $df = 3$.)

Red	Yellow	Green	Blue	
25	5	10	?	50
?	?	?	?	150
100	20	40	40	

EXAMPLE 17.2

Research has demonstrated strong gender differences in teenagers' approaches to dealing with mental health issues (Chandra & Minkovitz, 2006). In a typical study, eighth-grade students are asked to report their willingness to use mental health services in the event they were experiencing emotional or other mental health problems. Typical data for a sample of $n = 150$ students are shown in Table 17.8. Do the data show a significant relationship between gender and willingness to seek mental health assistance?

STEP 1 State the hypotheses, and select a level of significance. According to the null hypothesis, the two variables are independent. This general hypothesis can be stated in two different ways:

Version 1

H_0: In the general population, there is no relationship between gender and willingness to use mental health services.

This version of H_0 emphasizes the similarity between the chi-square test and a correlation. The corresponding alternative hypothesis would state:

H_1: In the general population, there is a consistent, predictable relationship between gender and willingness to use mental health services.

Version 2

H_0: In the general population, the distribution of reported willingness to use mental health services has the same proportions for males and for females.

The corresponding alternative hypothesis would state:

H_1: In the general population, the distribution of reported willingness to use mental health services for males has proportions that are different from those in the distribution for females.

The second version of H_0 emphasizes the similarity between the chi-square test and the independent-measures t test.

Remember that the two versions for the hypotheses are equivalent. The choice between them is largely determined by how the researcher wants to describe the outcome. For example, a researcher may want to emphasize the *relationship* between variables or the *difference* between groups.

For this test, we use $\alpha = .05$.

STEP 2 Determine the degrees of freedom and locate the critical region. For the chi-square test for independence,

$$df = (R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$$

With $df = 2$ and $\alpha = .05$, the critical value for chi-square is 5.99 (see Table B.8, p. 711).

TABLE 17.8

A frequency distribution showing willingness to use mental health services according to gender for a sample of $n = 150$ students.

	Willingness to Use Mental Health Services			
	Probably No	Maybe	Probably Yes	
Males	17	32	11	60
Females	13	43	34	90
	30	75	45	$n = 150$

STEP 3 Determine the expected frequencies, and compute the chi-square statistic. The following table shows an empty matrix with the same row totals and column totals as the original data. The expected frequencies must maintain the same row totals and column totals, and create an ideal frequency distribution that perfectly represents the null hypothesis. Specifically, the proportions for the group of 60 males must be the same as the proportions for the group of 90 females.

Willingness to Use Mental Health Services

	Probably No	Maybe	Probably Yes	
Males				60
Females				90
	30	75	45	$n = 150$

The column totals describe the overall distribution of willingness. These totals indicate that 30 out of 150 students reported that they would probably not use mental health services. This proportion corresponds to $\frac{30}{150}$, or 20% of the total sample. Similarly, $\frac{75}{150} = 50\%$ reported that they might use mental health services. Finally, $\frac{45}{150} = 30\%$ reported that they probably would use the services. The null hypothesis (version 2) states that these proportions are the same for males and females. Therefore, we simply apply the proportions to each group to obtain the expected frequencies. For the group of 60 males (top row), we obtain

20% of 60 = 12 males who would probably not seek services

50% of 60 = 30 males who might seek services

30% of 60 = 18 males who probably would seek services

For the group of 90 females (bottom row), we expect

20% of 90 = 18 females who would probably not seek services

50% of 90 = 45 females who may seek services

30% of 90 = 27 females who probably would seek services

These expected frequencies are summarized in Table 17.9.

The chi-square statistic is now used to measure the discrepancy between the data (the observed frequencies in Table 17.8) and the null hypothesis that was used to generate the expected frequencies in Table 17.9.

TABLE 17.9

The expected frequencies (f_e values) of willingness to use mental services is completely independent of gender.

Willingness to Use Mental Health Services

	Probably No	Maybe	Probably Yes	
Males	12	30	18	60
Females	18	45	27	90
	30	75	45	

$$\chi^2 = \frac{(17-12)^2}{12} + \frac{(32-30)^2}{30} - \frac{(11-18)^2}{18} + \frac{(13-18)^2}{18} + \frac{(43-45)^2}{45} + \frac{(34-27)^2}{27}$$

$$= 2.08 + 0.13 + 2.72 + 1.39 + 0.09 + 1.82$$

$$= 8.23$$

STEP 4 Make a decision regarding the null hypothesis and the outcome of the study. The obtained chi-square value exceeds the critical value (5.99). Therefore, the decision is to reject the null hypothesis. In the literature, this would be reported as a significant result with $\chi^2(2, n = 150) = 8.23$, $p < .05$. According to version 1 of H_0, this means that we have decided there is a significant relationship between gender and willingness to use mental health services. Expressed in terms of version 2 of H_0, the data show a significant difference between males' and females' attitudes toward using mental health services. To describe the details of the significant result, you must compare the original data (Table 17.8) with the expected frequencies in Table 17.9. Looking at the two tables, it should be clear that males were less willing to use mental health services and females were more willing than would be expected if the two variables were independent.

LEARNING CHECK

1. A researcher would like to know which factors are most important to people who are buying a new car. A sample of $n = 200$ customers between the ages of 20 and 29 are asked to identify the most important factor in the decision process: Performance, Reliability, or Style. The researcher would like to know whether there is a difference between the factors identified by women compared to those identified by men. The data are as follows:

Observed Frequencies of Most Important factor According to Gender

	Performance	Reliability	Style	Totals
Male	21	33	26	80
Female	19	67	34	120
Totals	40	100	60	

a. State the null hypotheses.

b. Determine the value for *df* for the chi-square test.

c. Compute the expected frequencies.

ANSWERS

1. a. H_0: In the population, the distribution of preferred factors for men has the same proportions as the distribution for women.

b. $df = 2$

c. f_e values are as follows:

Expected Frequencies

	Performance	Reliability	Style
Male	16	40	24
Female	24	60	36

17.4 MEASURING EFFECT SIZE FOR THE CHI-SQUARE TEST FOR INDEPENDENCE

A hypothesis test, like the chi-square test for independence, evaluates the statistical significance of the results from a research study. Specifically, the intent of the test is to determine whether it is likely that the patterns or relationships observed in the sample data could have occurred without any corresponding patterns or relationships in the population. Tests of significance are influenced not only by the size or strength of the treatment effects but also by the size of the samples. As a result, even a small effect can be statistically significant if it is observed in a very large sample. Because a significant effect does not necessarily mean a large effect, it is generally recommended that the outcome of a hypothesis test be accompanied by a measure of the effect size. This general recommendation also applies to the chi-square test for independence.

THE PHI-COEFFICIENT AND CRAMÉR'S *V*

In Chapter 15 (p. 545), we introduced the *phi-coefficient* as a measure of correlation for data consisting of two dichotomous variables (both variables have exactly two values). This same situation exists when the data for a chi-square test for independence form a 2×2 matrix (again, each variable has exactly two values). In this case, it is possible to compute the correlation phi (ϕ) in addition to the chi-square hypothesis test for the same set of data. Because phi is a correlation, it measures the strength of the relationship, rather than the significance, and thus provides a measure of effect size. The value for the phi-coefficient can be computed directly from chi-square by the following formula:

Caution: The value of χ^2 is already a squared value. Do not square it again.

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{17.6}$$

The value of the phi-coefficient is determined entirely by the *proportions* in the 2×2 data matrix and is completely independent of the absolute size of the frequencies. The chi-square value, however, is influenced by the proportions and by the size of the frequencies. This distinction is demonstrated in the following example.

EXAMPLE 17.3

The following data show a frequency distribution evaluating the relationship between gender and preference between two candidates for student president.

Candidate

	A	B
Male	5	10
Female	10	5

Note that the data show that males prefer candidate B by a 2-to-1 margin and females prefer candidate A by 2 to 1. Also note that the sample includes a total of 15 males and 15 females. We will not perform all the arithmetic here, but these data produce chi-square equal to 3.33 (which is not significant) and a phi-coefficient of 0.333.

Next we keep exactly the same proportions in the data, but double all of the frequencies. The resulting data are as follows:

	Candidate A	Candidate B
Male	10	20
Female	20	10

Once again, males prefer candidate B by 2 to 1 and females prefer candidate A by 2 to 1. However, the sample now contains 30 males and 30 females. For these new data, the value of chi-square is 6.66, twice as big as it was before (and now significant with $\alpha = .05$), but the value of the phi-coefficient is still 0.333.

Because the proportions are the same for the two samples, the value of the phi-coefficient is unchanged. However, the larger sample provides more convincing evidence than the smaller sample, so the larger sample is more likely to produce a significant result.

The interpretation of ϕ follows the same standards used to evaluate a correlation (Table 9.3, p. 299 shows the standards for squared correlations): a correlation of 0.10 is a small effect, 0.30 is a medium effect, and 0.50 is a large effect. Occasionally, the value of ϕ is squared (ϕ^2) and is reported as a percentage of variance accounted for, exactly the same as r^2.

When the chi-square test involves a matrix larger than 2×2, a modification of the phi-coefficient, known as *Cramér's* V, can be used to measure effect size.

$$V = \sqrt{\frac{\chi^2}{n(df^*)}} \tag{17.7}$$

Note that the formula for Cramér's V (17.7) is identical to the formula for the phi-coefficient (17.6) except for the addition of df^* in the denominator. The df^* value is *not* the same as the degrees of freedom for the chi-square test, but it is related. Recall that the chi-square test for independence has $df = (R - 1)(C - 1)$, where R is the number of rows in the table and C is the number of columns. For Cramér's V, the value of df^* is the smaller of either $(R - 1)$ or $(C - 1)$.

Cohen (1988) has also suggested standards for interpreting Cramér's V that are shown in Table 17.10. Note that when $df^* = 1$, as in a 2×2 matrix, the criteria for interpreting V are exactly the same as the criteria for interpreting a regular correlation or a phi-coefficient.

We will use the results from Example 17.2 (p. 610) to demonstrate the calculation of Cramér's V. The example evaluated the relationship between gender and willingness to use mental health services. There were two levels of gender and three levels of willingness producing a 2×3 table with a total of $n = 150$ participants. The data produced $\chi^2 = 8.23$. Using these values, we obtain

$$V = \sqrt{\frac{\chi^2}{n(df^*)}} = \sqrt{\frac{8.23}{150(1)}} = \sqrt{0.055} = 0.23$$

TABLE 17.10

Standards for interpreting Cramér's *V* as proposed by Cohen (1988).

	Small Effect	Medium Effect	Large Effect
For $df^* = 1$	0.10	0.30	0.50
For $df^* = 2$	0.07	0.21	0.35
For $df^* = 3$	0.06	0.17	0.29

According to Cohen's guidelines (see Table 17.10), this value indicates a small or medium relationship.

In a research report, the measure of effect size appears immediately after the results of the hypothesis test. For the study in Example 17.2, the results would be reported as follows:

> The results showed a significant difference between males' and females' attitudes toward using mental health services, $\chi^2(2, n = 50) = 8.23$, $p < .05$, $V = 0.23$.

17.5 ASSUMPTIONS AND RESTRICTIONS FOR CHI-SQUARE TESTS

To use a chi-square test for goodness of fit or a test of independence, several conditions must be satisfied. For any statistical test, violation of assumptions and restrictions casts doubt on the results. For example, the probability of committing a Type I error may be distorted when assumptions of statistical tests are not satisfied. Some important assumptions and restrictions for using chi-square tests are the following:

1. **Independence of Observations.** This is *not* to be confused with the concept of independence between *variables,* as seen in the chi-square test for independence (Section 17.3). One consequence of independent observations is that each observed frequency is generated by a different individual. A chi-square test would be inappropriate if a person could produce responses that can be classified in more than one category or contribute more than one frequency count to a single category. (See p. 254 for more information on independence.)
2. **Size of Expected Frequencies.** A chi-square test should not be performed when the expected frequency of any cell is less than 5. The chi-square statistic can be distorted when f_e is very small. Consider the chi-square computations for a single cell. Suppose that the cell has values of $f_e = 1$ and $f_o = 5$. Note that there is a 4-point difference between the observed and expected frequencies. However, the total contribution of this cell to the total chi-square value is

$$\text{cell} = \frac{(f_o - f_e)^2}{f_e} = \frac{(5 - 1)^2}{1} = \frac{4^2}{1} = 16$$

Now consider another instance, in which $f_e = 10$ and $f_o = 14$. The difference between the observed and the expected frequencies is still 4, but the contribution of this cell to the total chi-square value differs from that of the first case:

$$\text{cell} = \frac{(f_o - f_e)^2}{f_e} = \frac{(14 - 10)^2}{10} = \frac{4^2}{10} = 1.6$$

It should be clear that a small f_e value can have a great influence on the chi-square value. This problem becomes serious when f_e values are less than 5. When f_e is very small, what would otherwise be a minor discrepancy between f_o and f_e results in large chi-square values. The test is too sensitive when f_e values are extremely small. One way to avoid small expected frequencies is to use large samples.

LEARNING CHECK

1. A researcher completes a chi-square test for independence and obtains $\chi^2 = 6.2$ for a sample of $n = 40$ participants.
   a. If the frequency data formed a 2×2 matrix, what is the phi-coefficient for the test?
   b. If the frequency data formed a 3×3 matrix, what is Cramér's V for the test?
2. Explain why a very small value for an expected frequency can distort the results of a chi-square test.

ANSWERS

1. a. $\phi = 0.394$
   b. $V = 0.278$
2. With a very small value for an expected frequency, even a minor discrepancy between the observed frequency and the expected frequency can produce a large number that is added into the chi-square statistic. This inflates the value of chi-square and can distort the outcome of the test.

17.6 SPECIAL APPLICATIONS OF THE CHI-SQUARE TESTS

At the beginning of this chapter, we introduced the chi-square tests as examples of nonparametric tests. Although nonparametric tests serve a function that is uniquely their own, they also can be viewed as alternatives to the common parametric tests that were examined in earlier chapters. In general, nonparametric tests are used as substitutes for parametric tests in situations in which one of the following occurs:

1. The data do not meet the assumptions needed for a standard parametric test.
2. The data consist of nominal or ordinal measurements, so that it is impossible to compute standard descriptive statistics such as the mean and standard deviation.

In this section, we examine some of the relationships between chi-square tests and the parametric procedures for which they may be substituted.

CHI-SQUARE AND THE PEARSON CORRELATION

The chi-square test for independence and the Pearson correlation are both statistical techniques intended to evaluate the relationship between two variables. The type of data obtained in a research study determines which of these two statistical procedures is appropriate. Suppose, for example, that a researcher is interested in the relationship between self-esteem and academic performance for 10-year-old children. If the researcher obtained numerical scores for both variables, then the resulting data would be similar to the values shown in Table 17.11(a) and the researcher could use a Pearson correlation to evaluate the relationship. On the other hand, if both variables are classified into non-numerical categories as in Table 17.11(b), then the data consist of frequencies and the relationship could be evaluated with a chi-square test for independence.

TABLE 17.11

Two possible data structures for research studies examining the relationship between self-esteem and academic performance. In part (a) there are numerical scores for both variables and the data are suitable for a correlation. In part (b) both variables are classified into categories and the data are frequencies suitable for a chi-square test.

(a)

Participant	Self-Esteem X	Academic Performance Y
A	13	73
B	19	88
C	10	71
D	22	96
E	20	90
F	15	82
.	.	.
.	.	.
.	.	.

(b)

		Level of Self-Esteem: High	Medium	Low	
Academic Performance	High	17	32	11	60
	Low	13	43	34	90
		30	75	45	$n = 150$

CHI-SQUARE AND THE INDEPENDENT-MEASURES *t* AND ANOVA

Once again, consider a researcher investigating the relationship between self-esteem and academic performance for 10-year-old children. This time, suppose that the researcher measured academic performance by simply classifying individuals into two categories, high and low, and then obtained a numerical score for each individual's self-esteem. The resulting data would be similar to the scores in Table 17.12(a), and an independent-measures *t* test would be used to evaluate the mean difference between the two groups of scores. Alternatively, the researcher could measure self-esteem by classifying individuals into three categories: high, medium, and low. If a numerical score is then obtained for each individual's academic performance, the resulting data would look like the scores in Table 17.12(b), and an ANOVA would be used to evaluate the mean differences among the three groups. Finally, if both variables are classified into non-numerical categories, then the data would look like the scores shown earlier in Table 17.11(b) and a chi-square test for independence would be used to evaluate the difference between the two academic-performance groups or the differences among the three self-esteem groups.

The point of these examples is that the chi-square test for independence, the Pearson correlation, and tests for mean differences can all be used to evaluate the relationship between two variables. One main distinction among the different statistical procedures is the form of the data. However, another distinction is the fundamental purpose of these different statistics. The chi-square test and the tests for mean differences (*t* and ANOVA) evaluate the *significance* of the relationship; that is, they determine whether the relationship observed in the sample provides enough evidence to conclude that there is a corresponding relationship in the population. You can also evaluate the significance of a Pearson correlation, however, the main purpose of a correlation is to measure the *strength* of the relationship. In particular, squaring the correlation, r^2,

TABLE 17.12

Data appropriate for an independent-measures *t* test or an ANOVA. In part (a), self-esteem scores are obtained for two groups of students differing in level of academic performance. In part (b), academic performance scores are obtained for three groups of students differing in level of self-esteem.

(a) Self-esteem scores for two groups of students.

Academic Performance	
High	Low
17	13
21	15
16	14
24	20
18	17
15	14
19	12
20	19
18	16

(b) Academic performance scores for three groups of students.

Self-esteem		
High	Medium	Low
94	83	80
90	76	72
85	70	81
84	81	71
89	78	77
96	88	70
91	83	78
85	80	72
88	82	75

provides a measure of effect size, describing the proportion of variance in one variable that is accounted for by its relationship with the other variable.

THE MEDIAN TEST FOR INDEPENDENT SAMPLES

The median is the score that divides the population in half, with 50% scoring at or below the median.

The *median test* provides a nonparametric alternative to the independent-measures *t* test (or ANOVA) to determine whether there are significant differences among two or more independent samples. The null hypothesis for the median test states that the different samples come from populations that share a common median (no differences). The alternative hypothesis states that the samples come from populations that are different and do not share a common median.

The logic behind the median test is that whenever several different samples are selected from the same population distribution, roughly half of the scores in each sample should be above the population median and roughly half should be below. That is, all of the separate samples should be distributed around the same median. On the other hand, if the samples come from populations with different medians, then the scores in some samples will be consistently higher and the scores in other samples will be consistently lower.

The first step in conducting the median test is to combine all of the scores from the separate samples and then find the median for the combined group (see Chapter 3, page 83, for instructions for finding the median). Next, a matrix is constructed with a column for each of the separate samples and two rows: one for individuals with scores above the median and one for individuals with scores below the median. Finally, for each sample, count how many individuals scored above the combined median and how many scored below. These values are the observed frequencies that are entered in the matrix.

The frequency-distribution matrix is evaluated using a chi-square test for independence. The expected frequencies and a value for chi-square are computed exactly as described in Section 17.3. A significant value for chi-square indicates that the discrepancy between the individual sample distributions is greater than would be expected by chance.

The median test is demonstrated in the following example.

EXAMPLE 17.4

The following data represent self-esteem scores obtained from a sample of $n = 40$ children. The children are then separated into three groups based on their level of academic performance (high, medium, low). The median test evaluates whether there is a significant relationship between self-esteem and level of academic performance.

Self-Esteem Scores for Children at Three Levels of Academic Performance

High		Medium				Low	
22	14	22	13	24	20	11	19
19	18	18	22	10	16	13	15
12	21	19	15	14	19	20	16
20	18	11	18	11	10	10	18
23	20	12	19	15	12	15	11

The median for the combined group of $n = 40$ scores is $X = 17$ (exactly 20 scores are above this value and 20 are below). For the high performers, 8 out of 10 scores are above the median. For the medium performers, 9 out of 20 are above the median, and for the low performers, only 3 out of 10 are above the median. These observed frequencies are shown in the following matrix:

	Academic Performance		
	High	Medium	Low
Above Median	8	9	3
Below Median	2	11	7

The expected frequencies for this test are as follows:

	Academic Performance		
	High	Medium	Low
Above Median	5	10	5
Below Median	5	10	5

The chi-square statistic is

$$\chi^2 = \frac{9}{5} + \frac{9}{5} + \frac{1}{10} + \frac{1}{10} + \frac{4}{5} + \frac{4}{5} = 5.40$$

With $df = 2$ and $\alpha = .05$, the critical value for chi-square is 5.99. The obtained chi-square of 5.40 does not fall in the critical region, so we would fail to reject the null hypothesis. These data do not provide sufficient evidence to conclude that there are significant differences among the self-esteem distributions for these three groups of students.

A few words of caution are in order concerning the interpretation of the median test. First, the median test is *not* a test for mean differences. Remember: The mean for a distribution can be strongly affected by a few extreme scores. Therefore, the mean and the median for a distribution are not necessarily the same, and they may not even be related. The results from a median test *cannot* be interpreted as indicating that there is (or is not) a difference between means.

Second, you may have noted that the median test does not directly compare the median from one sample with the median from another. Thus, the median test is not a test for significant differences between medians. Instead, this test compares the distribution of scores for one sample to the distribution for another sample. If the samples are distributed evenly around a common point (the group median), then you can conclude that there is no significant difference. On the other hand, finding a significant difference simply indicates that the samples are not distributed evenly around the common median. Thus, the best interpretation of a significant result is that there is a *difference in the distributions* of the samples.

SUMMARY

1. Chi-square tests are nonparametric techniques that test hypotheses about the form of the entire frequency distribution. Two types of chi-square tests are the test for goodness of fit and the test for independence. The data for these tests consist of the frequency or number of individuals who are located in each category.
2. The test for goodness of fit compares the frequency distribution for a sample to the population distribution that is predicted by H_0. The test determines how well the observed frequencies (sample data) fit the expected frequencies (data predicted by H_0).
3. The expected frequencies for the goodness-of-fit test are determined by

$$\text{expected frequency} = f_e = pn$$

where p is the hypothesized proportion (according to H_0) of observations falling into a category and n is the size of the sample.

4. The chi-square statistic is computed by

$$\chi^2 = \Sigma\frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency for a particular category and f_e is the expected frequency for that category. Large values for χ^2 indicate that there is a large discrepancy between the observed (f_o) and the expected (f_e) frequencies and may warrant rejection of the null hypothesis.

5. Degrees of freedom for the test for goodness of fit are

$$df = C - 1$$

where C is the number of categories in the variable. Degrees of freedom measure the number of categories for which f_e values can be freely chosen. As can be seen from the formula, all but the last f_e value to be determined are free to vary.

6. The chi-square distribution is positively skewed and begins at the value of zero. Its exact shape is determined by degrees of freedom.
7. The test for independence is used to assess the relationship between two variables. The null hypothesis states that the two variables in question are independent of each other. That is, the frequency distribution for one variable does not depend on the categories of the second variable. On the other hand, if a relationship does exist, then the form of the distribution for one variable depends on the categories of the other variable.
8. For the test for independence, the expected frequencies for H_0 can be directly calculated from the marginal frequency totals,

$$f_e = \frac{f_c f_r}{n}$$

where f_c is the total column frequency and f_r is the total row frequency for the cell in question.

9. Degrees of freedom for the test for independence are computed by

$$df = (R - 1)(C - 1)$$

where R is the number of row categories and C is the number of column categories.

10. For the test for independence, a large chi-square value means there is a large discrepancy between the f_o and f_e values. Rejecting H_0 in this test provides support for a relationship between the two variables.
11. Both chi-square tests (for goodness of fit and independence) are based on the assumption that each observation is independent of the others. That is, each observed frequency reflects a different individual, and no individual can produce a response that would be

classified in more than one category or more than one frequency in a single category.

12. The chi-square statistic is distorted when f_e values are very small. Chi-square tests, therefore, should not be performed when the expected frequency of any cell is less than 5.

13. The effect size for a chi-square test for independence is measured by computing a phi-coefficient for data that form a 2×2 matrix or computing Cramér's V for a matrix that is larger than 2×2.

$$\text{phi} = \sqrt{\frac{\chi^2}{n}} \qquad \text{Cramér's } V = \sqrt{\frac{\chi^2}{n(df^*)}}$$

where df^* is the smaller of $(R - 1)$ and $(C - 1)$. Both phi and Cramér's V are evaluated using the criteria in Table 17.10.

KEY TERMS

parametric test (593)
nonparametric test (593)
chi-square test for goodness-of-fit (594)
observed frequencies (597)
expected frequencies (597)
chi-square statistic (598)
chi-square distribution (599)
chi-square test for independence (605)
phi-coefficient (613)
Cramér's V (614)
median test (618)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter
You can find a tutorial quiz and other learning exercises for Chapter 17 on the book companion website. The website also provides access to a workshop entitled *Chi-Square* that reviews the chi-square tests presented in this chapter.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Chi-Square Tests for Goodness of Fit and for Independence** that are presented in this chapter.

The Chi-Square Test for Goodness of Fit

Data Entry

1. Enter the set of observed frequencies in the first column of the SPSS data editor. If there are four categories, for example, enter the four observed frequencies.
2. In the second column, enter the numbers 1, 2, 3, and so on, so that there is a number beside each of the observed frequencies in the first column.

Data Analysis

1. Click **Data** on the tool bar at the top of the page and select **weight cases** at the bottom of the list.
2. Click the **Weight cases by** circle, then highlight the label for the column containing the observed frequencies (VAR00001) on the left and move it into the **Frequency Variable** box by clicking on the arrow.
3. Click **OK.**
4. Click **Analyze** on the tool bar, select **Nonparametric Tests,** and click on **Chi-Square.**
5. Highlight the label for the column containing the digits 1, 2, 3, and move it into the Test Variables box by clicking on the arrow.
6. To specify the expected frequencies, you can either use the **all categories equal** option, which automatically computes expected frequencies, or you can enter your own values. To enter your own expected frequencies, click on the **values** option, and, one by one, enter the expected frequencies into the small box and click **Add** to add each new value to the bottom of the list.
7. Click **OK.**

SPSS Output

The program produces a table showing the complete set of observed and expected frequencies. A second table provides the value for the chi-square statistic, the degrees of freedom, and the level of significance (the p value, or alpha level, for the test).

The Chi-Square Test for Independence

Data Entry

1. Enter the complete set of observed frequencies in one column of the SPSS data editor (VAR00001).
2. In a second column, enter a number (1, 2, 3, etc.) that identifies the row corresponding to each observed frequency. For example, enter a 1 beside each observed frequency that came from the first row.
3. In a third column, enter a number (1, 2, 3, etc.) that identifies the column corresponding to each observed frequency. Each value from the first column gets a 1, and so on.

Data Analysis

1. Click **Data** on the tool bar at the top of the page and select **weight cases** at the bottom of the list.
2. Click the **Weight cases by** circle, then highlight the label for the column containing the observed frequencies (VAR00001) on the left and move it into the **Frequency Variable** box by clicking on the arrow.

3. Click **OK.**
4. Click **Analyze** on the tool bar at the top of the page, select **Descriptive Statistics,** and click on **Crosstabs.**
5. Highlight the label for the column containing the rows (VAR00002) and move it into the **Rows** box by clicking on the arrow.
6. Highlight the label for the column containing the columns (VAR00003) and move it into the **Columns** box by clicking on the arrow.
7. Click on **Statistics,** select **Chi-Square,** and click **Continue.**
8. Click **OK.**

SPSS Output

We used SPSS to conduct the chi-square test for independence for the data in Example 17.2, examining the relationship between gender and willingness to use mental health services, and the output is shown in Figure 17.5. The first table in the output simply lists the variables and is not shown in the figure. The **Crosstabulation** table simply shows the matrix of observed frequencies. The final table, labeled **Chi-Square Tests,** reports the results. Focus on the top row, the **Pearson Chi-Square,** which reports the calculated chi-square value, the degrees of freedom, and the level of significance (the *p* value, or the alpha level, for the test).

VAR00002*VAR00003 Crosstabulation

Count

		VAR00003			
		1.00	2.00	3.00	Total
VAR00002	1.00	17	32	11	60
	2.00	13	43	34	90
Total		30	75	45	150

Chi-Square Tests

	Value	df	Asymp.Sig. (2-sided)
Pearson Chi-Square	8.231[a]	2	.016
Likelihood Ratio	8.443	2	.015
Linear by Linear Association	8.109	1	.004
N of Valid Cases	150		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.00.

FIGURE 17.5

The SPSS output for the chi-square test for independence in Example 17.2.

FOCUS ON PROBLEM SOLVING

1. The expected frequencies that you calculate must satisfy the constraints of the sample. For the goodness-of-fit test, $\Sigma f_e = \Sigma f_o = n$. For the test for independence, the row totals and column totals for the expected frequencies should be identical to the corresponding totals for the observed frequencies.
2. It is entirely possible to have fractional (decimal) values for expected frequencies. Observed frequencies, however, are always whole numbers.
3. Whenever $df = 1$, the difference between observed and expected frequencies ($f_o = f_e$) is identical (the same value) for all cells. This makes the calculation of chi-square easier.
4. Although you are advised to compute expected frequencies for all categories (or cells), you should realize that it is not essential to calculate all f_e values separately. Remember that df for chi-square identifies the number of f_e values that are free to vary. Once you have calculated that number of f_e values, the remaining f_e values are determined. You can get these remaining values by subtracting the calculated f_e values from their corresponding row or column totals.
5. Remember that, unlike previous statistical tests, the degrees of freedom (df) for a chi-square test are *not* determined by the sample size (n). Be careful!

DEMONSTRATION 17.1

TEST FOR INDEPENDENCE

A manufacturer of watches would like to examine preferences for digital versus analog watches. A sample of $n = 200$ people is selected, and these individuals are classified by age and preference. The manufacturer would like to know whether there is a relationship between age and watch preference. The observed frequencies (f_o) are as follows:

	Digital	Analog	Undecided	Totals
Younger than 30	90	40	10	140
30 or Older	10	40	10	60
Column totals	100	80	20	$n = 200$

STEP 1 State the hypotheses, and select an alpha level.

The null hypothesis states that there is no relationship between the two variables.

> H_0: Preference is independent of age. That is, the frequency distribution of preference has the same form for people younger than 30 as for people 30 or older.

The alternative hypothesis states that there is a relationship between the two variables.

H_1: Preference is related to age. That is, the type of watch preferred depends on a person's age.

We set alpha to $\alpha = .05$.

STEP 2 Locate the critical region.

Degrees of freedom for the chi-square test for independence are determined by

$$df = (C - 1)(R - 1)$$

For these data,

$$df = (3 - 1)(2 - 1) = 2(1) = 2$$

For $df = 2$ with $\alpha = .05$, the critical chi-square value is 5.99. Thus, our obtained chi-square must exceed 5.99 to be in the critical region and to reject H_0.

STEP 3 Compute the test statistic. Two calculations are required: finding the expected frequencies and calculating the chi-square statistic.

Expected frequencies, f_e. For the test for independence, the expected frequencies can be found using the column totals (f_c), the row totals (f_r), and the following formula:

$$f_e = \frac{f_c f_r}{n}$$

For people younger than 30, we obtain the following expected frequencies:

$$f_e = \frac{100(140)}{200} = \frac{14{,}000}{200} = 70 \text{ for digital}$$

$$f_e = \frac{80(140)}{200} = \frac{11{,}200}{200} = 56 \text{ for analog}$$

$$f_e = \frac{20(140)}{200} = \frac{2800}{200} = 14 \text{ for undecided}$$

For individuals 30 or older, the expected frequencies are as follows:

$$f_e = \frac{100(60)}{200} = \frac{6000}{200} = 30 \text{ for digital}$$

$$f_e = \frac{80(60)}{200} = \frac{4800}{200} = 24 \text{ for analog}$$

$$f_e = \frac{20(60)}{200} = \frac{1200}{200} = 6 \text{ for undecided}$$

The following table summarizes the expected frequencies:

	Digital	Analog	Undecided
Younger than 30	70	56	14
30 or Older	30	24	6

The chi-square statistic. The chi-square statistic is computed from the formula

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

The following table summarizes the calculations:

Cell	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Younger than 30—digital	90	70	20	400	5.71
Younger than 30—analog	40	56	−16	256	4.57
Younger than 30—undecided	10	14	−4	16	1.14
30 or Older—digital	10	30	−20	400	13.33
30 or Older—analog	40	24	16	256	10.67
30 or Older—undecided	10	6	4	16	2.67

Finally, add the values in the last column to get the chi-square statistic.

$$\chi^2 = 5.71 + 4.57 + 1.14 + 13.33 + 10.67 + 2.67$$
$$= 38.09$$

STEP 4 Make a decision about H_0, and state the conclusion.

The chi-square value is in the critical region. Therefore, we reject the null hypothesis. There is a relationship between watch preference and age, $\chi^2(2, n = 200) = 38.09$, $p < .05$.

DEMONSTRATION 17.2

EFFECT SIZE WITH CRAMÉR'S *V*

Because the data matrix is larger than 2 × 2, we compute Cramér's V to measure effect size.

$$\text{Cramér's } V = \sqrt{\frac{\chi^2}{n(df^*)}} = \sqrt{\frac{38.09}{200(1)}} = \sqrt{0.19} = 0.436$$

PROBLEMS

1. Parametric tests (such as t or ANOVA) differ from nonparametric tests (such as chi-square) primarily in terms of the assumptions they require and the data they use. Explain these differences.

2. The student population at the state college consists of 55% females and 45% males.

a. The college theater department recently staged a production of a modern musical. A researcher recorded the gender of each student entering the theater and found a total of 385 females and 215 males. Is the gender distribution for theater goers significantly different from the distribution for the general college? Test at the .05 level of significance.

b. The same researcher also recorded the gender of each student watching a men's basketball game in the college gym and found a total of 83 females and

97 males. Is the gender distribution for basketball fans significantly different from the distribution for the general college? Test at the .05 level of significance.

3. A developmental psychologist would like to determine whether infants display any color preferences. A stimulus consisting of four color patches (red, green, blue, and yellow) is projected onto the ceiling above a crib. Infants are placed in the crib, one at a time, and the psychologist records how much time each infant spends looking at each of the four colors. The color that receives the most attention during a 100-second test period is identified as the preferred color for that infant. The preferred colors for a sample of 60 infants are shown in the following table:

Red	Green	Blue	Yellow
20	12	18	10

a. Do the data indicate any significant preferences among the four colors? Test at the .05 level of significance.

b. Write a sentence demonstrating how the outcome of the hypothesis test would appear in a research report.

4. Data from the department of motor vehicles indicate that 80% of all licensed drivers are older than age 25.

a. In a sample of $n = 60$ people who recently received speeding tickets, 38 were older than 25 years and the other 22 were age 25 or younger. Is the age distribution for this sample significantly different from the distribution for the population of licensed drivers? Use $\alpha = .05$.

b. In a sample of $n = 60$ people who recently received parking tickets, 43 were older than 25 years and the other 17 were age 25 or younger. Is the age distribution for this sample significantly different from the distribution for the population of licensed drivers? Use $\alpha = .05$.

5. To investigate the phenomenon of "home-team advantage," a researcher recorded the outcomes from 64 college football games on one Saturday in October. Of the 64 games, 42 were won by home teams. Does this result provide enough evidence to conclude that home teams win significantly more than would be expected by chance? Assume that winning and losing are equally likely events if there is no home-team advantage. Use $\alpha = .05$.

6. Research has demonstrated that people tend to be attracted to others who are similar to themselves. One study demonstrated that individuals are disproportionately more likely to marry those with surnames that begin with the same last letter as their own (Jones, Pelham, Carvallo, & Mirenberg, 2004). The researchers began by looking at marriage records and recording the surname for each groom and the maiden name of each bride. From these records it is possible to calculate the probability of randomly matching a bride and a groom whose last names begin with the same letter. Suppose that this probability is only 6.5%. Next, a sample of $n = 200$ married couples is selected and the number who shared the same last initial at the time they were married is counted. The resulting observed frequencies are as follows:

Same Initial	Different Initials	
19	181	200

Do these data indicate that the number of couples with the same last initial is significantly different that would be expected if couples were matched randomly? Test with $\alpha = .05$.

7. Suppose that the researcher from the previous problem repeated the study of married couples' initials using twice as many participants and obtaining observed frequencies that exactly double the original values. The resulting data are as follows:

Same Initial	Different Initials	
38	362	400

a. Use a chi-square test to determine whether the number of couples with the same last initial is significantly different than would be expected if couples were matched randomly. Test with $\alpha = .05$.

b. You should find that the data lead to rejecting the null hypothesis. However, in problem 6 the decision was fail to reject. How do you explain the fact that the two samples have the same proportions but lead to different conclusions?

8. A professor in the psychology department would like to determine whether there has been a significant change in grading practices over the years. It is known that the overall grade distribution for the department in 1985 had 14% As, 26% Bs, 31% Cs, 19% Ds, and 10% Fs. A sample of $n = 200$ psychology students from last semester produced the following grade distribution:

A	B	C	D	F
32	61	64	31	12

Do the data indicate a significant change in the grade distribution? Test at the .05 level of significance.

9. Automobile insurance is much more expensive for teenage drivers than for older drivers. To justify this cost difference, insurance companies claim that the younger drivers are much more likely to be involved in costly accidents. To test this claim, a researcher obtains information about registered drivers from the department of motor vehicles (DMV) and selects a sample of $n = 300$ accident reports from the police department. The DMV reports the percentage of registered drivers in each age category as follows: 16% are younger than age 20; 28% are 20 to 29 years old; and 56% are age 30 or older. The number of accident reports for each age group is as follows:

Under age 20	Age 20–29	Age 30 or older
68	92	140

a. Do the data indicate that the distribution of accidents for the three age groups is significantly different from the distribution of drivers? Test with $\alpha = .05$.

b. Write a sentence demonstrating how the outcome of the hypothesis test would appear in a research report.

10. The color red is often associated with anger and male dominance. Based on this observation, Hill and Barton (2005) monitored the outcome of four combat sports (boxing, tae kwan do, Greco-Roman wrestling, and freestyle wrestling) during the 2004 Olympic games and found that participants wearing red outfits won significantly more often than those wearing blue.

a. In 50 wrestling matches involving red versus blue, suppose that the red outfit won 31 times and lost 19 times. Is this result sufficient to conclude that red wins significantly more than would be expected by chance? Test at the .05 level of significance.

b. In 100 matches, suppose red won 62 times and lost 38. Is this sufficient to conclude that red wins significantly more than would be expected by chance? Again, use $\alpha = .05$.

c. Note that the winning percentage for red uniforms in part a is identical to the percentage in part b (31 out of 50 is 62%, and 62 out of 100 is also 62%). Although the two samples have an identical winning percentage, one is significant and the other is not. Explain why the two samples lead to different conclusions.

11. A communications company has developed three new designs for a cell phone. To evaluate consumer response, a sample of 120 college students is selected and each student is given all three phones to use for 1 week. At the end of the week, the students must identify which of the three designs they prefer. The distribution of preference is as follows:

Design 1	Design 2	Design 3
54	38	28

Do the results indicate any significant preferences among the three designs?

12. In problem 11, a researcher asked college students to evaluate three new cell phone designs. However, the researcher suspects that college students may have criteria that are different from those used by older adults. To test this hypothesis, the researcher repeats the study using a sample of $n = 60$ older adults in addition to a sample of $n = 60$ students. The distribution of preference is as follows:

	Design 1	Design 2	Design 3	
Student	27	20	13	60
Older Adult	21	34	5	60
	48	54	18	

Do the data indicate that the distribution of preferences for older adults is significantly different from the distribution for college students? Test with $\alpha = .05$.

13. Research suggests that romantic background music increases the likelihood that a woman will give her phone number to a man she has just met (Guéguen & Jacoby, 2010). In the study, women spent time in a waiting room with background music playing. In one condition, the music was a popular love song and for the other condition the music was a neutral song. The participant was then moved to another room in which she was instructed to discuss two food products with a young man. The men were part of the study and were selected because they had been rated as average in attractiveness. The experimenter returned to end the study and asked the pair to wait alone for a few minutes. During this time, the man used a scripted line to ask the woman for her phone number. The following table presents data similar to those obtained in the study, showing the number of women who did or did not give their numbers for each music condition.

	Phone Number	No Number	
Romantic Music	21	19	40
Neutral Music	9	31	40
	30	50	

Is there a significant difference between the two types of music? Test with $\alpha = .05$

14. Mulvihill, Obuseh, and Caldwell (2008) conducted a survey evaluating healthcare providers' perception of a new state children's insurance program. One question

asked the providers whether they viewed the reimbursement from the new insurance as higher, lower, or the same as private insurance. Another question assessed the providers' overall satisfaction with the new insurance. The following table presents observed frequencies similar to the study results.

	Satisfied	Not Satisfied	
Less Reimbursement	46	54	100
Same or More Reimbursement	42	18	60
	88	72	

Do the results indicate that the providers' satisfaction of the new program is related to their perception of the reimbursement rates? Test with $\alpha = .05$.

15. A local county is considering a budget proposal that would allocate extra funding toward the renovation of city parks. A survey is conducted to measure public opinion concerning the proposal. A total of 150 individuals respond to the survey: 50 who live within the city limits and 100 from the surrounding suburbs. The frequency distribution is as follows:

	Opinion: Favor	Opinion: Oppose	
City	35	15	50
Suburb	55	45	100
	90	60	

a. Is there a significant difference in the distribution of opinions for city residents compared to those in the suburbs? Test at the .05 level of significance.

b. The relationship between home location and opinion can also be evaluated using the phi-coefficient. If the phi-coefficient were computed for these data, what value would be obtained for phi?

16. The data from problem 15 show no significant difference between the opinions for city residents and those who live in the suburbs. To construct the following data, we simply doubled the sample size from problem 15 so that all of the individual frequencies are twice as big. Notice that the sample *proportions* have not changed.

	Opinion: Favor	Opinion: Oppose	
City	70	30	100
Suburb	110	90	200
	180	120	

a. Test for a significant difference between the city distribution and the suburb distribution using $\alpha = .05$. How does the decision compare with the decision in problem 14? You should find that a larger sample increases the likelihood of a significant result.

b. Compute the phi-coefficient for these data and compare it with the result from problem 15. You should find that the sample size has no effect on the strength of the relationship.

17. In the Preview for this chapter, we discussed a study investigating the relationship between memory for eyewitnesses and the questions they are asked (Loftus & Palmer, 1974). In the study, participants watched a film of an automobile accident and then were questioned about the accident. One group was asked how fast the cars were going when they "smashed into" each other. A second group was asked about the speed when the cars "hit" each other, and a third group was not asked any question about the speed of the cars. A week later, the participants returned to answer additional questions about the accident, including whether they recalled seeing any broken glass. Although there was no broken glass in the film, several students claimed to remember seeing it. The following table shows the frequency distribution of responses for each group.

Verb Used to Ask About the Speed of the Cars	Response to the Question "Did You See Any Broken Glass?": Yes	Response to the Question "Did You See Any Broken Glass?": No
Smashed into	16	34
Hit	7	43
Control (not asked)	6	44

a. Does the proportion of participants who claim to remember broken glass differ significantly from group to group? Test with $\alpha = .05$.

b. Compute Cramérs *V* to measure the size of the treatment effect.

c. Describe how the phrasing of the question influenced the participants' memories.

d. Write a sentence demonstrating how the outcome of the hypothesis test and the measure of effect size would be reported in a journal article.

18. In a study investigating freshman weight gain, the researchers also looked at gender differences in weight (Kasparek, Corwin, Valois, Sargent, & Morris, 2008). Using self-reported heights and weights, they computed the body mass index (BMI) for each student. Based on the BMI scores, the students were classified as either desirable weight or overweight. When the students were further classified by gender,

the researchers found results similar to the frequencies in the following table.

	Desirable Weight	Overweight
Males	74	46
Females	62	18

a. Do the data indicate that the proportion of overweight men is significantly different from the proportion of overweight women? Test with $\alpha = .05$.
b. Compute the phi-coefficient to measure the strength of the relationship.
c. Write a sentence demonstrating how the outcome of the hypothesis test and the measure of effect size would be reported in a journal article.

19. Research results suggest that IQ scores for boys are more variable than IQ scores for girls (Arden & Plomin, 2006). A typical study looking at 10-year-old children classifies participants by gender and by low, average, or high IQ. Following are hypothetical data representing the research results. Do the data indicate a significant difference between the frequency distributions for males and females? Test at the .05 level of significance and describe the difference.

	IQ			
	Low	Average	High	
Boys	18	42	20	80
Girls	12	54	14	80
				$n = 160$

20. Gender differences in dream content are well documented (see Winget & Kramer, 1979). Suppose a researcher studies aggression content in the dreams of men and women. Each participant reports his or her most recent dream. Then each dream is judged by a panel of experts to have low, medium, or high aggression content. The observed frequencies are shown in the following matrix:

		Aggression Content		
		Low	Medium	High
Gender	Female	18	4	2
	Male	4	17	15

Is there a relationship between gender and the aggression content of dreams? Test with $\alpha = .01$.

21. In a study similar to one conducted by Fallon and Rozin (1985), a psychologist prepared a set of silhouettes showing different female body shapes ranging from somewhat thin to somewhat heavy and asked a group of women to indicate which body figure they thought men would consider the most attractive. Then a group of men were shown the same set of profiles and asked which image they considered the most attractive. The following hypothetical data show the number of individuals who selected each of the four body image profiles.
a. Do the data indicate a significant difference between the actual preferences for the men and the preferences predicted by the women? Test at the .05 level of significance.
b. Compute the phi-coefficient to measure the strength of the relationship.

	Body Image Profiles				
	Somewhat Thin	Slightly Thin	Slightly Heavy	Somewhat Heavy	
Women	29	25	18	8	80
Men	11	15	22	12	60
	40	40	40	20	

22. A recent study indicates that people tend to select video game avatars with characteristics similar to those of their creators (Bélisle & Onur, 2010). Participants who had created avatars for a virtual community game completed a questionnaire about their personalities. An independent group of viewers examined the avatars and recorded their impressions of the avatars. One personality characteristic considered was introverted/extroverted. The following frequency distribution of personalities for participants and the avatars they created.

	Participant Personality		
	Introverted	Extroverted	
Introverted Avatar	22	23	45
Extroverted Avatar	16	39	55
	38	62	

a. Is there a significant relationship between the personalities of the participants and the personalities of their avatars? Test with $\alpha = .05$.
b. Compute the phi-coefficient to measure the size of the effect.

23. Research indicates that people who volunteer to participate in research studies tend to have higher intelligence than nonvolunteers. To test this

phenomenon, a researcher obtains a sample of 200 high school students. The students are given a description of a psychological research study and asked whether they would volunteer to participate. The researcher also obtains an IQ score for each student and classifies the students into high, medium, and low IQ groups. Do the following data indicate a significant relationship between IQ and volunteering? Test at the .05 level of significance.

	IQ			
	High	Medium	Low	
Volunteer	43	73	34	150
Not Volunteer	7	27	16	50
	50	100	50	

24. Cialdini, Reno, and Kallgren (1990) examined how people conform to norms concerning littering. The researchers wanted to determine whether a person's tendency to litter depended on the amount of litter already in the area. People were handed a handbill as they entered an amusement park. The entrance area had already been prepared with either no litter, a small amount of litter, or a lot of litter lying on the ground. The people were observed to determine whether they dropped their handbills. The frequency data are as follows:

	Amount of Litter		
	None	Small Amount	Large Amount
Littering	17	28	49
Not Littering	73	62	41

a. Do the data indicate that people's tendency to litter depends on the amount of litter already on the ground? That is, is there a significant relationship between littering and the amount of existing litter? Test at the .05 level of significance.

b. Compute Cramér's V to measure the size of the treatment effect.

25. Although the phenomenon is not well understood, it appears that people born during the winter months are slightly more likely to develop schizophrenia than people born at other times (Bradbury & Miller, 1985). The following hypothetical data represent a sample of 50 individuals diagnosed with schizophrenia and a sample of 100 people with no psychotic diagnosis. Each individual is also classified according to season in which he or she was born. Do the data indicate a significant relationship between schizophrenia and the season of birth? Test at the .05 level of significance.

	Season of Birth				
	Summer	Fall	Winter	Spring	
No Disorder	26	24	22	28	100
Schizophrenia	9	11	18	12	50
	35	35	40	40	

This page intentionally left blank

CHAPTER

18

The Binomial Test

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Binomial distribution (Chapter 6)
- *z*-score hypothesis tests (Chapter 8)
- Chi-square test for goodness of fit (Chapter 17)

Preview

In 1960, Gibson and Walk designed a classic piece of apparatus to test depth perception. Their device, called a *visual cliff*, consisted of a wide board with a deep drop (the cliff) to one side and a shallow drop on the other side. An infant was placed on the board and then observed to see whether he or she crawled off the shallow side or crawled off the cliff. Infants who moved to the deep side actually crawled onto a sheet of heavy glass, which prevented them from falling. Thus, the deep side only appeared to be a cliff—hence the name *visual cliff*.

Gibson and Walk reasoned that if infants are born with the ability to perceive depth, they would recognize the deep side and not crawl off the cliff. On the other hand, if depth perception is a skill that develops over time through learning and experience, then infants should not be able to perceive any difference between the shallow and the deep sides.

Out of 27 infants who moved off the board, only 3 ventured onto the deep side at any time during the experiment. The other 24 infants stayed exclusively on the shallow side. Gibson and Walk interpreted these data as convincing evidence that depth perception is innate. The infants showed a systematic preference for the shallow side.

The Problem: You should notice immediately that the data from this experiment are different from the data that we usually encounter. There are no scores. Gibson and Walk simply counted the number of infants who went off the deep side and the number who went to the shallow side. Still, we would like to use these data to make statistical decisions. Do these sample data provide sufficient evidence to make a confident conclusion about depth perception in the population? Suppose that 8 of the 27 infants had crawled to the deep side. Would you still be convinced that there is a significant preference for the shallow side? What about 12 out of 27?

The Solution: We are asking a question about statistical significance and need a hypothesis test to obtain an answer. The null hypothesis for the Gibson and Walk study would state that infants have no depth perception and cannot perceive a difference between the shallow and deep sides. In this case, their movement should be random with half going to either side. Notice that the data and the hypothesis both concern frequencies or proportions. This situation is perfect for the chi-square test introduced in Chapter 17, and a chi-square test can be used to evaluate the data. However, when individuals are classified into exactly two categories (for example, shallow and deep), a special statistical procedure exists. In this chapter, we introduce the *binomial test*, which is used to evaluate and interpret frequency data involving exactly two categories of classification.

18.1 OVERVIEW

Data with exactly two categories are also known as dichotomous data.

In Chapter 6, we introduced the concept of *binomial data*. You should recall that binomial data exist whenever a measurement procedure classifies individuals into exactly two distinct categories. For example, the outcomes from tossing a coin can be classified as heads and tails; people can be classified as male or female; plastic products can be classified as recyclable or non-recyclable. In general, binomial data exist when

1. The measurement scale consists of exactly two categories.
2. Each individual observation in a sample is classified in only one of the two categories.
3. The sample data consist of the frequency of, or number of individuals in, each category.

The traditional notation system for binomial data identifies the two categories as A and B and identifies the probability (or proportion) associated with each category as p and q, respectively. For example, a coin toss results in either heads (A) or tails (B), with probabilities $p = \frac{1}{2}$ and $q = \frac{1}{2}$.

In this chapter, we examine the statistical process of using binomial data for testing hypotheses about the values of p and q for the population. This type of hypothesis test is called a *binomial test*.

DEFINITION

A **binomial test** uses sample data to evaluate hypotheses about the values of p and q for a population consisting of binomial data.

Consider the following two situations:

1. In a sample of $n = 34$ color-blind students, 30 are male, and only 4 are female. Does this sample indicate that color blindness is significantly more common for males in the general population?
2. In 2005, only 10% of American families had incomes below the poverty level. This year, in a sample of 100 families, 19 were below the poverty level. Does this sample indicate that there has been a significant change in the population proportions?

Notice that both of these examples have binomial data (exactly two categories). Although the data are relatively simple, we are asking a statistical question about significance that is appropriate for a hypothesis test: Do the sample data provide sufficient evidence to make a conclusion about the population?

HYPOTHESES FOR THE BINOMIAL TEST

In the binomial test, the null hypothesis specifies exact values for the population proportions p and q. Theoretically, you could choose any proportions for H_0, but usually there is a clear reason for the values that are selected. The null hypothesis typically falls into one of the following two categories:

1. **Just Chance.** Often the null hypothesis states that the two outcomes, A and B, occur in the population with the proportions that would be predicted simply by chance. If you were tossing a coin, for example, the null hypothesis might specify $p(\text{heads}) = \frac{1}{2}$ and $p(\text{tails}) = \frac{1}{2}$. Notice that this hypothesis states the usual, chance proportions for a balanced coin. Also notice that it is not necessary to specify both proportions. Once the value of p is identified, the value of q is determined by $1 - p$. For the coin toss example, the null hypothesis would simply state

   H_0: $p = p(\text{heads}) = \dfrac{1}{2}$ (The coin is balanced.)

   Similarly, if you were selecting cards from a deck and trying to predict the suit on each draw, the probability of predicting correctly would be $p = \frac{1}{4}$ for any given trial. (With four suits, you have a 1-out-of-4 chance of guessing correctly.) In this case, the null hypothesis would state

   H_0: $p = p(\text{guessing correctly}) = \dfrac{1}{4}$ (The outcome is simply the result of chance.)

   In each case, the null hypothesis states that there is nothing unusual about the proportions in the population; that is, the outcomes are occurring by chance.
2. **No Change or No Difference.** Often you may know the proportions for one population and want to determine whether the same proportions apply to a different population. In this case, the null hypothesis would simply specify that there is no difference between the two populations. Suppose that national

statistics indicate that 1 out of 12 drivers will be involved in a traffic accident during the next year. Does this same proportion apply to 16-year-olds who are driving for the first time? According to the null hypothesis,

H_0: For 16-year-olds, $p = p(\text{accident}) = \frac{1}{12}$ (Not different from the general population)

Similarly, suppose that last year, 30% of the freshman class failed the college writing test. This year, the college is requiring all freshmen to take a writing course. Will the course have any effect on the number who fail the test? According to the null hypothesis,

H_0: For this year, $p = p(\text{fail}) = 30\%$ (Not different from last year's class)

THE DATA FOR THE BINOMIAL TEST

For the binomial test, a sample of n individuals is obtained and you simply count how many are classified in category A and how many are classified in category B. We focus attention on category A and use the symbol X to stand for the number of individuals classified in category A. Recall from Chapter 6 that X can have any value from 0 to n and that each value of X has a specific probability. The distribution of probabilities for each value of X is called the *binomial distribution*. Figure 18.1 shows an example of a binomial distribution for which X is the number of heads obtained in four tosses of a balanced coin.

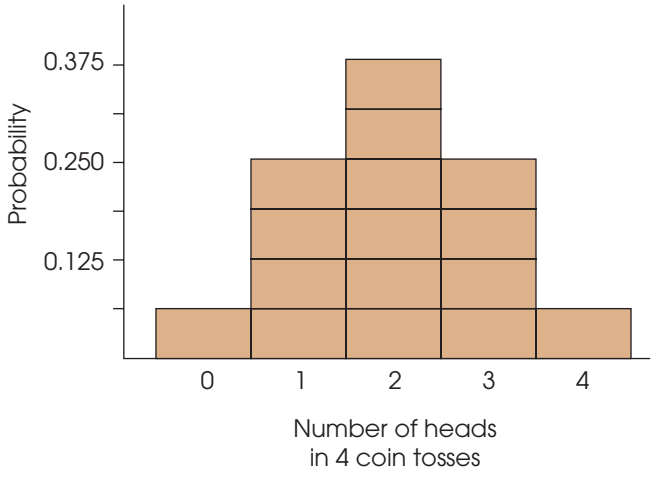
THE TEST STATISTIC FOR THE BINOMIAL TEST

As we noted in Chapter 6, when the values pn and qn are both equal to or greater than 10, the binomial distribution approximates a normal distribution. This fact is important because it allows us to compute z-scores and use the unit normal table to answer probability questions about binomial events. In particular, when pn and qn are both at least 10, the binomial distribution has the following properties:

1. The shape of the distribution is approximately normal.
2. The mean of the distribution is $\mu = pn$.

FIGURE 18.1

A binomial distribution for the number of heads obtained in four tosses of a balanced coin.



3. The standard deviation of the distribution is

$$\sigma = \sqrt{npq}$$

With these parameters in mind, it is possible to compute a z-score corresponding to each value of X in the binomial distribution.

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}} \qquad \text{(See Equation 6.3.)} \tag{18.1}$$

This is the basic z-score formula that is used for the binomial test. However, the formula can be modified slightly to make it more compatible with the logic of the binomial hypothesis test. The modification consists of dividing both the numerator and the denominator of the z-score by n. (You should realize that dividing both the numerator and the denominator by the same value does not change the value of the z-score.) The resulting equation is

$$z = \frac{X/n - p}{\sqrt{pq/n}} \tag{18.2}$$

For the binomial test, the values in this formula are defined as follows:

1. X/n is the proportion of individuals in the sample who are classified in category A.
2. p is the hypothesized value (from H_0) for the proportion of individuals in the population who are classified in category A.
3. $\sqrt{pq/n}$ is the standard error for the sampling distribution of X/n and provides a measure of the standard distance between the sample statistic (X/n) and the population parameter (p).

Thus, the structure of the binomial z-score (Equation 18.2) can be expressed as

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{\begin{matrix}\text{sample} \\ \text{proportion} \\ \text{(data)}\end{matrix} - \begin{matrix}\text{hypothesized} \\ \text{population} \\ \text{proportion}\end{matrix}}{\text{standard error}}$$

The logic underlying the binomial test is exactly the same as we encountered with the original z-score hypothesis test in Chapter 8. The hypothesis test involves comparing the sample data with the hypothesis. If the data are consistent with the hypothesis, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the hypothesis, then we reject the hypothesis. The value of the standard error provides a benchmark for determining whether the discrepancy between the data and the hypothesis is more than would be expected by chance. The alpha level for the test provides a criterion for deciding whether the discrepancy is significant. The hypothesis-testing procedure is demonstrated in the following section.

LEARNING CHECK

1. In the Preview, we described a research study using a visual cliff. State the null hypothesis for this study in words and as a probability value (p) that an infant will crawl off the deep side.
2. If the visual cliff study had used a sample of $n = 15$ infants, would it be appropriate to use the normal approximation to the binomial distribution? Explain why or why not.

3. If the results from the visual cliff study showed that 9 out of 36 infants crawled off the deep side, what z-score value would be obtained using Equation 18.1?

ANSWERS

1. The null hypothesis states that the probability of choosing between the deep side and the shallow side is just chance: $p(\text{deep side}) = \frac{1}{2}$.

2. The normal approximation to the binomial distribution requires that both pn and qn are at least 10. With $n = 15$, $pn = qn = 7.5$. The normal approximation should not be used.

3. With $n = 36$ and $p = \frac{1}{2}$, the binomial distribution has $\mu = \frac{1}{2}(36) = 18$, and $\sigma = \sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)(36)} = \sqrt{9} = 3$. $X = 9$ corresponds $z = -9/3 = -3.00$

18.2 THE BINOMIAL TEST

The binomial test follows the same four-step procedure presented earlier with other examples for hypothesis testing. The four steps are summarized as follows.

STEP 1 *State the hypotheses*. In the binomial test, the null hypothesis specifies values for the population proportions p and q. Typically, H_0 specifies a value only for p, the proportion associated with category A. The value of q is directly determined from p by the relationship $q = 1 - p$. Finally, you should realize that the hypothesis, as always, addresses the probabilities or proportions for the *population*. Although we use a sample to test the hypothesis, the hypothesis itself always concerns a population.

STEP 2 *Locate the critical region*. When both values for pn and qn are greater than or equal to 10, then the z-scores defined by Equation 18.1 or 18.2 form an approximately normal distribution. Thus, the unit normal table can be used to find the boundaries for the critical region. With $\alpha = .05$, for example, you may recall that the critical region is defined as z-score values greater than $+1.96$ or less than -1.96.

STEP 3 *Compute the test statistic (z-score)*. At this time, you obtain a sample of n individuals (or events) and count the number of times category A occurs in the sample. The number of occurrences of A in the sample is the X value for Equation 18.1 or 18.2. Because the two z-score equations are equivalent, you may use either one for the hypothesis test. Usually Equation 18.1 is easier to use because it involves larger numbers (fewer decimals) and it is less likely to be affected by rounding error.

STEP 4 *Make a decision*. If the z-score for the sample data is in the critical region, then you reject H_0 and conclude that the discrepancy between the sample proportions and the hypothesized population proportions is significantly greater than chance. That is, the data are not consistent with the null hypothesis, so H_0 must be wrong. On the other hand, if the z-score is not in the critical region, then you fail to reject H_0.

The following example demonstrates a complete binomial test.

EXAMPLE 18.1

In the Preview section, we described the *visual cliff* experiment designed to examine depth perception in infants. To summarize briefly, an infant is placed on a wide board that appears to have a deep drop on one side and a relatively shallow drop on the

other. An infant who is able to perceive depth should avoid the deep side and move toward the shallow side. Without depth perception, the infant should show no preference between the two sides. Of the 27 infants in the experiment, 24 stayed exclusively on the shallow side and only 3 moved onto the deep side. The purpose of the hypothesis test is to determine whether these data demonstrate that infants have a significant preference for the shallow side.

This is a binomial hypothesis-testing situation. The two categories are

A = move onto the deep side

B = move onto the shallow side

STEP 1 The null hypothesis states that, for the general population of infants, there is no preference between the deep and the shallow sides; the direction of movement is determined by chance. In symbols,

$$H_0: \quad p = p(\text{deep side}) = \frac{1}{2} \quad \left(\text{and } q = \frac{1}{2}\right)$$

$$H_1: \quad p \neq \frac{1}{2} \quad (\text{There is a preference.})$$

We use $\alpha = .05$.

STEP 2 With a sample of $n = 27$, $pn = 13.5$ and $qn = 13.5$. Both values are greater than 10, so the distribution of z-scores is approximately normal. With $\alpha = .05$, the critical region is determined by boundaries of $z = \pm 1.96$.

STEP 3 For this experiment, the data consist of $X = 3$ out of $n = 27$. Using Equation 18.1, these data produce a z-score value of

$$z = \frac{X - pn}{\sqrt{npq}} = \frac{3 - 13.5}{\sqrt{27\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{-10.5}{2.60} = -4.04$$

To use Equation 18.2, you first compute the sample proportion, $X/n = 3/27 = 0.111$. The z-score is then

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{0.111 - 0.5}{\sqrt{\frac{1}{2}\left(\frac{1}{2}\right)/27}} = \frac{-0.389}{0.096} = -4.05$$

Within rounding error, the two equations produce the same result.

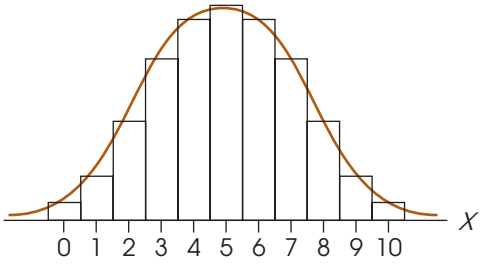
STEP 4 Because the data are in the critical region, our decision is to reject H_0. These data do provide sufficient evidence to conclude that there is a significant preference for the shallow side. Gibson and Walk (1960) interpreted these data as convincing evidence that depth perception is innate.

REAL LIMITS AND THE BINOMIAL TEST

In Chapter 6, we noted that a binomial distribution forms a discrete histogram (see Figure 18.1), whereas the normal distribution is a continuous curve. The difference between the two distributions was illustrated in Figure 6.18,

FIGURE 18.2

The relationship between the binomial distribution and the normal distribution. The binomial distribution is always a discrete histogram, and the normal distribution is a continuous, smooth curve. Each X value is represented by a bar in the histogram or a section of the normal distribution.



which is repeated here as Figure 18.2. In the figure, note that each score in the binomial distribution is represented by a bar in the histogram. For example, a score of $X = 6$ actually corresponds to a bar that reaches from a lower real limit of $X = 5.5$ to an upper real limit of 6.5.

When conducting a hypothesis test with the binomial distribution, the basic question is whether a specific score is located in the critical region. However, because each score actually corresponds to an interval, it is possible that part of the score is in the critical region and part is not. Fortunately, this is usually not an issue. When pn and qn are both equal to or greater than 10 (the criteria for using the normal approximation), each interval in the binomial distribution is extremely small and it is very unlikely that the interval overlaps the critical boundary. For example, the experiment in Example 18.1 produced a score of $X = 3$, and we computed a z-score of $z = -4.04$. Because this value is in the critical region, beyond $z = -1.96$, we rejected H_0. If we had used the real limit boundaries of $X = 2.5$ and $X = 3.5$, instead of $X = 3$, we would have obtained z-scores of

$$z = \frac{2.5 - 13.5}{2.60} \qquad \text{and} \qquad z = \frac{3.5 - 13.5}{2.60}$$

$$= -4.23 \qquad = -3.85$$

Thus, a score of $X = 3$ actually corresponds to an interval of z-scores ranging from $z = -3.85$ to $z = -4.23$. However, this entire interval is in the critical region beyond $z = -1.96$, so the decision is still to reject H_0.

In most situations, if the whole number X value (in this case, $X = 3$) is in the critical region, then the entire interval is in the critical region and the correct decision is to reject H_0. The only exception to this general rule occurs when an X value produces a z-score that is barely past the boundary into the critical region. In this situation, you should compute the z-scores corresponding to both real limits to determine whether any part of the z-score interval is not located in the critical region. Suppose, for example, that the researchers in Example 18.1 found that 8 out of 27 infants in the visual cliff experiment moved onto the deep side. A score of $X = 8$ corresponds to

$$z = \frac{8 - 13.5}{2.60}$$

$$z = \frac{-5.5}{2.60}$$

$$= -2.12$$

Because this value is beyond the − 1.96 boundary, it appears that we should reject H_0. However, this z-score is only slightly beyond the critical boundary, so it would be wise to check both ends of the interval. For $X = 8$, the real-limit boundaries are 7.5 and 8.5, which correspond to z-scores of

$$z = \frac{7.5 - 13.5}{2.60} = -2.31 \qquad \text{and} \qquad z = \frac{8.5 - 13.5}{2.60} = -1.92$$

Thus, a score of $X = 8$ corresponds to an interval extending from $z = -1.92$ to $z = -2.31$. However, the critical boundary is $z = -1.96$, which means that part of the interval (and part of the score) is not in the critical region for $\alpha = .05$. Because $X = 8$ is not completely beyond the critical boundary, the probability of obtaining $X = 8$ is greater than $\alpha = .05$. Therefore, the correct decision is to fail to reject H_0.

In general, it is safe to conduct a binomial test using the whole-number value for X. However, if you obtain a z-score that is only slightly beyond the critical boundary, you also should compute the z-scores for both real limits. If any part of the z-score interval is not in the critical region, the correct decision is to fail to reject H_0.

IN THE LITERATURE

REPORTING THE RESULTS OF A BINOMIAL TEST

Reporting the results of the binomial test typically consists of describing the data and reporting the z-score value and the probability that the results are caused by chance. It is also helpful to note that a binomial test was used because z-scores are used in other hypothesis-testing situations (see, for example, Chapter 8). For Example 18.1, the report might state:

> Three out of 27 infants moved to the deep side of the visual cliff. A binomial test revealed that there is a significant preference for the shallow side of the cliff, $z = -4.04$, $p < .05$.

Once again, p is *less than* .05. We have rejected the null hypothesis because it is very unlikely, probability less than 5%, that these results are simply caused by chance.

ASSUMPTIONS FOR THE BINOMIAL TEST

The binomial test requires two very simple assumptions:

1. The sample must consist of *independent* observations (see Chapter 8, page 254).
2. The values for pn and qn must both be greater than or equal to 10 to justify using the unit normal table for determining the critical region.

LEARNING CHECK

1. For a binomial test, the null hypothesis always states that $p = 1/2$. (True or false?)
2. The makers of brand X beer claim that people like their beer more than the leading brand. The basis for this claim is an experiment in which 64 beer drinkers compared the two brands in a side-by-side taste test. In this sample, 40 preferred brand X, and 24 preferred the leading brand.
   a. If you compute the z-score for $X = 40$, do these data support the claim that there is a significant preference? Test at the .05 level.
   b. If you compute z-scores for the real limits for $X = 40$, do the data support the claim that there is a significant preference? Test at the .05 level.

ANSWERS

1. False.

2. a. H_0: $p = \frac{1}{2} = q$, $X = 38$, $\mu = 32$, $\sigma = 4$, $z = 2.00$, reject H_0. Conclude that there is a significant preference.

b. The real limits of 39.5 and 40.5 correspond to z-scores of 1.88 and 2.13. The entire interval is not in the critical region so fail to reject H_0 and conclude that there is not a significant preference.

18.3 THE RELATIONSHIP BETWEEN CHI-SQUARE AND THE BINOMIAL TEST

You may have noticed that the binomial test evaluates the same basic hypotheses as the chi-square test for goodness of fit; that is, both tests evaluate how well the sample proportions fit a hypothesis about the population proportions. When an experiment produces binomial data, these two tests are equivalent, and either may be used. The relationship between the two tests can be expressed by the equation

$$\chi^2 = z^2$$

where χ^2 is the statistic from the chi-square test for goodness of fit and z is the z-score from the binomial test.

To demonstrate the relationship between the goodness-of-fit test and the binomial test, we reexamine the data from Example 18.1.

STEP 1 *Hypotheses.* In the visual cliff experiment from Example 18.1, the null hypothesis states that there is no preference between the shallow side and the deep side. For the binomial test, the null hypothesis states

$$H_0: \ p = p(\text{deep side}) = q = p(\text{shallow side}) = \frac{1}{2}$$

The chi-square test for goodness of fit would state the same hypothesis, specifying the population proportions as

	Shallow Side	Deep Side
H_0:	$\frac{1}{2}$	$\frac{1}{2}$

STEP 2 *Critical region.* For the binomial test, the critical region is located by using the unit normal table. With $\alpha = .05$, the critical region consists of any z-score value beyond ± 1.96. The chi-square test would have $df = 1$, and with $\alpha = .05$, the critical region consists of chi-square values greater than 3.84. Notice that the basic relationship, $\chi^2 = z^2$, holds:

$$3.84 = (1.96)^2$$

STEP 3 *Test statistic.* For the binomial test (Example 18.1), we obtained a z-score of $z = -4.04$. For the chi-square test, the expected frequencies are

	Shallow Side	Deep Side
f_e	13.5	13.5

With observed frequencies of 24 and 3, respectively, the chi-square statistic is

$$\chi^2 = \frac{(24 - 13.5)^2}{13.5} + \frac{(3 - 13.5)^2}{13.5}$$

$$= \frac{(10.5)^2}{13.5} + \frac{(-10.5)^2}{13.5}$$

$$= 8.167 + 8.167$$

$$= 16.33$$

With a little rounding error, the values obtained for the z-score and chi-square are related by the equation

Caution: The χ^2 value is already squared. Do not square it again.

$$\chi^2 = z^2$$

$$16.33 = (-4.04)^2$$

STEP 4 *Decision.* Because the critical values for both tests are related by the equation $\chi^2 = z^2$ and the test statistics are related in the same way, these two tests *always* result in the same statistical conclusion.

18.4 THE SIGN TEST

Although the binomial test can be used in many different situations, there is one specific application that merits special attention. For a repeated-measures study that compares two conditions, it is often possible to use a binomial test to evaluate the results. You should recall that a repeated-measures study involves measuring each individual in two different treatment conditions or at two different points in time. When the measurements produce numerical scores, the researcher can simply subtract to determine the difference between the two scores and then evaluate the data using a repeated-measures t test (see Chapter 11). Occasionally, however, a researcher may record only the *direction* of the difference between the two observations. For example, a clinician may observe patients before therapy and after therapy and simply note whether each patient got better or worse. Note that there is no measurement of how much change occurred; the clinician is simply recording the direction of change. Also note that the direction of change is a binomial variable; that is, there are only two values. In this situation it is possible to use a binomial test to evaluate the data. Traditionally, the two possible directions of change are coded by signs, with a positive sign indicating an increase and a negative sign indicating a decrease. When the binomial test is applied to signed data, it is called a *sign test*.

An example of signed data is shown in Table 18.1. Notice that the data can be summarized by saying that seven out of eight patients showed a decrease in symptoms after therapy.

The null hypothesis for the sign test states that there is no difference between the two treatments. Therefore, any change in a participant's score is the result of chance. In terms of probabilities, this means that increases and decreases are equally likely, so

$$p = p(\text{increase}) = \frac{1}{2}$$

$$q = p(\text{decrease}) = \frac{1}{2}$$

A complete example of a sign test follows.

TABLE 18.1

Hypothetical data from a research study evaluating the effectiveness of a clinical therapy. For each patient, symptoms are assessed before and after treatment and the data record whether there is an increase or a decrease in symptoms following therapy.

Patient	Direction of Change After Treatment
A	− (decrease)
B	− (decrease)
C	− (decrease)
D	+ (increase)
E	− (decrease)
F	− (decrease)
G	− (decrease)
H	− (decrease)

EXAMPLE 18.2

A researcher testing the effectiveness of acupuncture for treating the symptoms of arthritis obtains a sample of 36 people who have been diagnosed with arthritis. Each person's pain level is measured before treatment starts, and measured again after 4 months of acupuncture treatment. For this sample, 25 people experienced a reduction in pain, and 11 people had more pain after treatment. Do these data indicate a significant treatment effect?

STEP 1 State the hypothesis. The null hypothesis states that acupuncture has no effect. Any change in the level of pain is caused by chance, so increases and decreases are equally likely. Expressed as probabilities, the hypotheses are

$$H_0: \; p = p(\text{increased pain}) = \frac{1}{2} \text{ and } q = p(\text{decreased pain}) = \frac{1}{2}$$

H_1: $p \neq q$ (Changes tend to be consistently in one direction.)

Set $\alpha = .05$.

STEP 2 Locate the critical region. With $n = 36$ people, both pn and qn are greater than 10, so the normal approximation to the binomial distribution is appropriate. With $\alpha = .05$, the critical region consists of z-scores greater than $+1.96$ at one extreme and z-scores less than -1.96 at the other.

STEP 3 Compute the test statistic. For this sample we have $X = 25$ people with decreased pain. This score corresponds to a z-score of

$$z = \frac{X - pn}{\sqrt{npq}} = \frac{25 - 18}{\sqrt{36\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{7}{3} = 2.33$$

Because the z-score is only slightly beyond the 1.96 critical boundary, we consider the real limits for $X = 25$ to be certain that the entire interval is beyond the boundary. For $X = 25$, the real limits are 24.5 and 25.5, which correspond to z-scores of

$$z = \frac{24.5 - 18}{3} = 2.17 \quad \text{and} \quad z = \frac{25.5 - 18}{3} = 2.50$$

Thus, a score of $X = 25$ corresponds to an interval of z-scores ranging from $z = 2.17$ to $z = 2.50$. Note that this entire interval is beyond the 1.96 critical boundary.

STEP 4 Make a decision. Because the data are in the critical region, we reject H_0 and conclude that acupuncture treatment has a significant effect on arthritis pain, $z = 2.33$, $p < .05$.

ZERO DIFFERENCES IN THE SIGN TEST

You should notice that the null hypothesis in the sign test refers only to those individuals who show some difference between treatment 1 versus treatment 2. The null hypothesis states that if there is any change in an individual's score, then the probability of an increase is equal to the probability of a decrease. Stated in this form, the null hypothesis does not consider individuals who show zero difference between the two treatments. As a result, the usual recommendation is that these individuals be discarded from the data and the value of n be reduced accordingly. However, if the null hypothesis is interpreted more generally, it states that there is no difference between the two treatments. Phrased this way, it should be clear that individuals who show no difference actually are supporting the null hypothesis and should not be discarded. Therefore, an alternative approach to the sign test is to divide individuals who show zero differences equally between the positive and negative categories. (With an odd number of zero differences, discard one, and divide the rest evenly.) This alternative results in a more conservative test; that is, the test is more likely to fail to reject the null hypothesis.

EXAMPLE 18.3

It has been demonstrated that stress or exercise causes an increase in the concentration of certain chemicals in the brain called endorphins. Endorphins are similar to morphine and produce a generally relaxed feeling and a sense of well-being. The endorphins may explain the "high" experienced by long-distance runners. To demonstrate this phenomenon, a researcher tested pain tolerance for 40 athletes before and after they completed a mile run. Immediately after running, the ability to tolerate pain increased for 21 of the athletes, decreased for 12, and showed no change for the remaining 7.

Following the standard recommendation for handling zero differences, you would discard the 7 participants who showed no change and conduct a sign test with the remaining $n = 33$ athletes. With the more conservative approach, only 1 of the 7 who showed no difference would be discarded and the other 6 would be divided equally between the two categories. This would result in a total sample of $n = 39$ athletes with $21 + 3 = 24$ in the increased-tolerance category and $12 + 3 = 15$ in the decreased-tolerance category.

WHEN TO USE THE SIGN TEST

In many cases, data from a repeated-measures experiment can be evaluated using either a sign test or a repeated-measures t test. In general, you should use the t test whenever possible. Because the t test uses the actual difference scores (not just the signs), it makes maximum use of the available information and results in a more powerful test. However, there are some cases in which a t test cannot or should not be used, and in these situations, the sign test can be valuable. Four specific cases in which a t test is inappropriate or inconvenient are described below.

Before	After	Difference
20	23	+3
14	39	+25
27	Failed	+??
.	.	.
.	.	.
.	.	.

1. When you have infinite or undetermined scores, a *t* test is impossible, and the sign test is appropriate. Suppose, for example, that you are evaluating the effects of a sedative drug on problem-solving ability. A sample of rats is obtained, and each animal's performance is measured before and after receiving the drug. Hypothetical data are shown in the margin. Note that the third rat in this sample failed to solve the problem after receiving the drug. Because there is no score for this animal, it is impossible to compute a sample mean, an *SS*, or a *t* statistic. However, you could do a sign test because you know that the animal made more errors (an increase) after receiving the drug.
2. Often it is possible to describe the difference between two treatment conditions without precisely measuring a score in either condition. In a clinical setting, for example, a doctor can say whether a patient is improving, growing worse, or showing no change even though the patient's condition is not precisely measured by a score. In this situation, the data are sufficient for a sign test, but you could not compute a *t* statistic without individual scores.
3. Often a sign test is done as a preliminary check on an experiment before serious statistical analysis begins. For example, a researcher may predict that scores in treatment 2 should be consistently greater than scores in treatment 1. However, examination of the data after 1 week indicates that only 8 of 15 subjects showed the predicted increase. On the basis of these preliminary results, the researcher may choose to reevaluate the experiment before investing additional time.
4. Occasionally, the difference between treatments is not consistent across participants. This can create a very large variance for the difference scores. As we have noted in the past, large variance decreases the likelihood that a *t* test will produce a significant result. However, the sign test only considers the direction of each difference score and is not influenced by the variance of the scores.

LEARNING CHECK

1. A researcher used a chi-square test for goodness of fit to determine whether people had any preferences among three leading brands of potato chips. Could the researcher have used a binomial test instead of the chi-square test? Explain why or why not.
2. A researcher used a chi-square test to evaluate preferences between two logo designs for a minor-league hockey team. With a sample of $n = 100$ people, the researcher obtained a chi-square of 9.00. If a binomial test had been used instead of chi-square, what value would have been obtained for the *z*-score?
3. A developmental psychologist is using a behavior-modification program to help control the disruptive behavior of 40 children in a local school. After 1 month, 26 of the children have improved, 10 are worse, and 4 show no change in behavior. On the basis of these data, can the psychologist conclude that the program is working? Test at the .05 level.

ANSWERS

1. No, the binomial test cannot be used when there are three categories.
2. The *z*-score would be $\sqrt{9} = 3.00$.
3. Discarding the four participants who showed zero difference, $X = 26$ increases out of $n = 36$; $z = 2.67$; reject H_0; the program is working. If the four participants showing no change are divided between the two groups, then $X = 28$ out of $n = 40$; $z = 2.53$ and H_0 is still rejected.

SUMMARY

1. The binomial test is used with dichotomous data—that is, when each individual in the sample is classified in one of two categories. The two categories are identified as A and B, with probabilities of

$$p(A) = p \text{ and } p(B) = q$$

2. The binomial distribution gives the probability for each value of X, where X equals the number of occurrences of category A in a sample of n events. For example, X equals the number of heads in $n = 10$ tosses of a coin.

3. When pn and qn are both at least 10, then the binomial distribution is closely approximated by a normal distribution with

$$\mu = pn \qquad \sigma = \sqrt{npq}$$

   By using this normal approximation, each value of X has a corresponding z-score:

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}} \quad \text{or} \quad z = \frac{X/n - p}{\sqrt{pq/n}}$$

4. The binomial test uses sample data to test hypotheses about the binomial proportions, p and q, for a population. The null hypothesis specifies p and q, and the binomial distribution (or the normal approximation) is used to determine the critical region.

5. Usually the z-score in a binomial test is computed using the whole-number X value from the sample. However, if the z-score is only marginally in the critical region, you should compute the z-scores corresponding to both real limits of the score. If either one of these real-limit z-scores is not in the critical region, then the correct decision is to fail to reject the null hypothesis.

6. One common use of the binomial distribution is for the sign test. This test evaluates the difference between two treatments using the data from a repeated measures design. The difference scores are coded as being either increases (+) or decreases (−). Without a consistent treatment effect, the increases and decreases should be mixed randomly, so the null hypothesis states that

$$p(\text{increase}) = \frac{1}{2} = p(\text{decrease})$$

   With dichotomous data and hypothesized values for p and q, this is a binomial test.

KEY TERMS

binomial data (634)
binomial test (635)
binomial distribution (636)
sign test (643)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 18 on the book companion website.

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are found in Appendix D. If you are testing a null hypothesis specifying that $p = q = \frac{1}{2}$, then you can use SPSS to perform The Binomial Test presented in this chapter. Following are detailed instructions for the test. For other versions of the null hypothesis, use the equivalent chi-square test for goodness of fit presented in Chapter 17 (p. 594). The chi-square test allows you to specify expected frequencies, which is equivalent to specifying values for p and q.

Data Entry

1. Enter the category labels A and B in the first column of the SPSS data editor.
2. In the second column, enter the frequencies obtained for the two binomial categories. For example, if 21 out of 25 people were classified in category *A* (and only 4 people in category *B*), you would enter the values 21 and 4 in the second column.

Data Analysis

1. Click **Data** on the tool bar at the top of the page and select **weight cases** at the bottom of the list.
2. Click the **Weight cases by** circle, then highlight the label for the column containing the frequencies for the two categories and move it into the **Frequency Variable** box by clicking on the arrow.
3. Click **OK.**
4. Click **Analyze** on the tool bar, select **Nonparametric Tests,** and click on **One Sample.**
5. Select **Automatically compare observed data to hypothesis**.
6. Click **RUN.**

SPSS Output

We used SPSS to analyze the data from Example 18.1 and the output is shown in Figure 18.3. The output reports the null hypothesis for the test and the level of significance, which is rounded to .000 for this example.

FIGURE 18.3

The SPSS output for the binomial test in Example 18.1.

Hypothesis Test Summary

Null Hypothesis	Test	Sig.	Decision
The categories defined by VAR00001 = A and B occur with probabilities 0.5 and 0.5.	One-Sample Binomial Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

FOCUS ON PROBLEM SOLVING

1. For all binomial tests, the values of p and q must add up to 1.00 (or 100%).
2. Remember that both pn and qn must be at least 10 before you can use the normal distribution to determine critical values for a binomial test.

DEMONSTRATION 18.1

THE BINOMIAL TEST

The population of students in the psychology department at State College consists of 60% females and 40% males. Last semester, the Psychology of Gender course had a total of 36 students, of whom 26 were female and only 10 were male. Are the proportions of females and males in this class significantly different from what would be expected by chance from a population with 60% females and 40% males? Test at the .05 level of significance.

STEP 1 ***State the hypotheses, and specify alpha.*** The null hypothesis states that the male/female proportions for the class are not different from what is expected for a population with these proportions. In symbols,

$$H_0: \quad p = p(\text{female}) = 0.60 \text{ and } q = p(\text{male}) = 0.40$$

The alternative hypothesis is that the proportions for this class are different from what is expected for these population proportions.

$$H_1: \quad p \neq 0.60 \text{ (and } q \neq 0.40)$$

We set alpha at $\alpha = .05$.

STEP 2 ***Locate the critical region.*** Because pn and qn are both greater than 10, we can use the normal approximation to the binomial distribution. With $\alpha = .05$, the critical region is defined as a z-score value greater than -1.96 or less than -1.96.

STEP 3 ***Calculate the test statistic.*** The sample has 26 females out of 36 students, so the sample proportion is

$$\frac{X}{n} = \frac{26}{36} = 0.72$$

The corresponding z-score (using Equation 18.2) is

$$z = \frac{X/n - p}{\sqrt{pq/n}} = \frac{0.72 - 0.60}{\sqrt{\frac{0.60(0.40)}{36}}} = \frac{0.12}{0.0816} = 1.47$$

STEP 4 ***Make a decision about H_0, and state a conclusion.*** The obtained z-score is not in the critical region. Therefore, we fail to reject the null hypothesis. On the basis of these data, you conclude that the male/female proportions in the gender class are not significantly different from the proportions in the psychology department as a whole.

PROBLEMS

1. To investigate the phenomenon of "home team advantage," a researcher recorded the outcomes from 64 college football games on one Saturday in October. Of the 64 games, 42 were won by home teams. Does this result provide enough evidence to conclude that home teams win significantly more than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

2. Insurance companies charge young drivers more for automobile insurance because they tend to have more accidents than older drivers. To make this point, an insurance representative first determines that only 16% of licensed drivers are age 20 or younger. Because this age group makes up only 16% of the drivers, it is reasonable to predict that they should be involved in only 16% of the accidents. In a random sample of 100 accident reports, however, the representative finds 31 accidents that involved drivers who were 20 or younger. Is this sample sufficient to show that younger drivers have significantly more accidents than would be expected from the percentage of young drivers? Use a two-tailed test with $\alpha = .05$.

3. Güven, Elaimis, Binokay, and Tan (2003) studied the distribution of paw preferences in rats using a computerized food-reaching test. For a sample of $n = 144$ rats, they found 104 right-handed animals. Is this significantly more than would be expected if right- and left-handed rats are equally common in the population? Use a two-tailed test with $\alpha = .01$.

4. During the 2004 Olympic Games, Hill and Barton (2005) monitored contests in four combat sports: Greco-Roman wrestling, freestyle wrestling, boxing, and tae kwon do. Half of the competitors were assigned red outfits and half were assigned blue. The results indicate that participants wearing red won significantly more contests. Suppose that a sample of $n = 100$ contests produced 60 winners wearing red and only 40 wearing blue.

a. Is this enough evidence to justify a conclusion that wearing red produces significantly more wins than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

b. Because the outcome of the binomial test is a borderline z-score, use the real limits for $X = 60$ to determine if the entire z-score interval is located in the critical region. If any part of the interval is not in the critical region, the correct decision is to fail to reject the null hypothesis.

5. Problem 6 in Chapter 17 cited a study showing that people tend to choose partners who are similar to themselves. Jones, Pelham, Carvallo, & Mirenberg, (2004) demonstrated that people have a tendency to select marriage partners with surnames that begin with the same last letter as their own. The probability of randomly matching two last names beginning with the same letter is only $p = 0.065$ (6.5%). The researchers looked at marriage records and found that 38 out of 400 brides and grooms had surnames beginning with the same last letter. Is this significantly more than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

6. A researcher would like to determine whether people really can tell the difference between bottled water and tap water. Participants are asked to taste two unlabeled glasses of water, one bottled and one tap, and identify the one they thought tasted better. Out of 40 people in the sample, 28 picked the bottled water. Was the bottled water selected significantly more often than would be expected by chance? Use a two-tailed test with $\alpha = .05$.

7. In 1985, only 8% of the students in the city school district were classified as being learning disabled. A school psychologist suspects that the proportion of learning-disabled children has changed dramatically over the years. To demonstrate this point, a random sample of $n = 300$ students is selected. In this sample there are 42 students who have been identified as learning-disabled. Is this sample sufficient to indicate that there has been a significant change in the proportion of learning-disabled students since 1985? Use the .05 level of significance.

8. In the Preview section for Chapter 17, we discussed a study by Loftus and Palmer (1974) examining how different phrasing of questions can influence eyewitness testimony. In the study, students watched a video of an automobile accident and then were questioned about what they had seen. One group of participants was asked to estimate the speed of the cars when they "smashed into" each other. Another group of was asked to estimate the speed of the cars when they "hit" each other. Suppose that the actual speed of the cars was 22 miles per hour.

a. For the 50 people in the "smashed-into" group, assume that 32 overestimated the actual speed, 17 underestimated the speed, and 1 was exactly right. Is this result significantly different from what would be expected by chance? Use a two-tailed test with $\alpha = .05$.

b. For the 50 people in the "hit" group, assume that 27 overestimated the actual speed, 22 underestimated the speed, 1 was exactly right. Again, use a two-tailed test with $\alpha = .05$ to determine whether this result significantly different from what would be expected by chance.

9. A recent survey of practicing psychotherapists revealed that 25% of the individuals responding agreed with the

statement, "Hypnosis can be used to recover accurate memories of past lives" (Yapko, 1994). A researcher would like to determine whether this same level of belief exists in the general population. A sample of 192 adults is surveyed and 65 believe that hypnosis can be used to recover accurate memories of past lives. Based on these data, can you conclude that beliefs held by the general population are significantly different from beliefs held by psychotherapists? Test with $\alpha = .05$.

10. In 2005, Fung et al. published a study reporting that patients prefer technical quality versus interpersonal skills when selecting a primary care physician. Participants were presented with report cards describing pairs of hypothetical physicians and were asked to select the one that they preferred. Suppose that this study is repeated with a sample of $n = 150$ participants, and the results show that physicians with greater technical skill are preferred by 92 participants and physicians with greater interpersonal skills are selected by 58. Are these results sufficient to conclude that there is a significant preference for technical skill?

11. Danner and Phillips (2008) report the results from a county-wide study showing that delaying high school start times by one hour significantly reduced the motor vehicle crash rate for teen drivers in the study. Suppose that the researchers monitored 500 student drivers for 1 year after the start time was delayed and found that 44 were involved in automobile accidents. Before delaying the start time, the accident rate was 12%. Use a binomial test to determine whether these results indicate a significant change in the accident rate following the change in school start time. Use a two-tailed test with $\alpha = .05$

12. For each of the following, assume that a two-tailed test using the normal approximation to the binomial distribution with $\alpha = .05$ is being used to evaluate the significance of the result.

a. For a true-false test with 20 questions, how many would you have to get right to do significantly better than chance? That is, what X value is needed to produce a z-score greater than 1.96?

b. How many would you need to get right on a 40-question true-false test?

c. How many would you need to get right on a 100-question true-false test?

Remember that each X value corresponds to an interval with real limits. Be sure that the entire interval is in the critical region.

13. On a multiple-choice exam with 100 questions and 4 possible answers for each question, you get a score of $X = 32$. Is your score significantly better than would be expected by chance (by simply guessing for each question)? Use a two tailed test with $\alpha = .05$.

14. For each of the following, assume that a two-tailed test using the normal approximation to the binomial distribution with $\alpha = .05$ is being used to evaluate the significance of the result.

a. For a multiple-choice test with 48 questions, each with 4 possible answers, how many would you have to get right to do significantly better than chance? That is, what X value is needed to produce a z-score greater than 1.96?

b. How many would you need to get right on a multiple-choice test with 192 questions to be significantly better than chance?

Remember that each X value corresponds to an interval with real limits. Be sure that the entire interval is in the critical region.

15. Reed, Vernon, and Johnson (2004) examined the relationship between brain nerve conduction velocity and intelligence in normal adults. Brain nerve conduction velocity was measured three separate ways and nine different measures were used for intelligence. The researchers then correlated each of the three nerve velocity measures with each of the nine intelligence measures for a total of 27 separate correlations. Unfortunately, none of the correlations were significant.

a. For the 186 males in the study, however, 25 of the 27 correlations were positive. Is this significantly more than would be expected if positive and negative correlations were equally likely? Use a two-tailed test with $\alpha = .05$.

b. For the 201 females in the study, 20 of the 27 correlations were positive. Is this significantly more than would be expected if positive and negative correlations were equally likely? Use a two-tailed test with $\alpha = .05$.

16. In the Preview section for Chapter 11, we presented a study showing that swearing can help relieve pain (Stephens, Atkins, & Kingston, 2009). In the study, participants placed one hand in freezing cold water for as long as they could bear the pain. In one condition, they shouted a swear word over and over while the hand was in the water. In the other condition, they shouted a neutral word. Suppose that 18 of the 25 participants tolerated the pain longer while swearing than while shouting neutral words. Is this result significantly different from chance? Use a two-tailed test with $\alpha = .01$.

17. Thirty percent of the students in the local elementary school are classified as only children (no siblings). However, in the special program for talented and gifted children, 43 out of 90 students are only

children. Is the proportion of only children in the special program significantly different from the proportion for the school? Test at the .05 level of significance.

18. Stressful or traumatic experiences can often worsen other health-related problems such as asthma or rheumatoid arthritis. However, if patients are instructed to write about their stressful experiences, it can often lead to improvement in health (Smyth, Stone, Hurewitz, & Kaell, 1999). In a typical study, patients with asthma or arthritis are asked to write about the "most stressful event of your life." In a sample of $n = 112$ patients, suppose that 64 showed improvement in their symptoms, 12 showed no change, and 36 showed worsening symptoms.

a. If the 12 patients showing no change are discarded, are these results sufficient to conclude that the writing had a significant effect? Use a two-tailed test with $\alpha = .05$.

b. If the 12 patients who showed no change are split between the two groups, are the results sufficient to demonstrate a significant change? Use a two-tailed test with $\alpha = .05$.

19. Langewitz, Izakovic, and Wyler (2005) reported that self-hypnosis can significantly reduce hay-fever symptoms. Patients with moderate to severe allergic reactions were trained to focus their minds on specific locations where their allergies did not bother them, such as a beach or a ski resort. In a sample of 64 patients who received this training, suppose that 47 showed reduced allergic reactions and 17 showed an increase in allergic reactions. Are these results sufficient to conclude that the self-hypnosis has a significant effect? Use a two-tailed test with $\alpha = .05$.

20. Group-housed laying hens appear to prefer having more floor space than height in their cages. Albentosa and Cooper (2005) tested hens in groups of 10. The birds in each group were given free choice between a cage with a height of 38 cm (low) and a cage with a height of 45 cm (high). The results showed a tendency for the hens in each group to distribute themselves evenly between the two cages, showing no preference for either height. Suppose that a similar study tested a sample of $n = 80$ hens and found that 47 preferred the taller cage. Does this result indicate a significant preference? Use a two-tailed test with $\alpha = .05$.

21. In Problem 21 in Chapter 11, we described a study showing that students are likely to improve their test scores if they go back and change answers after reconsidering some of the questions on the exam (Johnston, 1975). In the study, one group of students was encouraged to reconsider each question and to change answers whenever they felt it was appropriate. The students were asked to record their original answers as well as the changes. For each student, the exam was graded based on the original answers and on the changed answers. For a group of $n = 40$ students, suppose that 29 had higher scores for the changed-answer version and only 11 had higher scores for the original-answer version. Is this result significantly different from chance? Use a two-tailed test with $\alpha = .01$.

22. The habituation technique is one method that is used to examine memory for infants. The procedure involves presenting a stimulus to an infant (usually projected on the ceiling above the crib) for a fixed time period and recording how long the infant spends looking at the stimulus. After a brief delay, the stimulus is presented again. If the infant spends less time looking at the stimulus during the second presentation, it is interpreted as indicating that the stimulus is remembered and, therefore, is less novel and less interesting than it was on the first presentation. This procedure is used with a sample of $n = 30$ 2-week-old infants. For this sample, 22 infants spent less time looking at the stimulus during the second presentation than during the first. Do these data indicate a significant difference? Test at the .01 level of significance.

23. Most children and adults are able to learn the meaning of new words by listening to sentences in which the words appear. Shulman and Guberman (2007) tested the ability of children to learn word meaning from syntactical cues for three groups: children with autism, children with specific language impairment (SLI), and children with typical language development (TLD). Although the researchers used relatively small samples, their results indicate that the children with TLD and those with autism were able to learn novel words using the syntactical cues in sentences. The children with SLI, on the other hand, experienced significantly more difficulty. Suppose that a similar study is conducted in which each child listens to a set of sentences containing a novel word and then is given a choice of three definitions for the word.

a. If 25 out of 36 autistic children select the correct definition, is this significantly more than would be expected if they were simply guessing? Use a two-tailed test with $\alpha = .05$.

b. If only 16 out of 36 children with SLI select the correct definition, is this significantly more than would be expected if they were simply guessing? Use a two-tailed test with $\alpha = .05$.

24. A researcher is testing the effectiveness of a skills-mastery imagery program for soccer players. A sample of $n = 25$ college varsity players is selected and each player is tested on a ball-handling obstacle course before beginning the imagery program and again after

completing the 5-week program. Of the 25 players, 18 showed improved performance on the obstacle course after the imagery program and 7 were worse.

a. Is this result sufficient to conclude that there is a significant change in performance following the imagery program? Use a two-tailed test with $\alpha = .05$.

b. Because the outcome of the binomial test is a borderline z-score, use the real limits for $X = 18$ and verify that the entire z-score interval is located in the critical region.

25. Last year the college counseling center offered a workshop for students who claimed to suffer from extreme exam anxiety. Of the 45 students who attended the workshop, 31 had higher grade-point averages this semester than they did last year. Do these data indicate a significant difference from what would be expected by chance? Test at the .01 level of significance.

26. Trying to fight a drug-resistant bacteria, a researcher tries an experimental drug on infected subjects. Out of 70 monkeys, 42 showed improvement, 22 got worse, and 6 showed no change. Is this researcher working in the right direction? Is there a significant effect of the drug on the infection? Use a two-tailed test at the .05 level of significance.

27. Biofeedback training is often used to help people who suffer migraine headaches. A recent study found that 29 out of 50 participants reported a decrease in the frequency and severity of their headaches after receiving biofeedback training. Of the remaining participants in this study, 10 reported that their headaches were worse, and 11 reported no change.

a. Discard the zero-difference participants, and use a sign test with $\alpha = .05$ to determine whether the biofeedback produced a significant difference.

b. Divide the zero-difference participants between the two groups, and use a sign test to evaluate the effect of biofeedback training.

REVIEW

After completing this part, you should be able to calculate and interpret correlations, find linear regression equations, conduct the chi-square tests for goodness of fit and for independence, and do a binomial test.

The most commonly used correlation is the Pearson correlation, which measures the direction and degree of linear relationship between two variables (X and Y) that have been measured on interval or ratio scales (numerical scores). The regression equation determines the best fitting line to describe the relationship between X and Y, and to compute predicted Y values for each value of X. A partial correlation can be used to reveal the underlying relationship between X and Y when the influence of a third variable is eliminated.

The Pearson formula is also used in a variety of other situations to compute special correlations. The Spearman correlation uses the Pearson formula when X and Y are both measured on ordinal scales (ranks). The Spearman correlation measures the direction and the degree to which the relationship is consistently one directional. When one of the variables consists of numerical scores and the other has only two values, the two values of the dichotomous variable can be coded as 0 and 1, and the Pearson formula can be used to find the point-biserial correlation. The point-biserial correlation measures the strength of the relationship between X and Y, and can be squared to produce the same r^2 value that is used to measure effect size for the independent-measures t test. When both variables are dichotomous, they can both be coded as 0 and 1, and the Pearson formula can be used to find the phi-coefficient. As a correlation, the phi-coefficient measures the strength of the relationship and is often used as a measure of effect size to accompany a chi-square test for independence for a 2×2 data matrix.

The chi-square test for goodness of fit uses the frequency distribution from a sample to evaluate a hypothesis about the corresponding population distribution. The null hypothesis for the goodness-of-fit test typically falls into one of two categories:

1. Equal proportions: The null hypothesis states that the population is equally distributed across the set of categories.
2. No difference: The null hypothesis states that the distribution for one population is not different from the known distribution for another population.

The chi-square test for independence uses frequency data from a sample to evaluate a hypothesis about the relationship between two variables in the population. The null hypothesis for this test can be phrased two different ways:

1. No relationship: The null hypothesis states that there is no relationship between the two variables in the population.
2. No difference: One variable is viewed as defining a set of different populations. The null hypothesis states that the frequency distribution for the second variable has the same shape (same proportions) for all the different populations.

The binomial test uses the frequencies or proportions from a sample to test a hypothesis about the corresponding population proportions for a binomial variable in the population. Because the binomial distribution approximates the normal distribution when pn and qn are both at least 10, it uses z-scores and proportions from the unit normal table for the test.

REVIEW PROBLEMS

1. The following scores are related by the equation $Y = X^2$. Note that this is not a linear relationship, but every time X increases, Y also increases.

X	Y
2	4
4	16
6	36
8	64
10	100

a. Compute the Pearson correlation between X and Y. You should find a positive, but not perfect, correlation.

b. Convert the scores to ranks and compute the Spearman correlation. You should find a perfect, positive correlation.

2. It is well known that similarity in attitudes, beliefs, and interests plays an important role in interpersonal attraction (see Byrne, 1971, for example). Thus, correlations for attitudes between married couples should be strong. Suppose that a researcher developed a questionnaire that measures how liberal or conservative one's attitudes are. Low scores indicate that the person has liberal attitudes, whereas high scores indicate conservatism. The following hypothetical data are scores for married couples.

Couple	Wife	Husband
A	11	14
B	6	7
C	16	15
D	4	7
E	1	3
F	10	9
G	5	9
H	3	8

a. Compute the Pearson correlation for these data.
b. Find the regression equation for predicting the husband's score from the wife's.

3. A researcher is investigating the physical characteristics that influence whether a person's face is judged as beautiful. The researcher selects a photograph of a woman and then creates two modifications of the photo by (1) moving the eyes slightly farther apart and (2) moving the eyes slightly closer together. The original photograph and the two modifications are then shown to a sample of $n = 150$ college students, and each student is asked to select the "most beautiful" of the three faces. The distribution of responses was as follows:

Original Photo	Eyes Moved Apart	Eyes Moved Together
51	72	27

Do the data indicate any significant preferences among the three versions of the photograph? Test at the .05 level of significance.

4. Friedman and Rosenman (1974) have suggested that personality type is related to heart disease. Specifically, type A people, who are competitive, driven, pressured, and impatient, are more prone to heart disease. On the other hand, type B individuals, who are less competitive and more relaxed, are less likely to have heart disease. Suppose that an investigator would like to examine the relationship between personality type and disease. For a random sample of individuals, personality type is assessed with a standardized test. These individuals are then examined and categorized as to whether they have a heart disorder. The observed frequencies are as follows:

	No Heart Disease	Heart Disease	
Type A	32	18	50
Type B	128	22	150
	160	40	

a. Is there a relationship between personality and disorder? Test at the .05 level of significance.
b. Compute the phi-coefficient to measure the strength of the relationship.

5. One of the original methods for testing ESP (extrasensory perception) used Zener cards, which were designed specifically for the testing process. Each card shows one of five symbols (square, circle, star, wavy lines, cross). The person being tested must predict the symbol before the card is turned over. Chance performance on this task would produce 1 out of 5 (20%) correct predictions. Use a binomial test to determine whether 27 correct predictions out of 100 attempts is significantly different from chance performance? Use a two-tailed test with $\alpha = .05$.

This page intentionally left blank

CHAPTER

19

Choosing the Right Statistics

Tools You Will Need

This chapter provides an organized overview for most of the statistical procedures presented in this book. The following items are considered background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Descriptive statistics
- Mean (Chapter 3)
- Standard deviation (Chapter 4)
- Correlation (Chapter 15)
- Inferential Statistics (Chapters 9, 10, 11, 12, 13, 14, 15, 16, 17, 18)

Preview

After students have completed a statistics course, they occasionally are confronted with situations in which they have to apply the statistics they have learned. For example, in the context of a research methods course, or while working as a research assistant, students are presented with the results from a study and asked to do the appropriate statistical analysis. The problem is that many of these students have no idea where to begin. Although they have learned the individual statistics, they cannot match the statistical procedures to a specific set of data. Our goal for this chapter is to provide some help with the problem.

We assume that you know (or can anticipate) what your data look like. Therefore, we begin by presenting some basic categories of data so you can find that one that matches your own data. For each data category, we then present the potential statistical procedures and identify the factors that determine which are appropriate for you based on the specific characteristics of your data.

19.1 THREE BASIC DATA STRUCTURES

Most research data can be classified in one of three basic categories.

Category 1: A single group of participants with one score per participant.

Category 2: A single group of participants with two (or more) variables measured for each participant.

Category 3: Two (or more) groups of scores with each score a measurement of the same variable.

In this section, we present examples of each structure. Once you match your own data to one of the examples, you can proceed to the section of the chapter in which we describe the statistical procedures that apply to that example.

SCALES OF MEASUREMENT

Before we begin discussion of the three categories of data, there is one other factor that differentiates data within each category and helps to determine which statistics are appropriate. In Chapter 1, we introduced four scales of measurement and noted that different measurement scales allow different kinds of mathematical manipulation, which result in different statistics. For most statistical applications, however, ratio and interval scales are equivalent so we group them together for the following review.

Ratio scales and **interval scales** produce numerical scores that are compatible with the full range of mathematical manipulation. Examples include measurements of height in inches, weight in pounds, the number of errors on a task, and IQ scores.

Ordinal scales consist of ranks or ordered categories. Examples include classifying cups of coffee as small, medium, and large or ranking job applicants as 1st, 2nd, and 3rd.

Nominal scales consist of named categories. Examples include gender (male/female), academic major, or occupation.

Within each category of data, we present examples representing these three measurement scales and discuss the statistics that apply to each.

CATEGORY 1: A SINGLE GROUP OF PARTICIPANTS WITH ONE SCORE PER PARTICIPANT

This type of data often exists in research studies that are conducted simply to describe individual variables as they exist naturally. For example, a recent news report stated that half of American teenagers, ages 12 through 17, send 50 or more text messages a day. To get this number, the researchers had to measure the number of text messages for each individual in a large sample of teenagers. The resulting data consist of one score per participant for a single group.

It is also possible that the data are a portion of the results from a larger study examining several variables. For example, a college administrator may conduct a survey to obtain information describing the eating, sleeping, and study habits of the college's students. Although several variables are being measured, the intent is to look at them one at a time. For example, the administrator will look at the number of hours each week that each student spends studying. These data consist of one score for each individual in a single group. The administrator will then shift attention to the number of hours per day that each student spends sleeping. Again, the data consist of one score for each person in a single group. The identifying feature for this type of research (and this type of data) is that there is no attempt to examine relationships between different variables. Instead, the goal is to describe individual variables, one at a time.

Table 19.1 presents three examples of data in this category. Note that the three data sets differ in terms of the scale of measurement used to obtain the scores. The first set (a) shows numerical scores measured on an interval or ratio scale. The second set (b) consists of ordinal, or rank ordered categories, and the third set shows nominal measurements. The statistics used for data in this category are discussed in Section 19.2.

CATEGORY 2: A SINGLE GROUP OF PARTICIPANTS WITH TWO (OR MORE) VARIABLES MEASURED FOR EACH PARTICIPANT

These research studies are specifically intended to examine relationships between variables. Note that different variables are being measured, so each participant has two or more scores, each representing a different variable. Typically, there is no attempt to manipulate or control the variables; they are simply observed and recorded as they exist naturally.

Although several variables may be measured, researchers usually select pairs of variables to evaluate specific relationships. Therefore, we present examples showing pairs of variables. Table 19.2 presents four examples of data in this category. Once again, the four data sets differ in terms of the scales of measurement that are used. The first set of data (a) shows numerical scores for each set of measurements. For the second set (b), we have ranked the scores from the first set and show the resulting ranks. The third data set (c) shows numerical scores for one variable and nominal scores for the second variable. In the fourth set (d), both scores are measured on a nominal scale

TABLE 19.1

Three examples of data with one score per participant for one group of participants.

(a) Number of Text Messages Sent in Past 24 Hours	(b) Rank in Class for High School Graduation	(c) Got a Flu Shot Last Season
X	*X*	*X*
6	23rd	No
13	18th	No
28	5th	Yes
11	38th	No
9	17th	Yes
31	42nd	No
18	32nd	No

TABLE 19.2

Examples of data with two scores for each participant for one group of participants.

(a) SAT Score (*X*) and College Freshman GPA (*Y*)		(b) Ranks for the Scores in Set (a)	
X	*Y*	*X*	*Y*
620	3.90	7	8
540	3.12	3	2
590	3.45	6	5
480	2.75	1	1
510	3.20	2	3
660	3.85	8	7
570	3.50	5	6
560	3.24	4	4

(c) Age (*X*) and Wrist Watch Preference (*Y*)		(d) Gender (*X*) and Academic Major (*Y*)	
X	*Y*	*X*	*Y*
27	digital	M	Sciences
43	analog	M	Humanities
19	digital	F	Arts
34	digital	M	Professions
37	digital	F	Professions
49	analog	F	Humanities
22	digital	F	Arts
65	analog	M	Sciences
46	digital	F	Humanities

of measurement. The appropriate statistical analyses for these data are discussed in Section 19.3.

CATEGORY 3: TWO OR MORE GROUPS OF SCORES WITH EACH SCORE A MEASUREMENT OF THE SAME VARIABLE

A second method for examining relationships between variables is to use the categories of one variable to define different groups and then measure a second variable to obtain a set of scores within each group. The first variable, defining the groups, usually falls into one of the following general categories:

a. Participant characteristic: For example, gender or age.

b. Time: For example, before versus after treatment.

c. Treatment conditions: For example, with caffeine versus without caffeine.

If the scores in one group are consistently different from the scores in another group, then the data indicate a relationship between variables. For example, if the performance scores for a group of females are consistently higher than the scores for a group of males, then there is a relationship between performance and gender.

Another factor that differentiates data sets in this category is the distinction between independent-measures and repeated-measures designs. Independent-measures designs were introduced in Chapters 10 and 12, and repeated-measures designs were presented in Chapters 11 and 13. You should recall that an *independent-measures design,* also known as a *between-subjects design,* requires a separate group of participants for each group of scores. For example, a study comparing scores for males with scores for females would require two groups of participants. On the other hand, a *repeated-measures design,* also known as a *within-subjects design,* obtains several

groups of scores from the same group of participants. A common example of a repeated-measures design is a before/after study in which one group of individuals is measured before a treatment and then measured again after the treatment.

Examples of data sets in this category are presented in Table 19.3. The table includes a sampling of independent-measures and repeated-measures designs as well as examples representing measurements from several different scales of measurement. The appropriate statistical analyses for data in this category are discussed in Section 19.4.

19.2 STATISTICAL PROCEDURES FOR DATA FROM A SINGLE GROUP OF PARTICIPANTS WITH ONE SCORE PER PARTICIPANT

One feature of this data category is that the researcher typically does not want to examine a relationship between variables but rather simply intends to describe individual variables as they exist naturally. Therefore, the most commonly used statistical procedures for these data are descriptive statistics that are used to summarize and describe the group of scores.

We should note that the same descriptive statistics used to describe a single group are also used to describe groups of scores that are a part of more complex data sets. For example, a researcher may want to compare a set of scores for males with a set of scores

TABLE 19.3

Examples of data comparing two or more groups of scores with all scores measuring the same variable.

(a) Attractiveness Ratings for a Woman in a Photograph Shown on a Red or a White Background

White	Red
5	7
4	5
4	4
3	5
4	6
3	4
4	5

(b) Performance Scores Before and After 24 Hours of Sleep Deprivation

Participant	Before	After
A	9	7
B	7	6
C	7	5
D	8	8
E	5	4
F	9	8
G	8	5

(c) Success or Failure on a Task for Participants Working Alone or in a Group

Alone	Group
Fail	Succeed
Succeed	Succeed
Succeed	Succeed
Succeed	Succeed
Fail	Fail
Fail	Succeed
Succeed	Succeed
Fail	Succeed

(d) Amount of Time Spent on Facebook (Small, Medium, Large) for Students from Each High School Class

Freshman	Sophomore	Junior	Senior
med	small	med	large
small	large	large	med
small	med	large	med
med	med	large	large
small	med	med	large
large	large	med	large
med	large	small	med
small	med	large	large

for females (data from category 3). However, the statistics used to describe the group of males would be the same as the descriptive statistics that would be used if the males were the only group in the study.

SCORES FROM RATIO OR INTERVAL SCALES: NUMERICAL SCORES

When the data consist of numerical values from interval or ratio scales, there are several options for descriptive and inferential statistics. We consider the most likely statistics and mention some alternatives.

Descriptive Statistics The most often used descriptive statistics for numerical scores are the mean (Chapter 3) and the standard deviation (Chapter 4). If there are a few extreme scores or the distribution is strongly skewed, the median (Chapter 3) may be better than the mean as a measure of central tendency.

Inferential Statistics If there is a basis for a null hypothesis concerning the mean of the population from which the scores were obtained, a single-sample *t* test (Chapter 9) can be used to evaluate the hypothesis. Some potential sources for a null hypothesis are as follows:

1. If the scores are from a measurement scale with a well-defined neutral point, then the *t* test can be used to determine whether the sample mean is significantly different from (higher than or lower than) the neutral point. On a 7-point rating scale, for example, a score of $X = 4$ is often identified as neutral. The null hypothesis would state that the population mean is equal to (greater than or less than) $\mu = 4$.
2. If the mean is known for a comparison population, then the *t* test can be used to determine whether the sample mean is significantly different from (higher than or lower than) the known value. For example, it may be known that the average score on a standardized reading achievement test for children finishing first grade is $\mu = 20$. If the sample consists of test scores for first grade children who are all the only child in the household, the null hypothesis would state that the mean for children in this population is also equal to 20. The known mean could also be from an earlier time, for example 10 years ago. The hypothesis test would then determine whether a sample from today's population indicates a significant change in the mean during the past 10 years.

The single-sample *t* test evaluates the statistical significance of the results. A significant result means that the data are very unlikely ($p < \alpha$) to have been produced by random, chance factors. However, the test does not measure the size or strength of the effect. Therefore, a *t* test should be accompanied by a measure of effect size such as Cohen's *d* or the percentage of variance accounted for, r^2.

SCORES FROM ORDINAL SCALES: RANKS OR ORDERED CATEGORIES

Descriptive Statistics Occasionally, the original scores are measurements on an ordinal scale. It is also possible that the original numerical scores have been transformed into ranks or ordinal categories (for example, small, medium, and large). In either case, the median is appropriate for describing central tendency for ordinal measurements and proportions can be used to describe the distribution of individuals across categories. For example, a researcher might report that 60% of the students were in the high self-esteem category, 30% in the moderate self-esteem category, and only 10% in the low self-esteem category.

Inferential Statistics If there is a basis for a null hypothesis specifying the proportions in each ordinal category for the population from which the scores were obtained, then a chi-square test for goodness of fit (Chapter 17) can be used to evaluate the hypothesis. With only two categories, the binomial test (Chapter 18) also can be used. For example, it may be reasonable to hypothesize that the categories occur equally often (equal proportions) in the population and the test would determine whether the sample proportions are significantly different. If the original data were converted from numerical values into ordered categories using z-score values to define the category boundaries, then the null hypothesis could state that the population distribution is normal, using proportions obtained from the unit normal table. A chi-square test for goodness of fit would determine whether the shape of the sample distribution is significantly different from a normal distribution. For example, the null hypothesis would state that the distribution has the following proportions, which describe a normal distribution according to the unit normal table:

$z < -1.5$	$-1.5 < z < -0.5$	$-0.5 < z < 0.5$	$0.5 < z < 1.5$	$z > 1.50$
6.68%	24.17%	38.30%	24.17%	6.68%

SCORES FROM A NOMINAL SCALE

For these data, the scores simply indicate the nominal category for each individual. For example, individuals could be classified as male/female or grouped into different occupational categories.

Descriptive Statistics The only descriptive statistics available for these data are the mode (Chapter 3) for describing central tendency or using proportions (or percentages) to describe the distribution across categories.

Inferential Statistics If there is a basis for a null hypothesis specifying the proportions in each category for the population from which the scores were obtained, then a chi-square test for goodness of fit (Chapter 17) can be used to evaluate the hypothesis. With only two categories, the binomial test (Chapter 18) also can be used. For example, it may be reasonable to hypothesize that the categories occur equally often (equal proportions) in the population. If proportions are known for a comparison population or for a previous time, the null hypothesis could specify that the proportions are the same for the population from which the scores were obtained. For example, if it is known that 35% of the adults in the United States get a flu shot each season, then a researcher could select a sample of college students and count how many got a shot and how many did not [see the data in Table 19.1(c)]. The null hypothesis for a chi-square test or a binomial test would state that the distribution for college students is not different from the distribution for the general population. For the chi-square test

	Flu Shot	No Flu Shot
H_0:	35%	65%

For the binomial test, H_0: $p = p(\text{shot}) = 0.35$ and $q = p(\text{no shot}) = 0.65$

Figure 19.1 summarizes the statistical procedures used for data in category 1.

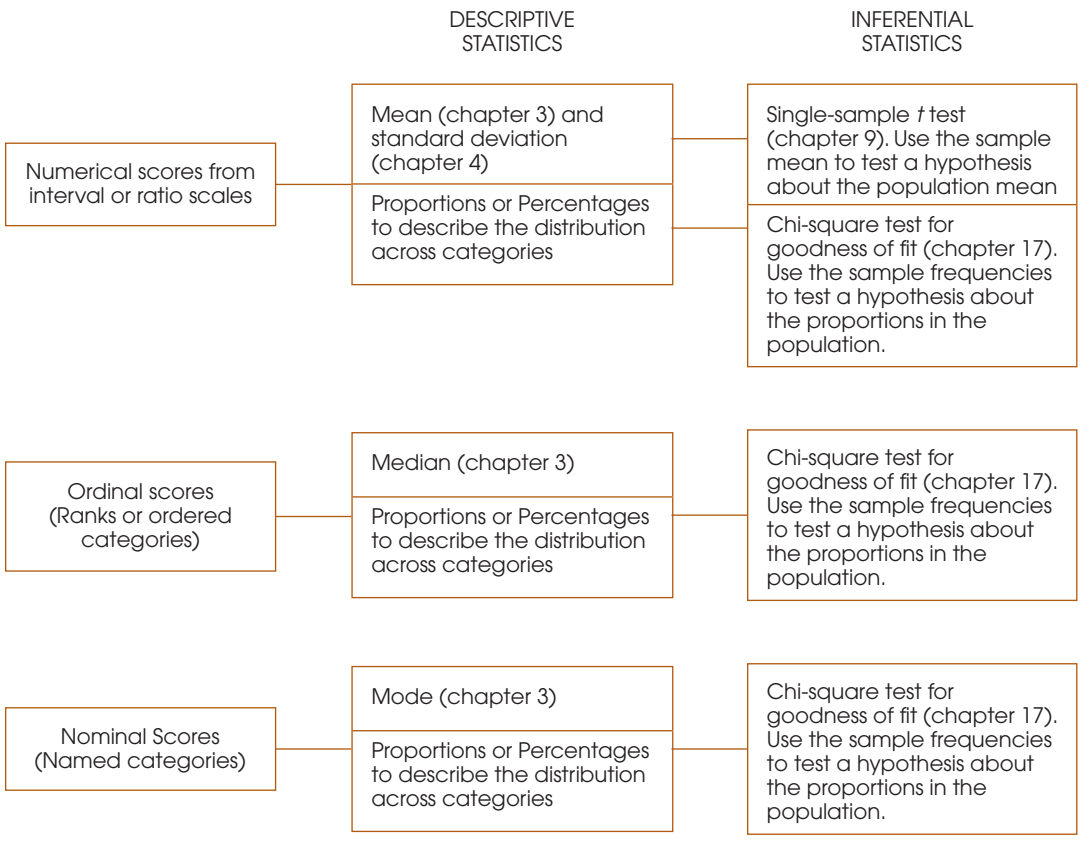


FIGURE 19.1

Statistics for Category 1 data. A single group of participants with one score per participant. The goal is to describe the variable as it exists naturally.

19.3 STATISTICAL PROCEDURES FOR DATA FROM A SINGLE GROUP OF PARTICIPANTS WITH TWO (OR MORE) VARIABLES MEASURED FOR EACH PARTICIPANT

The goal of the statistical analysis for data in this category is to describe and evaluate the relationships between variables, typically focusing on two variables at a time. With only two variables, the appropriate statistics are correlations (Chapter 15), linear regression (Chapter 16), and the chi-square test for independence (Chapter 17). With three or more variables, the applicable statistics are partial correlation (Chapter 15) and multiple regression (Chapter 16).

TWO NUMERICAL VARIABLES FROM INTERVAL OR RATIO SCALES

The Pearson correlation measures the degree and direction of linear relationship between the two variables (see Example 15.3 on p. 517). Linear regression determines the equation for the straight line that gives the best fit to the data points. For each X value in the data, the equation produces a predicted Y value on the line so

that the squared distances between the actual Y values and the predicted Y values are minimized.

Descriptive Statistics The Pearson correlation serves as its own descriptive statistic. Specifically, the sign and magnitude of the correlation describe the linear relationship between the two variables. The squared correlation is often used to describe the strength of the relationship. The linear regression equation provides a mathematical description of the relationship between X values and Y. The slope constant describes the amount that Y changes each time the X value is increased by 1 point. The constant (Y intercept) value describes the value of Y when X is equal to zero.

Inferential Statistics The statistical significance of the Pearson correlation is evaluated by comparing the sample correlation with critical values listed in Table B6. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population. Analysis of regression is a hypothesis-testing procedure that evaluates the significance of the regression equation. Statistical significance means that the equation predicts more of the variance in the Y scores than would be reasonable to expect if there were not a real underlying relationship between X and Y.

TWO ORDINAL VARIABLES (RANKS OR ORDERED CATEGORIES)

The Spearman correlation is used when both variables are measured on ordinal scales (ranks). If one or both variables consist of numerical scores from an interval or ratio scale, then the numerical values can be transformed to ranks and the Spearman correlation can be computed.

Descriptive Statistics The Spearman correlation describes the degree and direction of monotonic relationship; that is the degree to which the relationship is consistently one directional.

Inferential Statistics The statistical significance of the Spearman correlation is evaluated by comparing the sample correlation with critical values listed in Table B7. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population.

ONE NUMERICAL VARIABLE AND ONE DICHOTOMOUS VARIABLE (A VARIABLE WITH EXACTLY 2 VALUES)

The point-biserial correlation measures the relationship between a numerical variable and a dichotomous variable. The two categories of the dichotomous variable are coded as numerical values, typically 0 and 1, to calculate the correlation.

Descriptive Statistics Because the point-biserial correlation uses arbitrary numerical codes, the direction of relationship is meaningless. However, the size of the correlation, or the squared correlation, describes the degree of relationship.

Inferential Statistics The data for a point-biserial correlation can be regrouped into a format suitable for an independent-measures t hypothesis test, or the t value can be computed directly from the point-biserial correlation (see the example on pages 542–544). The t value from the hypothesis test determines the significance of the relationship.

TWO DICHOTOMOUS VARIABLES

The phi-coefficient is used when both variables are dichotomous. For each variable, the two categories are numerically coded, typically as 0 and 1, to calculate the correlation.

Descriptive Statistics Because the phi-coefficient uses arbitrary numerical codes, the direction of relationship is meaningless. However, the size of the correlation, or the squared correlation, describes the degree of relationship.

Inferential Statistics The data from a phi-coefficient can be regrouped into a format suitable for a 2×2 chi-square test for independence, or the chi-square value can be computed directly from the phi-coefficient (see Example 17.3 on p. 613). The chi-square value determines the significance of the relationship.

TWO VARIABLES FROM ANY MEASUREMENT SCALES

The chi-square test for independence (Chapter 17) provides an alternative to correlations for evaluating the relationship between two variables. For the chi-square test, each of the two variables can be measured on any scale, provided that the number of categories is reasonably small. For numerical scores covering a wide range of value, the scores can be grouped into a smaller number of ordinal intervals. For example, IQ scores ranging from 93 to 137 could be grouped into three categories described as high, medium, and low IQ.

For the chi-square test, the two variables are used to create a matrix showing the frequency distribution for the data. The categories for one variable define the rows of the matrix and the categories of the second variable define the columns. Each cell of the matrix contains the frequency or number of individuals whose scores correspond to the row and column of the cell. For example, the gender and academic major scores in Table 19.2(d) could be reorganized in a matrix as follows:

	Arts	Humanities	Sciences	Professions
Female				
Male				

The value in each cell is the number of students with the gender and major identified by the cell's row and column. The null hypothesis for the chi-square test would state that there is no relationship between gender and academic major.

Descriptive Statistics The chi-square test is an inferential procedure that does not include the calculation of descriptive statistics. However, it is customary to describe the data by listing or showing the complete matrix of observed frequencies. Occasionally researchers describe the results by pointing out cells that have exceptionally large discrepancies. For example, in the Preview for Chapter 17 we described a study investigating eyewitness memory. Participants watched a video of an automobile accident and were questioned about what they saw. One group was asked to estimate the speed of the cars when they "smashed into" each other and another group was asked to estimate speed when the cars "hit" each other. A week later, they were asked additional questions, including whether they recalled seeing broken glass. Part of the description of the results focuses on cells reporting "Yes" responses. Specifically, the "smashed into" group had more than twice as many "Yes" responses than the "hit" group.

Inferential Statistics The chi-square test evaluates the significance of the relationship between the two variables. A significant result means that the distribution of frequencies in the data is very unlikely to occur ($p < \alpha$) if there is no underlying relationship between variables in the population. As with most hypothesis tests, a significant result does not provide information about the size or strength of the relationship. Therefore, either a phi-coefficient or Cramér's V is used to measure effect size.

THREE NUMERICAL VARIABLES FROM INTERVAL OR RATIO SCALES

To evaluate the relationship among three variables, the appropriate statistics are partial correlation (Chapter 15) and multiple regression (Chapter 16). A partial correlation measures the relationship between two variables while controlling the third variable. Multiple regression determines the linear equation that gives the best fit to the data points. For each pair of X values in the data, the equation produces a predicted Y value so that the squared distances between the actual Y values and the predicted Y values are minimized.

Descriptive Statistics A partial correlation describes the direction and degree of linear relationship between two variables while the influence of a third variable is controlled. This technique determines the degree to which the third variable is responsible for what appears to be a relationship between the first two. The multiple regression equation provides a mathematical description of the relationship between the two X values and Y. Each of the two slope constants describes the amount that Y changes each time the corresponding X value is increased by 1 point. The constant value describes the value of Y when both X values are equal to zero.

Inferential Statistics The statistical significance of a partial correlation is evaluated by comparing the sample correlation with critical values listed in Table B6 using $df = n - 3$ instead of the $n - 2$ value that is used for a routine Pearson correlation. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population. Analysis of regression evaluates the significance of the multiple regression equation. Statistical significance means that the equation predicts more of the variance in the Y scores than would be reasonable to expect if there were not a real underlying relationship between the two Xs and Y.

THREE VARIABLES INCLUDING NUMERICAL VALUES AND DICHOTOMOUS VARIABLES

Partial correlation (Chapter 15) and multiple regression (Chapter 16) can also be used to evaluate the relationship among three variables, including one or more dichotomous variables. For each dichotomous variable, the two categories are numerically coded, typically as 0 and 1, before the partial correlation or multiple regression is done. The *descriptive statistics* and the *inferential statistics* for the two statistical procedures are identical to those for numerical scores except that direction of relationship (or sign of the slope constant) is meaningless for the dichotomous variables.

Figure 19.2 summarizes the statistical procedures used for data in category 2.

19.4 STATISTICAL PROCEDURES FOR DATA CONSISTING OF TWO (OR MORE) GROUPS OF SCORES WITH EACH SCORE A MEASUREMENT OF THE SAME VARIABLE

Data in this category includes single-factor and two-factor designs. In a single-factor study, the values of one variable are used to define different groups and a second variable (the dependent variable) is measured to obtain a set of scores in each group. For a two-factor design, two variables are used to construct a matrix with the values of one variable defining the rows and the values of the second variable defining the columns. A third variable (the dependent variable) is measured to obtain a set of scores in each cell of the matrix. To simplify discussion, we focus on single-factor designs now and address two-factor designs in a separate section at the end of this chapter.

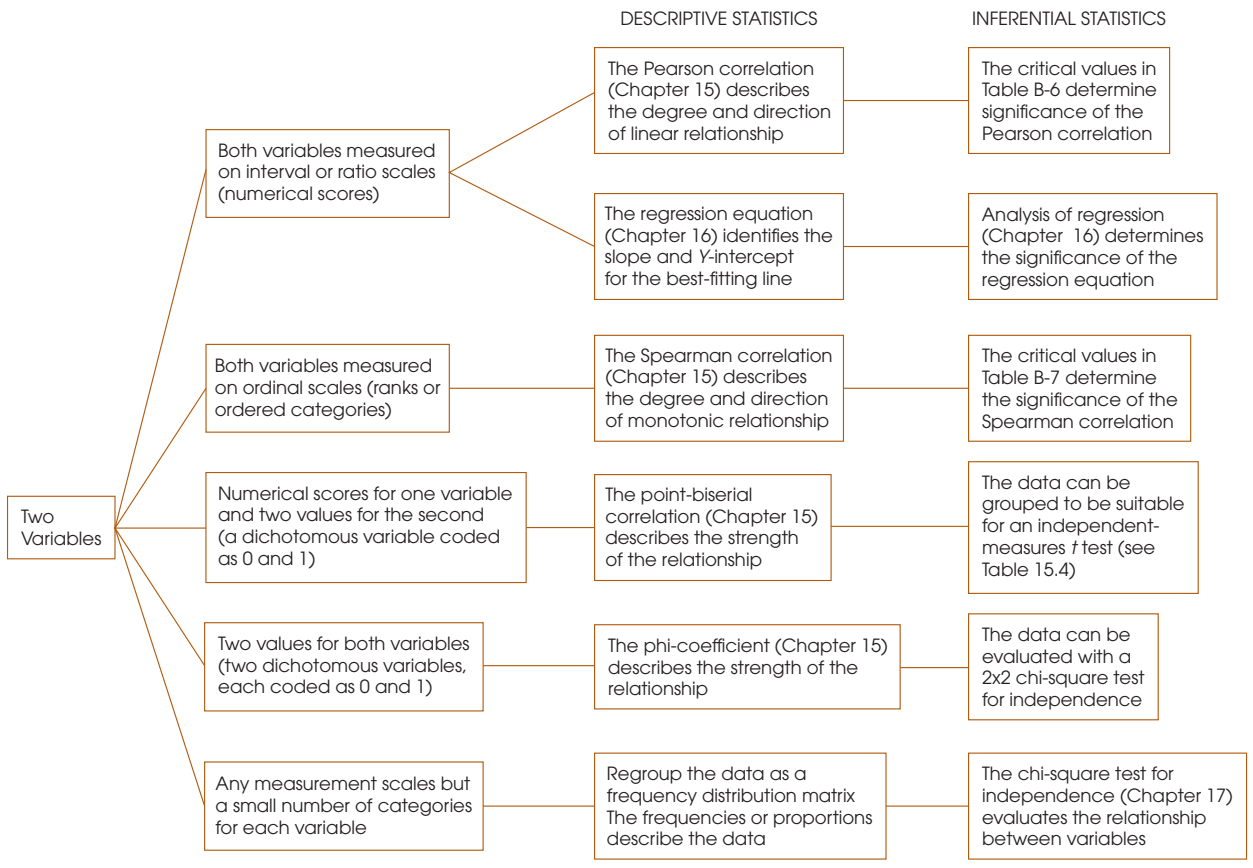


FIGURE 19.2

Statistics for Category 2 data. One group of participants with two (or more) variables measured for each participant. The goal is to describe and evaluate the relationship between variables.

The goal for a single-factor research design is to demonstrate a relationship between the two variables by showing consistent differences between groups. The scores in each group can be numerical values measured on interval or ratio scales, ordinal values (ranks), or simply categories on a nominal scale. The different measurement scales permit different types of mathematics and result in different statistical analyses.

SCORES FROM INTERVAL OR RATIO SCALES: NUMERICAL SCORES

Descriptive Statistics When the scores in each group are numerical values, the standard procedure is to compute the mean (Chapter 3) and the standard deviation (Chapter 4) as descriptive statistics to summarize and describe each group. For a repeated-measures study comparing exactly two groups, it also is common to compute the difference between the two scores for each participant and then report the mean and the standard deviation for the difference scores.

Inferential Statistics Analysis of variance (ANOVA) and *t* tests are used to evaluate the statistical significance of the mean differences between the groups of

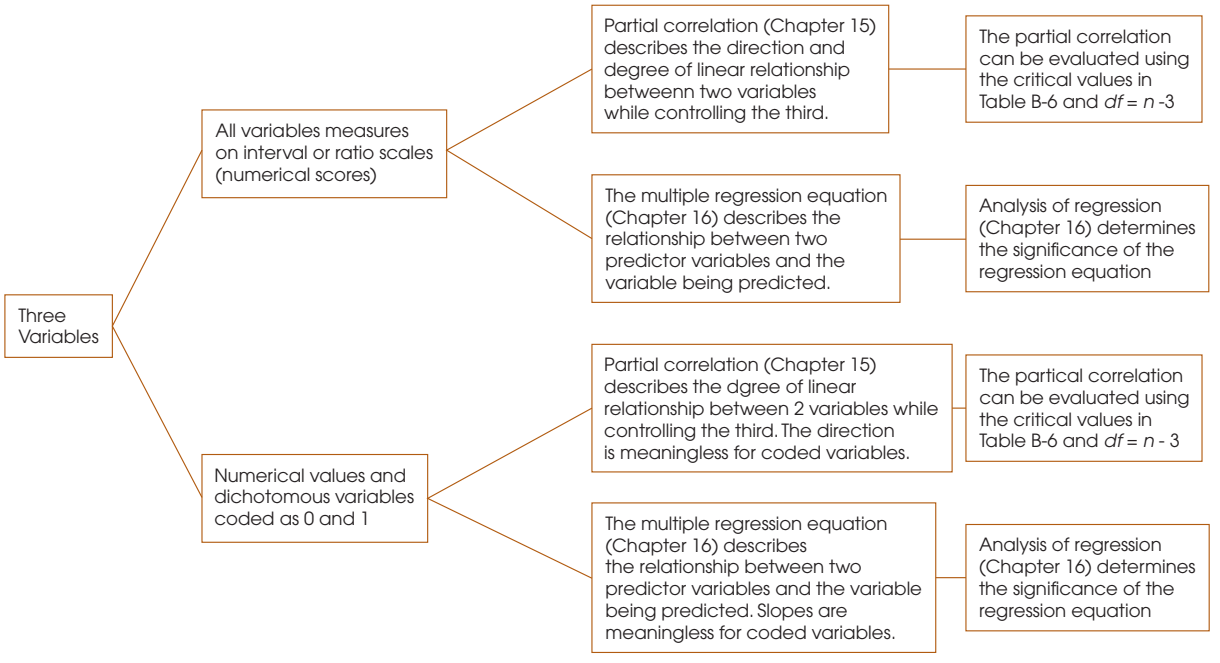


FIGURE 19.2—Cont'd

scores. With only two groups, the two tests are equivalent and either may be used. With more than two groups, mean differences are evaluated with an ANOVA. For independent-measures designs (between-subjects designs), the independent-measures t (Chapter 10) and independent-measures ANOVA (Chapter 12) are appropriate. For repeated-measures designs, the repeated-measures t (Chapter 11) and repeated-measures ANOVA (Chapter 13) are used. For all tests, a significant result indicates that the sample mean differences in the data are very unlikely ($p < \alpha$) to occur if there are not corresponding mean differences in the population. For an ANOVA comparing more than two means, a significant F-ratio indicates that post tests such as Scheffé or Tukey (Chapter 12) are necessary to determine exactly which sample means are significantly different. Significant results from a t test should be accompanied by a measure of effect size such as Cohen's d or r^2. For ANOVA, effect size is measured by computing the percentage of variance accounted for, η^2.

SCORES FROM ORDINAL SCALES: RANKS OR ORDERED CATEGORIES

For scores that are rank ordered, there are hypothesis tests developed specifically for ordinal data to determine whether there are significant differences in the ranks from one group to another. Also, if the scores are limited to a relatively small number of ordinal categories, then a chi-square test for independence can be used to determine whether there are significant differences in proportions from one group to another.

Descriptive Statistics Ordinal scores can be described by the set of ranks or ordinal categories within each group. For example, the ranks in one group may be consistently

larger (or smaller) than ranks in another group. Or, the "large" ratings may be concentrated in one group and the "small" ratings in another.

Inferential Statistics Appendix E presents a set of hypothesis tests developed for evaluating differences between groups of ordinal data.

1. The Mann-Whitney *U* test evaluates the difference between two groups of scores from an independent-measures design. The scores are the ranks obtained by combining the two groups and rank ordering the entire set of participants from smallest to largest.
2. The Wilcoxon signed ranks test evaluates the difference between two groups of scores from a repeated-measures design. The scores are the ranks obtained by rank ordering the magnitude of the differences, independent of sign (+ or −).
3. The Kruskal-Wallis test evaluates differences between three or more groups from an independent-measures design. The scores are the ranks obtained by combining all of the groups and rank ordering the entire set of participants from smallest to largest.
4. The Friedman test evaluates differences among three or more groups from a repeated-measures design. The scores are the ranks obtained by rank ordering the scores for each participant. With three conditions, for example, each participant is measured three times and would receive ranks of 1, 2, and 3.

For all tests, a significant result indicates that the differences between groups are very unlikely ($p < \alpha$) to have occurred unless there are consistent differences in the population.

A chi-square test for independence (Chapter 17) can be used to evaluate differences between groups for an independent-measures design with a relatively small number of ordinal categories for the dependent variable. In this case, the data can be displayed as a frequency distribution matrix with the groups defining the rows and the ordinal categories defining the columns. For example, a researcher could group high school students by class (Freshman, Sophomore, Junior, Senior) and measure the amount of time each student spends on Facebook by classifying students into three ordinal categories (small, medium, large). An example of the resulting data is shown in Table 19.3(d). However, the same data could be regrouped into a frequency-distribution matrix as follows:

	Amount of Time Spent on Facebook		
	Small	Medium	Large
Freshman			
Sophomore			
Junior			
Senior			

The value in each cell is the number of students, with the high school class and amount of Facebook time identified by the cell's row and column. A chi-square test for independence would evaluate the differences between groups. A significant result indicates that the frequencies (proportions) in the sample data would be very unlikely ($p < \alpha$) to occur unless the proportions for the population distributions are different from one group to another.

SCORES FROM A NOMINAL SCALE

Descriptive Statistics As with ordinal data, data from nominal scales are usually described by the distribution of individuals across categories. For example, the scores in one group may be clustered in one category or set of categories and the scores in another group may be clustered in different categories.

Inferential Statistics With a relatively small number of nominal categories, the data can be displayed as a frequency-distribution matrix with the groups defining the rows and the nominal categories defining the columns. The number in each cell is the frequency, or number of individuals in the group, identified by the cell's row, with scores corresponding to the cell's column. For example, the data in Table 19.3(c) show success or failure on a task for participants who are working alone or working in a group. These data could be regrouped as follows:

	Success	Failure
Work Alone		
Work in a Group		

A chi-square test for independence (Chapter 17) can be used to evaluate differences between groups. A significant result indicates that the two sample distributions would be very unlikely ($p < \alpha$) to occur if the two population distributions have the same proportions (same shape).

TWO-FACTOR DESIGNS WITH SCORES FROM INTERVAL OR RATIO SCALES

Research designs with two independent (or quasi-independent) variables are known as two-factor designs. These designs can be presented as a matrix with the levels of one factor defining the rows and the levels of the second factor defining the columns. A third variable (the dependent variable) is measured to obtain a group of scores in each cell of the matrix (see Example 14.1 on page 477).

Descriptive Statistics When the scores in each group are numerical values, the standard procedure is to compute the mean (Chapter 3) and the standard deviation (Chapter 4) as descriptive statistics to summarize and describe each group.

Inferential Statistics A two-factor ANOVA is used to evaluate the significance of the mean differences between cells. The ANOVA separates the mean differences into three categories and conducts three separate hypothesis tests:

1. The main effect for factor *A* evaluates the overall mean differences for the first factor; that is, the mean differences between rows in the data matrix.
2. The main effect for factor *B* evaluates the overall mean differences for the second factor; that is, the mean differences between columns in the data matrix.
3. The interaction between factors evaluates the mean differences between cells that are not accounted for by the main effects.

For each test, a significant result indicates that the sample mean differences in the data are very unlikely ($p < \alpha$) to occur if there are not corresponding mean differences in the population. For each of the three tests, effect size is measured by computing the percentage of variance accounted for, η^2.

Figure 19.3 summarizes the statistical procedures used for data in category 3.

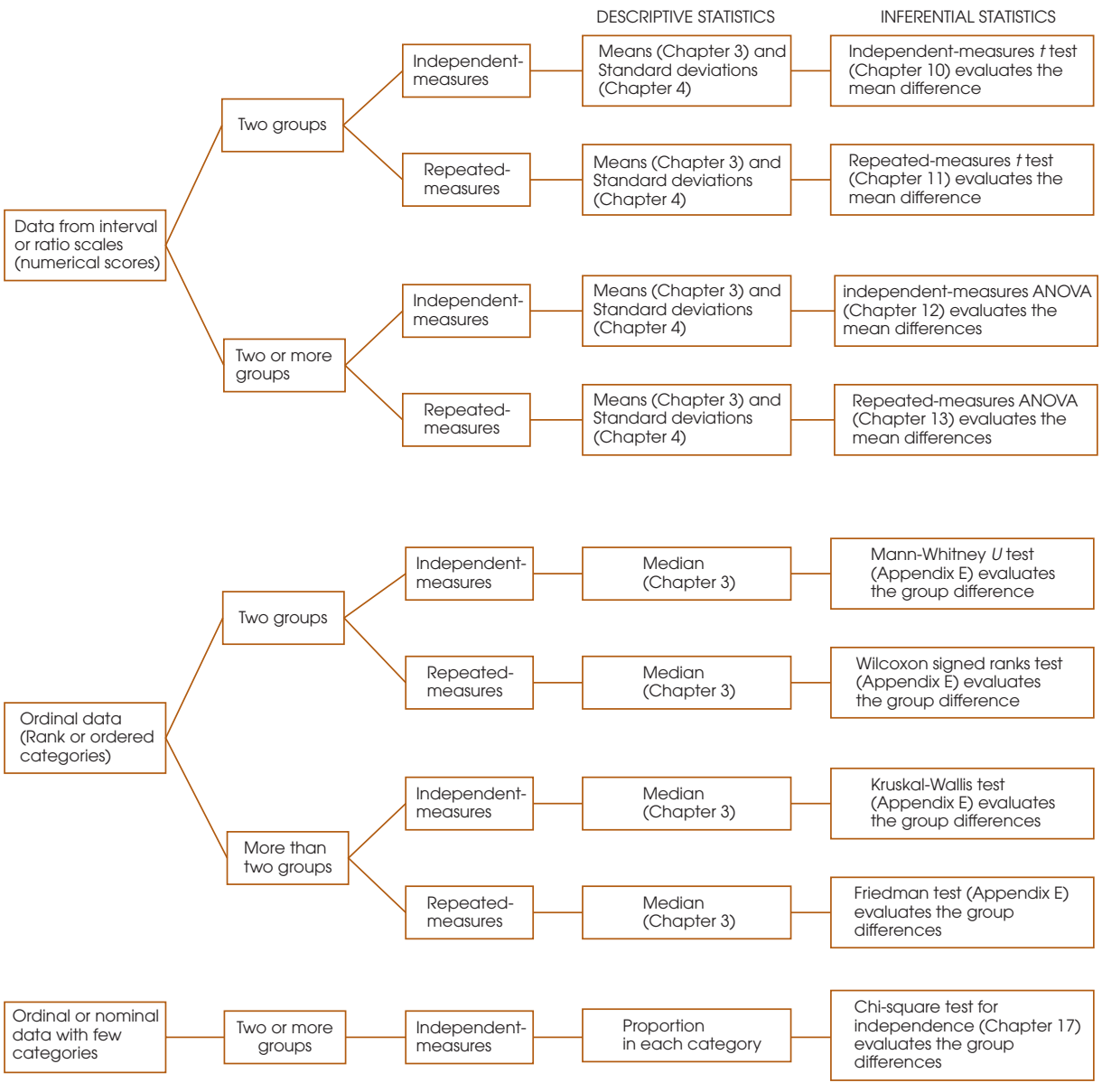


FIGURE 19.3

Statistics for Category 3 data. Two or more groups of scores with one score per participant. The goal is to describe and evaluate differences between groups of scores.

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravettter

You can find a tutorial quiz and other learning exercises for Chapter 19 on the book companion website.

aplia

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

PROBLEMS

Each problem describes a research situation and the data it produces. Identify the statistical procedures that are appropriate for the data. When possible, identify descriptive statistics, inferential statistics, and a measure of effect size.

1. Research suggests that the antioxidants in foods such as blueberries can reduce and even reverse age-related declines in cognitive functioning (Joseph et al., 1999). To test this phenomenon, a researcher selects a sample of $n = 25$ adults aged 70 to 75 and administers a cognitive function test to each participant. The participants then drink a blueberry supplement every day for 4 months before they are tested again. The researcher compares the scores before treatment with the scores after treatment to see if there is any change in cognitive function.
2. Recent budget cuts forced the city school district to increase class size in the elementary schools. To determine student reaction to the change, the district administered a survey to students asking whether the larger classes were "better, worse, or not different" from the classes the previous year. The results from the survey will be used to describe the students' attitude.
3. Last fall, the college introduced a peer-mentor program in which a sample of $n = 75$ freshmen was each assigned an upperclassman mentor. To evaluate the success of the program, the administration looked at the number of students who returned to the college to begin their second year. The data showed that 88% of the students in the peer-mentor program returned, compared to 72% for freshmen who were not in the program.
4. To examine the relationship between alcohol consumption and birth weight, a researcher selects a sample of $n = 20$ pregnant rats and mixes alcohol with their food for 2 weeks before the pups are born.

One newborn pup is randomly selected from each subject's litter and the average birth weight for the $n = 20$ pups is recorded. It is known that the average birth weight for regular rats (without exposure to alcohol) is $\mu = 5.6$ grams.

5. To examine the relationship between texting and driving skill, a researcher uses orange cones to set up a driving circuit in the high school parking lot. A group of students is then tested on the circuit, once while receiving and sending text messages and once without texting. For each student, the researcher records the number of orange cones hit while driving each circuit.

6. Childhood participation in sports, cultural groups, and youth groups appears to be related to improved self-esteem for adolescents (McGee, Williams, Howden-Chapman, Martin, & Kawachi, 2006). In a representative study, a researcher compares scores on a standardized self-esteem questionnaire for a sample of $n = 100$ adolescents with a history of group participation and a separate sample of $n = 100$ who have no history of group participation.

7. There is some evidence indicating that people with visible tattoos are viewed more negatively than people without visible tattoos (Resenhoeft, Villa, & Wiseman, 2008). In a similar study, a researcher showed male college students photographs of women and asked the students to rate the attractiveness of each woman using a 7-point scale. One of the women was selected as the target. For one group of participants, the target was photographed with a large tattoo on her shoulder and for a second group her photograph showed no tattoo. The researcher plans to compare the target's ratings for the two groups to determine whether the tattoo had any effect on perceived attractiveness.

8. A researcher investigated different combinations of temperature and humidity to examine how heat affects performance. The researcher compared three temperature conditions (70º, 80º, and 90º) with a high humidity and a low humidity condition for each temperature. A separate group of participants was tested in each of the six different conditions. For each participant, the researcher recorded the number of errors on a problem-solving task. The researcher would like to know how different combinations of temperature and humidity influence performance.

9. Hallam, Price, and Katsarou (2002) investigated the influence of background noise on classroom performance for children aged 10 to 12. In a similar study, students in one classroom worked on an arithmetic task with calming music in the background. Students in a second classroom heard aggressive, exciting music, and students in a third room had no music at all. The researchers measured the number of problems answered correctly for each student to determine whether the music conditions had any effect on performance.

10. A researcher is investigating the relationship between personality and birth order position. A sample of college students is classified into four birth-order categories (1st, 2nd, 3rd, 4th or later) and classified as being either extroverted or introverted.

11. A researcher is investigating the relationship between personality and birth order position. A sample of college students is classified into four birth-order categories (1st, 2nd, 3rd, 4th or later) and given a personality test that measures the degree of extroversion on a 50-point scale.

12. A survey of female high school seniors includes one question asking for the amount of time spent on clothes, hair, and makeup each morning before school. The researcher plans to use the results as part of a general description of today's high school students.

13. Brunt, Rhee, and Zhong (2008) surveyed 557 undergraduate college students to examine their weight status, health behaviors, and diet. In a similar study, researchers used body mass index (BMI) to classify a group of students into four categories: underweight, healthy weight, overweight, and obese. The students were also surveyed to determine the number of fatty and/or sugary snacks they eat each day. The researchers would like to use to data to determine whether there is a relationship between weight status and diet.

14. A researcher would like to determine whether infants, age 2 to 3 months, show any evidence of color preference. The babies are positioned in front of a screen on which a set of four colored patches is presented. The four colors are red, green, blue, and yellow. The researcher measures the amount of time each infant looks at each of the four colors during a 30 second test period. The color with the greatest time is identified as the preferred color for the child.

15. A researcher administers a survey to graduating seniors, asking them to rate their optimism about the current job market on a 7-point scale. The researcher plans to use the results as part of a description of today's graduating seniors.

16. Standardized measures seem to indicate that the average level of anxiety has increased gradually over the past 50 years (Twenge, 2000). In the 1950s, the average score on the Child Manifest Anxiety Scale was $\mu = 15.1$. A researcher administers the same test

to a sample of $n = 50$ of today's children to determine whether there has been a significant change in the average anxiety level.

17. Belsky, Weinraub, Owen, and Kelly (2001) reported on the effects of preschool childcare on the development of young children. One result suggests that children who spend more time away from their mothers are more likely to show behavioral problems in kindergarten. Suppose that a kindergarten teacher is asked rank order the degree of disruptive behavior for the $n = 20$ children in the class.

a. Researchers then separate the students into two groups: children with a history of preschool and children with little or no experience in preschool. The researchers plan to compare the ranks for the two groups.

b. The researchers interview each child's parents to determine how much time the child spent in preschool. The children are then ranked according to the amount of preschool experience. The researchers plan to use the data to determine whether there is a relationship between preschool experience and disruptive behavior.

18. McGee and Shevlin (2009) found that an individual's sense of humor had a significant effect on how attractive the individual was perceived to be by others. In a similar study, female college students were given brief descriptions of three potential romantic partners. One was identified as the target and was described positively as being single, ambitious, and having good job prospects. For half of the participants, the description also said that he had a great sense of humor. For another half, it said that he has no sense of humor. After reading the three descriptions, the participants were asked to rank the three men 1st, 2nd, and 3rd in terms of attractiveness. For each of the two groups, the researchers recorded the number of times the target was placed in each ordinal position.

19. Numerous studies have found that males report higher self-esteem than females, especially for adolescents (Kling, Hyde, Showers, & Buswell, 1999). A recent study found that males scored an average of 8 points higher than females on a standardized questionnaire measuring self-esteem. The researcher would like to know whether this is a significant difference.

20. Research has demonstrated that IQ scores have been increasing, generation by generation, for years (Flynn, 1999). A researcher would like to determine whether this trend can be described by a linear equation showing the relationship between age and IQ scores. The same IQ test is given to a sample of 100 adults who range in age from 20 to 85 years. The age and IQ score are recorded for each person.

21. A researcher is investigating the effectiveness of acupuncture treatment for chronic back pain. A sample of $n = 20$ participants is obtained from a pain clinic. Each individual rates the current level of pain and then begins a 6-week program of acupuncture treatment. At the end of the program, the pain level is rated again and the researcher records whether the pain has increased or decreased for each participant.

22. Research results indicate that physically attractive people are also perceived as being more intelligent (Eagly, Ashmore, Makhijani, & Longo, 1991). As a demonstration of this phenomenon, a researcher obtained a set of $n = 25$ photographs of male college students. The photographs were shown to a sample of female college students who used a 7-point scale to rate several characteristics, including intelligence and attractiveness, for the person in each photo. The average attractiveness rating and the average intelligence rating were computed for each photograph. The researcher plans to use the averages to determine whether there is relationship between perceived attractiveness and perceived intelligence.

23. Research has shown that people are more likely to show dishonest and self-interested behaviors in darkness than in a well-lit environment (Zhong, Bohns, & Gino, 2010). In a related experiment, students were given a quiz and then asked to grade their own papers while the teacher read the correct answers. One group of students was tested in a well-lit room and another group was tested in a dimly-lit room. The researchers recorded the number of correct answers reported by each student to determine whether there was a significant difference between the two groups.

24. There is some evidence suggesting that you are likely to improve your test score if you rethink and change answers on a multiple-choice exam (Johnston, 1975). To examine this phenomenon, a teacher encouraged students to reconsider their answers before turning in exams. Students were asked to record their original answers and the changes that they made. When the exams were collected, the teacher found that 18 students improved their grades by changing answers and only 7 students had lower grades with the changes. The teacher would like to know if this is a statistically significant result.

25. A researcher is evaluating customer satisfaction with the service and coverage of three phone carriers. Each individual in a sample of $n = 25$ uses one carrier for 2 weeks, then switches to another for 2 weeks, and

finally switches to the third for 2 weeks. Each participant then rates the three carriers.

a. Assume that each carrier was rated on a 10-point scale.

b. Assume that each participant ranked the three carriers 1st, 2nd and 3rd.

c. Assume the each participant simply identified the most preferred carrier of the three.

26. There is some research indicating that college students who use Facebook while studying tend to have lower grades than non-users (Kirschner & Karpinski, 2010). A representative study surveys students to determine the amount of Facebook use during the time they are studying or doing homework. Based on the amount of time spent on Facebook, students are classified into three groups (high, medium, and low time) and their grade point averages are recorded. The researcher would like to examine the relationship between grades and amount of time on Facebook.

27. To examine the effect of sleep deprivation on motor-skills performance, a sample of $n = 10$ participants was tested on a motor-skills task after 24 hours of sleep deprivation, tested again after 36 hours, and tested once more after 48 hours. The dependent variable is the number of errors made on the motor-skills task.

28. Ryan and Hemmes (2005) examined how homework assignments are related to learning. The participants were college students enrolled in a class with weekly homework assignments and quizzes. For some weeks, the homework was required and counted toward the student's grade. Other weeks, the homework was optional and did not count toward the student's grade. Predictably, most students completed the required homework assignments and did not do the optional assignments. For each student, the researchers recorded the average quiz grade for weeks with required homework and the average grade for weeks with optional homework to determine whether the grades were significantly higher when homework was required and actually done.

29. Ford and Torok (2008) found that motivational signs were effective in increasing physical activity on a college campus. In a similar study, researchers first counted the number of students and faculty who used the stairs and the number who used the elevators in a college building during a 30-minute observation period. The following week, signs such as "Step up to a healthier lifestyle" and "An average person burns 10 calories a minute walking up the stairs" were posted by the elevators and stairs and the researchers once again counted people to determine whether the signs had a significant effect on behavior.

APPENDIX A Basic Mathematics Review

Preview

This appendix reviews some of the basic math skills that are necessary for the statistical calculations presented in this book. Many students already will know some or all of this material. Others will need to do extensive work and review. To help you assess your own skills, we include a skills assessment exam here. You should allow approximately 30 minutes to complete the test. When you finish, grade your test using the answer key on page 697.

Notice that the test is divided into five sections. If you miss more than three questions in any section of the test, you probably need help in that area. Turn to the section of this appendix that corresponds to your problem area. In each section, you will find a general review, examples, and additional practice problems. After reviewing the appropriate section and doing the practice problems, turn to the end of the appendix. You will find another version of the skills assessment exam. If you still miss more than three questions in any section of the exam, continue studying. Get assistance from an instructor or a tutor if necessary. At the end of this appendix is a list of recommended books for individuals who need a more extensive review than can be provided here. We stress that mastering this material now will make the rest of the course much easier.

SKILLS ASSESSMENT PREVIEW EXAM

SECTION 1

(corresponding to Section A.1 of this appendix)

1. $3 + 2 \times 7 = ?$
2. $(3 + 2) \times 7 = ?$
3. $3 + 2^2 - 1 = ?$
4. $(3 + 2)^2 - 1 = ?$
5. $12/4 + 2 = ?$
6. $12/(4 + 2) = ?$
7. $12/(4 + 2)^2 = ?$
8. $2 \times (8 - 2^2) = ?$
9. $2 \times (8 - 2)^2 = ?$
10. $3 \times 2 + 8 - 1 \times 6 = ?$
11. $3 \times (2 + 8) - 1 \times 6 = ?$
12. $3 \times 2 + (8 - 1) \times 6 = ?$

SECTION 2

(corresponding to Section A.2 of this appendix)

1. The fraction $\frac{3}{4}$ corresponds to a percentage of _____.
2. Express 30% as a fraction.
3. Convert $\frac{12}{40}$ to a decimal.
4. $\frac{2}{13} + \frac{8}{13} = ?$
5. $1.375 + 0.25 = ?$
6. $\frac{2}{5} \times \frac{1}{4} = ?$
7. $\frac{1}{8} + \frac{2}{3} = ?$
8. $3.5 \times 0.4 = ?$
9. $\frac{1}{5} \div \frac{3}{4} = ?$
10. $3.75/0.5 = ?$
11. In a group of 80 students, 20% are psychology majors. How many psychology majors are in this group?
12. A company reports that two-fifths of its employees are women. If there are 90 employees, how many are women?

SECTION 3

(corresponding to Section A.3 of this appendix)

1. $3 + (-2) + (-1) + 4 = ?$
2. $6 - (-2) = ?$
3. $-2 - (-4) = ?$
4. $6 + (-1) - 3 - (-2) - (-5) = ?$
5. $4 \times (-3) = ?$
6. $-2 \times (-6) = ?$
7. $-3 \times 5 = ?$
8. $-2 \times (-4) \times (-3) = ?$
9. $12 \div (-3) = ?$
10. $-18 \div (-6) = ?$
11. $-16 \div 8 = ?$
12. $-100 \div (-4) = ?$

SECTION 4

(corresponding to Section A.4 of this appendix)

For each equation, find the value of X.

1. $X + 6 = 13$
2. $X - 14 = 15$
3. $5 = X - 4$
4. $3X = 12$
5. $72 = 3X$
6. $X/5 = 3$
7. $10 = X/8$
8. $3X + 5 = -4$
9. $24 = 2X + 2$
10. $(X + 3)/2 = 14$
11. $(X - 5)/3 = 2$
12. $17 = 4X - 11$

SECTION 5

(corresponding to Section A.5 of this appendix)

1. $4^3 = ?$
2. $\sqrt{25 - 9} = ?$
3. If $X = 2$ and $Y = 3$, then $XY^3 = ?$
4. If $X = 2$ and $Y = 3$, then $(X + Y)^2 = ?$
5. If $a = 3$ and $b = 2$, then $a^2 + b^2 = ?$
6. $(-3)^3 = ?$
7. $(-4)^4 = ?$
8. $\sqrt{4} \times 4 = ?$
9. $36/\sqrt{9} = ?$
10. $(9 + 2)2 = ?$
11. $5^2 + 2^3 = ?$
12. If $a = 3$ and $b = -1$, then $a^2b^3 = ?$

The answers to the skills assessment exam are at the end of the appendix (pages 697−698).

A.1 SYMBOLS AND NOTATION

Table A.1 presents the basic mathematical symbols that you should know, along with examples of their use. Statistical symbols and notation are introduced and explained throughout this book as they are needed. Notation for exponents and square roots is covered separately at the end of this appendix.

Parentheses are a useful notation because they specify and control the order of computations. Everything inside the parentheses is calculated first. For example,

$$(5 + 3) \times 2 = 8 \times 2 = 16$$

Changing the placement of the parentheses also changes the order of calculations. For example,

$$5 + (3 \times 2) = 5 + 6 = 11$$

ORDER OF OPERATIONS

Often a formula or a mathematical expression will involve several different arithmetic operations, such as adding, multiplying, squaring, and so on. When you encounter these situations, you must perform the different operations in the correct sequence. Following is a list of mathematical operations, showing the order in which they are to be performed.

1. Any calculation contained within parentheses is done first.
2. Squaring (or raising to other exponents) is done second.
3. Multiplying and/or dividing is done third. A series of multiplication and/or division operations should be done in order from left to right.
4. Adding and/or subtracting is done fourth.

The following examples demonstrate how this sequence of operations is applied in different situations.

To evaluate the expression

$$(3 + 1)^2 - 4 \times 7/2$$

first, perform the calculation within parentheses:

$$(4)^2 - 4 \times 7/2$$

Next, square the value as indicated:

$$16 - 4 \times 7/2$$

TABLE A.1

Symbol	Meaning	Example
$+$	Addition	$5 + 7 = 12$
$-$	Subtraction	$8 - 3 = 5$
$\times$, ( )	Multiplication	$3 \times 9 = 27$, $3(9) = 27$
$\div$, /	Division	$15 \div 3 = 5$, $15/3 = 5$, $\frac{15}{3} = 5$
$>$	Greater than	$20 > 10$
$<$	Less than	$7 < 11$
$\neq$	Not equal to	$5 \neq 6$

Then perform the multiplication and division:

$$16 - 14$$

Finally, do the subtraction:

$$16 - 14 = 2$$

A sequence of operations involving multiplication and division should be performed in order from left to right. For example, to compute $12/2 \times 3$, you divide 12 by 2 and then multiply the result by 3:

$$12/2 \times 3 = 6 \times 3 = 18$$

Notice that violating the left-to-right sequence can change the result. For this example, if you multiply before dividing, you will obtain

$$12/2 \times 3 = 12/6 = 2 \qquad \text{(This is wrong.)}$$

A sequence of operations involving only addition and subtraction can be performed in any order. For example, to compute $3 + 8 - 5$, you can add 3 and 8 and then subtract 5:

$$(3 + 8) - 5 = 11 - 5 = 6$$

or you can subtract 5 from 8 and then add the result to 3:

$$3 + (8 - 5) = 3 + 3 = 6$$

A mathematical expression or formula is simply a concise way to write a set of instructions. When you evaluate an expression by performing the calculation, you simply follow the instructions. For example, assume you are given these instructions:

1. First, add 3 and 8.
2. Next, square the result.
3. Next, multiply the resulting value by 6.
4. Finally, subtract 50 from the value you have obtained.

You can write these instructions as a mathematical expression.

1. The first step involves addition. Because addition is normally done last, use parentheses to give this operation priority in the sequence of calculations:

   $$(3 + 8)$$

2. The instruction to square a value is noted by using the exponent 2 beside the value to be squared:

   $$(3 + 8)^2$$

3. Because squaring has priority over multiplication, you can simply introduce the multiplication into the expression:

   $$6 \times (3 + 8)^2$$

4. Addition and subtraction are done last, so simply write in the requested subtraction:

$$6 \times (3 + 8)^2 - 50$$

To calculate the value of the expression, you work through the sequence of operations in the proper order:

$$\begin{aligned} 6 \times (3 + 8)^2 - 50 &= 6 \times (11)^2 - 50 \\ &= 6 \times (121) - 50 \\ &= 726 - 50 \\ &= 676 \end{aligned}$$

As a final note, you should realize that the operation of squaring (or raising to any exponent) applies only to the value that immediately precedes the exponent. For example,

$$2 \times 3^2 = 2 \times 9 = 18 \qquad \text{(Only the 3 is squared.)}$$

If the instructions require multiplying values and then squaring the product, you must use parentheses to give the multiplication priority over squaring. For example, to multiply 2 times 3 and then square the product, you would write

$$(2 \times 3)^2 = (6)^2 = 36$$

LEARNING CHECK

1. Evaluate each of the following expressions:

a. $4 \times 8/2^2$
b. $4 \times (8/2)^2$
c. $100 - 3 \times 12/(6 - 4)^2$
d. $(4 + 6) \times (3 - 1)^2$
e. $(8 - 2)/(9 - 8)^2$
f. $6 + (4 - 1)^2 - 3\_4^2$
g. $4 \times (8 - 3) + 8 - 3$

ANSWERS **1. a.** 8 **b.** 64 **c.** 91 **d.** 40 **e.** 6 **f.** −33 **g.** 25

A.2 PROPORTIONS: FRACTIONS, DECIMALS, AND PERCENTAGES

A proportion is a part of a whole and can be expressed as a fraction, a decimal, or a percentage. For example, in a class of 40 students, only 3 failed the final exam.

The proportion of the class that failed can be expressed as a fraction

$$\text{fraction} = \frac{3}{40}$$

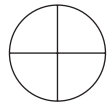
or as a decimal value

$$\text{decimal} = 0.075$$

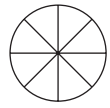
or as a percentage

$$\text{percentage} = 7.5\%$$

In a fraction, such as $\frac{3}{4}$, the bottom value (the denominator) indicates the number of equal pieces into which the whole is split. Here the "pie" is split into 4 equal pieces:

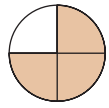


If the denominator has a larger value—say, 8—then each piece of the whole pie is smaller:



A larger denominator indicates a smaller fraction of the whole.

The value on top of the fraction (the numerator) indicates how many pieces of the whole are being considered. Thus, the fraction $\frac{3}{4}$ indicates that the whole is split evenly into 4 pieces and that 3 of them are being used:



A fraction is simply a concise way of stating a proportion: "Three out of four" is equivalent to $\frac{3}{4}$. To convert the fraction to a decimal, you divide the numerator by the denominator:

$$\frac{3}{4} = 3 \div 4 = 0.75$$

To convert the decimal to a percentage, simply multiply by 100, and place a percent sign (%) after the answer:

$$0.75 \times 100 = 75\%$$

The U.S. money system is a convenient way of illustrating the relationship between fractions and decimals. "One quarter," for example, is one-fourth $\left(\frac{1}{4}\right)$ of a dollar, and its decimal equivalent is 0.25. Other familiar equivalencies are as follows:

	Dime	Quarter	50 Cents	75 Cents
Fraction	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
Decimal	0.10	0.25	0.50	0.75
Percentage	10%	25%	50%	75%

FRACTIONS

1. **Finding Equivalent Fractions.** The same proportional value can be expressed by many equivalent fractions. For example,

$$\frac{1}{2} = \frac{2}{4} = \frac{10}{20} = \frac{50}{100}$$

To create equivalent fractions, you can multiply the numerator and denominator by the same value. As long as both the numerator and the denominator of the fraction are multiplied by the same value, the new fraction will be equivalent to the original. For example,

$$\frac{3}{10} = \frac{9}{30}$$

because both the numerator and the denominator of the original fraction have been multiplied by 3. Dividing the numerator and denominator of a fraction by the same value will also result in an equivalent fraction. By using division, you can reduce a fraction to a simpler form. For example,

$$\frac{40}{100} = \frac{2}{5}$$

because both the numerator and the denominator of the original fraction have been divided by 20.

You can use these rules to find specific equivalent fractions. For example, find the fraction that has a denominator of 100 and is equivalent to $\frac{3}{4}$. That is,

$$\frac{3}{4} = \frac{?}{100}$$

Notice that the denominator of the original fraction must be multiplied by 25 to produce the denominator of the desired fraction. For the two fractions to be equal, both the numerator and the denominator must be multiplied by the same number. Therefore, we also multiply the top of the original fraction by 25 and obtain

$$\frac{3 \times 25}{4 \times 25} = \frac{75}{100}$$

2. **Multiplying Fractions.** To multiply two fractions, you first multiply the numerators and then multiply the denominators. For example,

$$\frac{3}{4} \times \frac{5}{7} = \frac{3 \times 5}{4 \times 7} = \frac{15}{28}$$

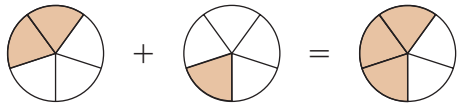
3. **Dividing Fractions.** To divide one fraction by another, you invert the second fraction and then multiply. For example,

$$\frac{1}{2} \div \frac{1}{4} = \frac{1}{2} \times \frac{4}{1} = \frac{1 \times 4}{2 \times 1} = \frac{4}{2} = \frac{2}{1} = 2$$

4. **Adding and Subtracting Fractions.** Fractions must have the same denominator before you can add or subtract them. If the two fractions already have a common denominator, you simply add (or subtract as the case may be) *only* the values in the numerators. For example,

$$\frac{2}{5} + \frac{1}{5} = \frac{3}{5}$$

Suppose you divided a pie into five equal pieces (fifths). If you first ate two-fifths of the pie and then another one-fifth, the total amount eaten would be three-fifths of the pie:



If the two fractions do not have the same denominator, you must first find equivalent fractions with a common denominator before you can add or subtract. The product of the two denominators will always work as a common denominator for equivalent fractions (although it may not be the lowest common denominator). For example,

$$\frac{2}{3} + \frac{1}{10} = ?$$

Because these two fractions have different denominators, it is necessary to convert each into an equivalent fraction and find a common denominator. We will use $3 \times 10 = 30$ as the common denominator. Thus, the equivalent fraction of each is

$$\frac{2}{3} = \frac{20}{30} \quad \text{and} \quad \frac{1}{10} = \frac{3}{30}$$

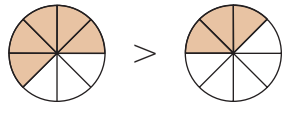
Now the two fractions can be added:

$$\frac{20}{30} + \frac{3}{30} = \frac{23}{30}$$

5. **Comparing the Size of Fractions.** When comparing the size of two fractions with the same denominator, the larger fraction will have the larger numerator. For example,

$$\frac{5}{8} > \frac{3}{8}$$

The denominators are the same, so the whole is partitioned into pieces of the same size. Five of these pieces are more than three of them:



When two fractions have different denominators, you must first convert them to fractions with a common denominator to determine which is larger. Consider the following fractions:

$$\frac{3}{8} \quad \text{and} \quad \frac{7}{16}$$

If the numerator and denominator of $\frac{3}{8}$ are multiplied by 2, the resulting equivalent fraction will have a denominator of 16:

$$\frac{3}{8} = \frac{3 \times 2}{8 \times 2} = \frac{6}{16}$$

Now a comparison can be made between the two fractions:

$$\frac{6}{16} < \frac{7}{16}$$

Therefore,

$$\frac{3}{8} < \frac{7}{16}$$

DECIMALS

1. **Converting Decimals to Fractions.** Like a fraction, a decimal represents part of the whole. The first decimal place to the right of the decimal point indicates how many tenths are used. For example,

$$0.1 = \frac{1}{10} \qquad 0.7 = \frac{7}{10}$$

The next decimal place represents $\frac{1}{100}$, the next $\frac{1}{1000}$, the next $\frac{1}{10{,}000}$, and so on. To change a decimal to a fraction, just use the number without the decimal point for the numerator. Use the denominator that the last (on the right) decimal place represents. For example,

$$0.32 = \frac{32}{100} \qquad 0.5333 = \frac{5333}{10{,}000} \qquad 0.05 = \frac{5}{100} \qquad 0.001 = \frac{1}{1000}$$

2. **Adding and Subtracting Decimals.** To add and subtract decimals, the only rule is that you must keep the decimal points in a straight vertical line. For example,

$$\begin{array}{r} 0.27 \\ +1.326 \\ \hline 1.596 \end{array} \qquad \begin{array}{r} 3.595 \\ -0.67 \\ \hline 2.925 \end{array}$$

3. **Multiplying Decimals.** To multiply two decimal values, you first multiply the two numbers, ignoring the decimal points. Then you position the decimal point in the answer so that the number of digits to the right of the decimal point is equal to the total number of decimal places in the two numbers being multiplied. For example,

$$\begin{array}{rl} 1.73 & \text{(two decimal places)} \\ \times 0.251 & \text{(three decimal places)} \\ \hline 173 & \\ 865 & \\ 346 & \\ \hline 0.43423 & \text{(five decimal places)} \end{array} \qquad \begin{array}{rl} 0.25 & \text{(two decimal places)} \\ \times 0.005 & \text{(three decimal places)} \\ \hline 125 & \\ 00 & \\ 00 & \\ \hline 0.00125 & \text{(five decimal places)} \end{array}$$

4. **Dividing Decimals.** The simplest procedure for dividing decimals is based on the fact that dividing two numbers is identical to expressing them as a fraction:

$$0.25 \div 1.6 \text{ is identical to } \frac{0.25}{1.6}$$

You now can multiply both the numerator and the denominator of the fraction by 10, 100, 1000, or whatever number is necessary to remove the decimal places. Remember that multiplying both the numerator and the denominator of a fraction by the *same* value will create an equivalent fraction. Therefore,

$$\frac{0.25}{1.6} = \frac{0.25 \times 100}{1.6 \times 100} = \frac{25}{160} = \frac{5}{32}$$

The result is a division problem without any decimal places in the two numbers.

PERCENTAGES

1. **Converting a Percentage to a Fraction or a Decimal.** To convert a percentage to a fraction, remove the percent sign, place the number in the numerator, and use 100 for the denominator. For example,

$$52\% = \frac{52}{100} \qquad 5\% = \frac{5}{100}$$

To convert a percentage to a decimal, remove the percent sign and divide by 100, or simply move the decimal point two places to the left. For example,

$$83\% = 83. = 0.83$$

$$14.5\% = 14.5 = 0.145$$

$$5\% = 5. = 0.05$$

2. **Performing Arithmetic Operations with Percentages.** There are situations in which it is best to express percent values as decimals in order to perform certain arithmetic operations. For example, what is 45% of 60? This question may be stated as

$$45\% \times 60 = ?$$

The 45% should be converted to decimal form to find the solution to this question. Therefore,

$$0.45 \times 60 = 27$$

LEARNING CHECK

1. Convert $\frac{3}{25}$ to a decimal.
2. Convert $\frac{3}{8}$ to a percentage.
3. Next to each set of fractions, write "True" if they are equivalent and "False" if they are not:
   - **a.** $\frac{3}{8} = \frac{9}{24}$ _____
   - **b.** $\frac{7}{9} = \frac{17}{19}$ _____
   - **c.** $\frac{2}{7} = \frac{4}{14}$ _____

4. Compute the following:

a. $\frac{1}{6} \times \frac{7}{10}$ **b.** $\frac{7}{8} - \frac{1}{2}$ **c.** $\frac{9}{10} \div \frac{2}{3}$ **d.** $\frac{7}{22} + \frac{2}{3}$

5. Identify the larger fraction of each pair:

a. $\frac{7}{10}, \frac{21}{100}$ **b.** $\frac{3}{4}, \frac{7}{12}$ **c.** $\frac{22}{3}, \frac{19}{3}$

6. Convert the following decimals into fractions:

a. 0.012 **b.** 0.77 **c.** 0.005

7. $2.59 \times 0.015 = ?$

8. $1.8 \div 0.02 = ?$

9. What is 28% of 45?

ANSWERS **1.** 0.12 **2.** 37.5% **3. a.** True **b.** False **c.** True

4. a. $\frac{7}{60}$ **b.** $\frac{3}{8}$ **c.** $\frac{27}{20}$ **d.** $\frac{65}{66}$ **5. a.** $\frac{7}{10}$ **b.** $\frac{3}{4}$ **c.** $\frac{22}{3}$

6. a. $\frac{12}{1000} = \frac{3}{250}$ **b.** $\frac{77}{100}$ **c.** $\frac{5}{1000} = \frac{1}{200}$ **7.** 0.03885 **8.** 90 **9.** 12.6

A.3 NEGATIVE NUMBERS

Negative numbers are used to represent values less than zero. Negative numbers may occur when you are measuring the difference between two scores. For example, a researcher may want to evaluate the effectiveness of a propaganda film by measuring people's attitudes with a test both before and after viewing the film:

	Before	After	Amount of Change
Person A	23	27	+4
Person B	18	15	−3
Person C	21	16	−5

Notice that the negative sign provides information about the direction of the difference: A plus sign indicates an increase in value, and a minus sign indicates a decrease.

Because negative numbers are frequently encountered, you should be comfortable working with these values. This section reviews basic arithmetic operations using negative numbers. You should also note that any number without a sign (+ or −) is assumed to be positive.

1. **Adding Negative Numbers.** When adding numbers that include negative values, simply interpret the negative sign as subtraction. For example,

$$3 + (-2) + 5 = 3 - 2 + 5 = 6$$

When adding a long string of numbers, it often is easier to add all the positive values to obtain the positive sum and then to add all of the negative values to obtain the negative sum. Finally, you subtract the negative sum from the positive sum. For example,

$$-1 + 3 + (-4) + 3 + (-6) + (-2)$$

positive sum = 6 negative sum = 13

Answer: $6 - 13 = -7$

2. **Subtracting Negative Numbers.** To subtract a negative number, change it to a positive number, and add. For example,

$$4 - (-3) = 4 + 3 = 7$$

This rule is easier to understand if you think of positive numbers as financial gains and negative numbers as financial losses. In this context, taking away a debt is equivalent to a financial gain. In mathematical terms, taking away a negative number is equivalent to adding a positive number. For example, suppose you are meeting a friend for lunch. You have \$7, but you owe your friend \$3. Thus, you really have only \$4 to spend for lunch. But your friend forgives (takes away) the \$3 debt. The result is that you now have \$7 to spend. Expressed as an equation,

$$\$4 \text{ minus a } \$3 \text{ debt} = \$7$$

$$4 - (-3) = 4 + 3 = 7$$

3. **Multiplying and Dividing Negative Numbers.** When the two numbers being multiplied (or divided) have the same sign, the result is a positive number. When the two numbers have different signs, the result is negative. For example,

$$3 \times (-2) = -6$$

$$-4 \times (-2) = +8$$

The first example is easy to explain by thinking of multiplication as repeated addition. In this case,

$$3 \times (-2) = (-2) + (-2) + (-2) = -6$$

You add three negative 2s, which results in a total of negative 6. In the second example, we are multiplying by a negative number. This amounts to repeated subtraction. That is,

$$\begin{aligned} -4 \times (-2) &= -(-2) - (-2) - (-2) - (-2) \\ &= 2 + 2 + 2 + 2 = 8 \end{aligned}$$

By using the same rule for both multiplication and division, we ensure that these two operations are compatible. For example,

$$-6 \div 3 = -2$$

which is compatible with

$$3 \times (-2) = -6$$

Also,

$$8 \div (-4) = -2$$

which is compatible with

$$-4 \times (-2) = +8$$

LEARNING CHECK

1. Complete the following calculations:

a. $3 + (-8) + 5 + 7 + (-1) + (-3)$
b. $5 - (-9) + 2 - (-3) - (-1)$
c. $3 - 7 - (-21) + (-5) - (-9)$
d. $4 - (-6) - 3 + 11 - 14$
e. $9 + 8 - 2 - 1 - (-6)$
f. $9 \times (-3)$
g. $-7 \times (-4)$
h. $-6 \times (-2) \times (-3)$
i. $-12 \div (-3)$
j. $18 \div (-6)$

ANSWERS **1. a.** 3 **b.** 20 **c.** 21 **d.** 4 **e.** 20
f. −27 **g.** 28 **h.** −36 **i.** 4 **j.** −3

A.4 BASIC ALGEBRA: SOLVING EQUATIONS

An equation is a mathematical statement that indicates two quantities are identical. For example,

$$12 = 8 + 4$$

Often an equation will contain an unknown (or variable) quantity that is identified with a letter or symbol, rather than a number. For example,

$$12 = 8 + X$$

In this event, your task is to find the value of X that makes the equation "true," or balanced. For this example, an X value of 4 will make a true equation. Finding the value of X is usually called *solving the equation*.

To solve an equation, there are two points to keep in mind:

1. Your goal is to have the unknown value (X) isolated on one side of the equation. This means that you need to remove all of the other numbers and symbols that appear on the same side of the equation as the X.
2. The equation remains balanced, provided you treat both sides exactly the same. For example, you could add 10 points to *both* sides, and the solution (the X value) for the equation would be unchanged.

FINDING THE SOLUTION FOR AN EQUATION

We will consider four basic types of equations and the operations needed to solve them.

1. **When X Has a Value Added to It.** An example of this type of equation is

$$X + 3 = 7$$

Your goal is to isolate X on one side of the equation. Thus, you must remove the $+3$ on the left-hand side. The solution is obtained by subtracting 3 from *both* sides of the equation:

$$X + 3 - 3 = 7 - 3$$

$$X = 4$$

The solution is $X = 4$. You should always check your solution by returning to the original equation and replacing X with the value you obtained for the solution. For this example,

$$X + 3 = 7$$

$$4 + 3 = 7$$

$$7 = 7$$

2. **When X Has a Value Subtracted From It.** An example of this type of equation is

$$X - 8 = 12$$

In this example, you must remove the -8 from the left-hand side. Thus, the solution is obtained by adding 8 to *both* sides of the equation:

$$X - 8 + 8 = 12 + 8$$

$$X = 20$$

Check the solution:

$$X - 8 = 12$$

$$20 - 8 = 12$$

$$12 = 12$$

3. **When X Is Multiplied by a Value.** An example of this type of equation is

$$4X = 24$$

In this instance, it is necessary to remove the 4 that is multiplied by X. This may be accomplished by dividing both sides of the equation by 4:

$$\frac{4X}{4} = \frac{24}{4}$$

$$X = 6$$

Check the solution:

$$4X = 24$$

$$4(6) = 24$$

$$24 = 24$$

4. **When X Is Divided by a Value.** An example of this type of equation is

$$\frac{X}{3} = 9$$

Now the X is divided by 3, so the solution is obtained by multiplying by 3. Multiplying both sides yields

$$3\left(\frac{X}{3}\right) = 9(3)$$

$$X = 27$$

For the check,

$$\frac{X}{3} = 9$$

$$\frac{27}{3} = 9$$

$$9 = 9$$

SOLUTIONS FOR MORE COMPLEX EQUATIONS

More complex equations can be solved by using a combination of the preceding simple operations. Remember that at each stage you are trying to isolate X on one side of the equation. For example,

$$3X + 7 = 22$$

$$3X + 7 - 7 = 22 - 7 \qquad \text{(Remove } +7 \text{ by subtracting 7 from both sides.)}$$

$$3X = 15$$

$$\frac{3X}{3} = \frac{15}{3} \qquad \text{(Remove 3 by dividing both sides by 3.)}$$

$$X = 5$$

To check this solution, return to the original equation, and substitute 5 in place of X:

$$3X + 7 = 22$$

$$3(5) + 7 = 22$$

$$15 + 7 = 22$$

$$22 = 22$$

Following is another type of complex equation frequently encountered in statistics:

$$\frac{X + 3}{4} = 2$$

First, remove the 4 by multiplying both sides by 4:

$$4\left(\frac{X + 3}{4}\right) = 2(4)$$

$$X + 3 = 8$$

Now remove the $+3$ by subtracting 3 from both sides:

$$X + 3 - 3 = 8 - 3$$

$$X = 5$$

To check this solution, return to the original equation, and substitute 5 in place of X:

$$\frac{X+3}{4} = 2$$

$$\frac{5+3}{4} = 2$$

$$\frac{8}{4} = 2$$

$$2 = 2$$

LEARNING CHECK

1. Solve for X, and check the solutions:

a. $3X = 18$ **b.** $X + 7 = 9$ **c.** $X - 4 = 18$ **d.** $5X - 8 = 12$

e. $\frac{X}{9} = 5$ **f.** $\frac{X+1}{6} = 4$ **g.** $X + 2 = -5$ **h.** $\frac{X}{5} = -5$

i. $\frac{2X}{3} = 12$ **j.** $\frac{X}{3} + 1 = 3$

ANSWERS **1. a.** $X = 6$ **b.** $X = 2$ **c.** $X = 22$ **d.** $X = 4$ **e.** $X = 45$

f. $X = 23$ **g.** $X = -7$ **h.** $X = -25$ **i.** $X = 18$ **j.** $X = 6$

A.5 EXPONENTS AND SQUARE ROOTS

EXPONENTIAL NOTATION

A simplified notation is used whenever a number is being multiplied by itself. The notation consists of placing a value, called an *exponent,* on the right-hand side of and raised above another number, called a *base*. For example,

$7^3 \leftarrow$ exponent
$\uparrow$
base

The exponent indicates how many times the base is used as a factor in multiplication. Following are some examples:

$7^3 = 7(7)(7)$	(Read "7 cubed" or "7 raised to the third power")
$5^2 = 5(5)$	(Read "5 squared")
$2^5 = 2(2)(2)(2)(2)$	(Read "2 raised to the fifth power")

There are a few basic rules about exponents that you will need to know for this course. They are outlined here.

1. **Numbers Raised to One or Zero.** Any number raised to the first power equals itself. For example,

$$6^1 = 6$$

Any number (except zero) raised to the zero power equals 1. For example,

$$9^0 = 1$$

2. **Exponents for Multiple Terms**. The exponent applies only to the base that is just in front of it. For example,

$$XY^2 = XYY$$

$$a^2b^3 = aabbb$$

3. **Negative Bases Raised to an Exponent.** If a negative number is raised to a power, then the result will be positive for exponents that are even and negative for exponents that are odd. For example,

$$\begin{aligned}(-4)^3 &= -4(-4)(-4)\\ &= 16(-4)\\ &= -64\end{aligned}$$

and

$$\begin{aligned}(-3)^4 &= -3(-3)(-3)(-3)\\ &= 9(-3)(-3)\\ &= 9(9)\\ &= 81\end{aligned}$$

Note: The parentheses are used to ensure that the exponent applies to the entire negative number, including the sign. Without the parentheses there is some ambiguity as to how the exponent should be applied. For example, the expression -3^2 could have two interpretations:

$$-3^2 = (-3)(-3) = 9 \quad \text{or} \quad -3^2 = -(3)(3) = -9$$

4. **Exponents and Parentheses.** If an exponent is present outside of parentheses, then the computations within the parentheses are done first, and the exponential computation is done last:

$$(3 + 5)^2 = 8^2 = 64$$

Notice that the meaning of the expression is changed when each term in the parentheses is raised to the exponent individually:

$$3^2 + 5^2 = 9 + 25 = 34$$

Therefore,

$$X^2 + Y^2 \neq (X + Y)^2$$

5. **Fractions Raised to a Power.** If the numerator and denominator of a fraction are each raised to the same exponent, then the entire fraction can be raised to that exponent. That is,

$$\frac{a^2}{b^2} = \left(\frac{a}{b}\right)^2$$

For example,

$$\frac{3^2}{4^2} = \left(\frac{3}{4}\right)^2$$
$$\frac{9}{16} = \frac{3}{4}\left(\frac{3}{4}\right)$$
$$\frac{9}{16} = \frac{9}{16}$$

SQUARE ROOTS

The square root of a value equals a number that when multiplied by itself yields the original value. For example, the square root of 16 equals 4 because 4 times 4 equals 16. The symbol for the square root is called a *radical,* $\sqrt{\ }$. The square root is taken for the number under the radical. For example,

$$\sqrt{16} = 4$$

Finding the square root is the inverse of raising a number to the second power (squaring). Thus,

$$\sqrt{a^2} = a$$

For example,

$$\sqrt{3^2} = \sqrt{9} = 3$$

Also,

$$(\sqrt{b})^2 = b$$

For example,

$$(\sqrt{64})^2 = 8^2 = 64$$

Computations under the same radical are performed *before* the square root is taken. For example,

$$\sqrt{9 + 16} = \sqrt{25} = 5$$

Note that with addition (or subtraction), separate radicals yield a different result:

$$\sqrt{9} + \sqrt{16} = 3 + 4 = 7$$

Therefore,

$$\sqrt{X} + \sqrt{Y} \neq \sqrt{X + Y}$$
$$\sqrt{X} - \sqrt{Y} \neq \sqrt{X - Y}$$

If the numerator and denominator of a fraction each have a radical, then the entire fraction can be placed under a single radical:

$$\frac{\sqrt{16}}{\sqrt{4}} = \sqrt{\frac{16}{4}}$$
$$\frac{4}{2} = \sqrt{4}$$
$$2 = 2$$

Therefore,

$$\frac{\sqrt{X}}{\sqrt{Y}} = \sqrt{\frac{X}{Y}}$$

Also, if the square root of one number is multiplied by the square root of another number, then the same result would be obtained by taking the square root of the product of both numbers. For example,

$$\sqrt{9} \times \sqrt{16} = \sqrt{9 \times 16}$$
$$3 \times 4 = \sqrt{144}$$
$$12 = 12$$

Therefore,

$$\sqrt{a} \times \sqrt{b} = \sqrt{ab}$$

LEARNING CHECK

1. Perform the following computations:

a. $(-6)^3$
b. $(3 + 7)^2$
c. a^3b^2 when $a = 2$ and $b = -5$
d. a^4b^3 when $a = 2$ and $b = 3$
e. $(XY)^2$ when $X = 3$ and $Y = 5$
f. $X^2 + Y^2$ when $X = 3$ and $Y = 5$
g. $(X + Y)^2$ when $X = 3$ and $Y = 5$
h. $\sqrt{5 + 4}$
i. $(\sqrt{9})^2$
j. $\dfrac{\sqrt{16}}{\sqrt{4}}$

ANSWERS **1. a.** -216 **b.** 100 **c.** 200 **d.** 432 **e.** 225 **f.** 34 **g.** 64 **h.** 3 **i.** 9 **j.** 2

PROBLEMS FOR APPENDIX A Basic Mathematics Review

1. $50/(10 - 8) = ?$

2. $(2 + 3)^2 = ?$

3. $20/10 \times 3 = ?$

4. $12 - 4 \times 2 + 6/3 = ?$

5. $24/(12 - 4) + 2 \times (6 + 3) = ?$

6. Convert $\frac{7}{20}$ to a decimal.

7. Express $\frac{9}{25}$ as a percentage.

8. Convert 0.91 to a fraction.

9. Express 0.0031 as a fraction.

10. Next to each set of fractions, write "True" if they are equivalent and "False" if they are not:

a. $\frac{4}{1000} = \frac{2}{100}$ _____

b. $\frac{5}{6} = \frac{52}{62}$ _____

c. $\frac{1}{8} = \frac{7}{56}$ _____

11. Perform the following calculations:

a. $\frac{4}{5} \times \frac{2}{3} = ?$ **b.** $\frac{7}{9} \div \frac{2}{3} = ?$

c. $\frac{3}{8} + \frac{1}{5} = ?$ **d.** $\frac{5}{18} - \frac{1}{6} = ?$

12. $2.51 \times 0.017 = ?$

13. $3.88 \times 0.0002 = ?$

14. $3.17 + 17.0132 = ?$

15. $5.55 + 10.7 + 0.711 + 3.33 + 0.031 = ?$

16. $2.04 \div 0.2 = ?$

17. $0.36 \div 0.4 = ?$

18. $5 + 3 - 6 - 4 + 3 = ?$

19. $9 - (-1) - 17 + 3 - (-4) + 5 = ?$

20. $5 + 3 - (-8) - (-1) + (-3) - 4 + 10 = ?$

21. $8 \times (-3) = ?$

22. $-22 \div (-2) = ?$

23. $-2(-4)\_(-3) = ?$

24. $84 \div (-4) = ?$

Solve the equations in problems 25–32 for X.

25. $X - 7 = -2$

26. $9 = X + 3$

27. $\frac{X}{4} = 11$

28. $-3 = \frac{X}{3}$

29. $\frac{X + 3}{5} = 2$

30. $\frac{X + 1}{3} = -8$

31. $6X - 1 = 11$

32. $2X + 3 = -11$

33. $(-5)^2 = ?$

34. $(-5)^3 = ?$

35. If $a = 4$ and $b = 3$, then $a^2 + b^4 = ?$

36. If $a = -1$ and $b = 4$, then $(a + b)^2 = ?$

37. If $a = -1$ and $b = 5$, then $ab^2 = ?$

38. $\frac{18}{\sqrt{4}} = ?$

39. $\sqrt{\frac{20}{5}} = ?$

SKILLS ASSESSMENT FINAL EXAM

SECTION 1

1. $4 + 8/4 = ?$

2. $(4 + 8)/4 = ?$

3. $4 \times 3^2 = ?$

4. $(4 \times 3)^2 = ?$

5. $10/5 \times 2 = ?$

6. $10/(5 \times 2) = ?$

7. $40 - 10 \times 4/2 = ?$

8. $(5 - 1)^2/2 = ?$

9. $3 \times 6 - 3^2 = ?$

10. $2 \times (6 - 3)^2 = ?$

11. $4 \times 3 - 1 + 8 \times 2 = ?$

12. $4 \times (3 - 1 + 8) \times 2 = ?$

SECTION 2

1. Express $\frac{14}{80}$ as a decimal.

2. Convert $\frac{6}{25}$ to a percentage.

3. Convert 18% to a fraction.

4. $\frac{3}{5} \times \frac{2}{3} = ?$

5. $\frac{5}{24} + \frac{5}{6} = ?$

6. $\frac{7}{12} \div \frac{5}{6} = ?$

7. $\frac{5}{9} - \frac{1}{3} = ?$

8. $6.11 \times 0.22 = ?$

9. $0.18 \div 0.9 = ?$

10. $8.742 + 0.76 = ?$

11. In a statistics class of 72 students, three-eighths of the students received a B on the first test. How many Bs were earned?

12. What is 15% of 64?

SECTION 3

1. $3 - 1 - 3 + 5 - 2 + 6 = ?$

2. $-8 - (-6) = ?$

3. $2 - (-7) - 3 + (-11) - 20 = ?$

4. $-8 - 3 - (-1) - 2 - 1 = ?$

5. $8(-2) = ?$

6. $-7(-7) = ?$

7. $-3(-2)(-5) = ?$

8. $-3(5)(-3) = ?$

9. $-24 \div (-4) = ?$

10. $36 \div (-6) = ?$

11. $-56/7 = ?$

12. $-7/(-1) = ?$

SECTION 4

Solve for X.

1. $X + 5 = 12$

2. $X - 11 = 3$

3. $10 = X + 4$

4. $4X = 20$

5. $\frac{X}{2} = 15$

6. $18 = 9X$

7. $\frac{X}{5} = 35$

8. $2X + 8 = 4$

9. $\frac{X + 1}{3} = 6$

10. $4X + 3 = -13$

11. $\frac{X + 3}{3} = -7$

12. $23 = 2X - 5$

SECTION 5

1. $5^3 = ?$

2. $(-4)^3 = ?$

3. $(-2)^5 = ?$

4. $(-2)^6 = ?$

5. If $a = 4$ and $b = 2$, then $ab^2 = ?$

6. If $a = 4$ and $b = 2$, then $(a + b)^3 = ?$

7. If $a = 4$ and $b = 2$, then $a^2 + b^2 = ?$

8. $(11 + 4)^2 = ?$

9. $\sqrt{7^2} = ?$

10. If $a = 36$ and $b = 64$, then $\sqrt{a + b} = ?$

11. $\frac{25}{\sqrt{25}} = ? = ?$

12. If $a = -1$ and $b = 2$, then $a^3b^4 = ?$

ANSWER KEY Skills Assessment Exams

PREVIEW EXAM

SECTION 1

1. 17 **2.** 35 **3.** 6
4. 24 **5.** 5 **6.** 2
7. $\frac{1}{3}$ **8.** 8 **9.** 72
10. 8 **11.** 24 **12.** 48

SECTION 2

1. 75% **2.** $\frac{30}{100}$, or $\frac{3}{10}$ **3.** 0.3
4. $\frac{10}{13}$ **5.** 1.625 **6.** $\frac{2}{20}$, or $\frac{1}{10}$
7. $\frac{19}{24}$ **8.** 1.4 **9.** $\frac{4}{15}$
10. 7.5 **11.** 16 **12.** 36

SECTION 3

1. 4 **2.** 8 **3.** 2
4. 9 **5.** -12 **6.** 12
7. -15 **8.** -24 **9.** -4
10. 3 **11.** -2 **12.** 25

FINAL EXAM

SECTION 1

1. 6 **2.** 3 **3.** 36
4. 144 **5.** 4 **6.** 1
7. 20 **8.** 8 **9.** 9
10. 18 **11.** 27 **12.** 80

SECTION 2

1. 0.175 **2.** 24% **3.** $\frac{18}{100}$, or $\frac{9}{50}$
4. $\frac{6}{15}$, or $\frac{2}{5}$ **5.** $\frac{25}{24}$ **6.** $\frac{42}{60}$, or $\frac{7}{10}$
7. $\frac{2}{9}$ **8.** 1.3442 **9.** 0.2
10. 9.502 **11.** 27 **12.** 9.6

SECTION 3

1. 8 **2.** -2 **3.** -25
4. -13 **5.** -16 **6.** 49
7. -30 **8.** 45 **9.** 6
10. -6 **11.** -8 **12.** 7

PREVIEW EXAM

SECTION 4

1. $X = 7$	**2.** $X = 29$	**3.** $X = 9$
4. $X = 4$	**5.** $X = 24$	**6.** $X = 15$
7. $X = 80$	**8.** $X = -3$	**9.** $X = 11$
10. $X = 25$	**11.** $X = 11$	**12.** $X = 7$

SECTION 5

1. 64	**2.** 4	**3.** 54
4. 25	**5.** 13	**6.** -27
7. 256	**8.** 8	**9.** 12
10. 121	**11.** 33	**12.** -9

FINAL EXAM

SECTION 4

1. $X = 7$	**2.** $X = 14$	**3.** $X = 6$
4. $X = 5$	**5.** $X = 30$	**6.** $X = 2$
7. $X = 175$	**8.** $X = -2$	**9.** $X = 17$
10. $X = -4$	**11.** $X = -24$	**12.** $X = 14$

SECTION 5

1. 125	**2.** -64	**3.** -32
4. 64	**5.** 16	**6.** 216
7. 20	**8.** 225	**9.** 7
10. 10	**11.** 5	**12.** -16

SOLUTIONS TO SELECTED PROBLEMS FOR APPENDIX A Basic Mathematics Review

1. 25

3. 6

5. 21

6. 0.35

7. 36%

9. $\frac{31}{10{,}000}$

10. b. False

11. **a.** $\frac{8}{15}$ **b.** $\frac{21}{18}$ **c.** $\frac{23}{40}$

12. 0.04267

14. 20.1832

17. 0.9

19. 5

21. -24

22. 11

25. $X = 5$

28. $X = -9$

30. $X = -25$

31. $X = 2$

34. -125

36. 9

37. -25

39. 2

SUGGESTED REVIEW BOOKS

There are many basic mathematics books available if you need a more extensive review than this appendix can provide. Several are probably available in your library. The following books are but a few of the many that you may find helpful:

Gustafson, R. D., Karr, R., & Massey, M. (2011). *Beginning Algebra* (9th ed.). Belmont, CA: Brooks/Cole.

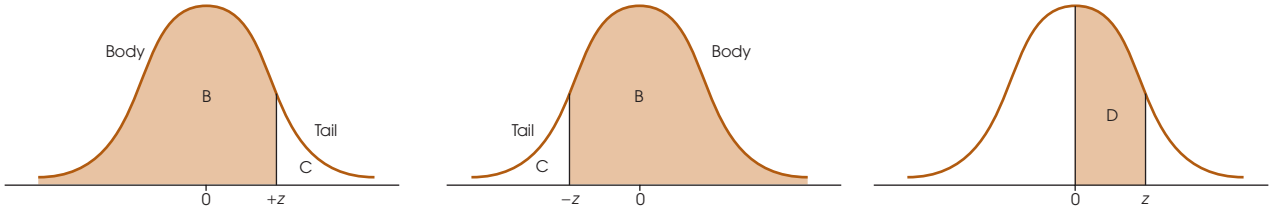
Lial, M. L., Salzman, S.A., & Hestwood, D.L. (2006). *Basic College Mathematics* (7th ed). Reading MA: Addison-Wesley.

McKeague, C. P. (2010). *Basic College Mathematics: A Text/Workbook.* (7th ed.). Belmont, CA: Brooks/Cole.

APPENDIX B Statistical Tables

TABLE B.1 THE UNIT NORMAL TABLE*

*Column A lists *z*-score values. A vertical line drawn through a normal distribution at a *z*-score location divides the distribution into two sections.
Column B identifies the proportion in the larger section, called the *body*.
Column C identifies the proportion in the smaller section, called the *tail*.
Column D identifies the proportion between the mean and the *z*-score.
Note: Because the normal distribution is symmetrical, the proportions for negative *z*-scores are the same as those for positive *z*-scores.



(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*	(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*
0.00	.5000	.5000	.0000	0.25	.5987	.4013	.0987
0.01	.5040	.4960	.0040	0.26	.6026	.3974	.1026
0.02	.5080	.4920	.0080	0.27	.6064	.3936	.1064
0.03	.5120	.4880	.0120	0.28	.6103	.3897	.1103
0.04	.5160	.4840	.0160	0.29	.6141	.3859	.1141
0.05	.5199	.4801	.0199	0.30	.6179	.3821	.1179
0.06	.5239	.4761	.0239	0.31	.6217	.3783	.1217
0.07	.5279	.4721	.0279	0.32	.6255	.3745	.1255
0.08	.5319	.4681	.0319	0.33	.6293	.3707	.1293
0.09	.5359	.4641	.0359	0.34	.6331	.3669	.1331
0.10	.5398	.4602	.0398	0.35	.6368	.3632	.1368
0.11	.5438	.4562	.0438	0.36	.6406	.3594	.1406
0.12	.5478	.4522	.0478	0.37	.6443	.3557	.1443
0.13	.5517	.4483	.0517	0.38	.6480	.3520	.1480
0.14	.5557	.4443	.0557	0.39	.6517	.3483	.1517
0.15	.5596	.4404	.0596	0.40	.6554	.3446	.1554
0.16	.5636	.4364	.0636	0.41	.6591	.3409	.1591
0.17	.5675	.4325	.0675	0.42	.6628	.3372	.1628
0.18	.5714	.4286	.0714	0.43	.6664	.3336	.1664
0.19	.5753	.4247	.0753	0.44	.6700	.3300	.1700
0.20	.5793	.4207	.0793	0.45	.6736	.3264	.1736
0.21	.5832	.4168	.0832	0.46	.6772	.3228	.1772
0.22	.5871	.4129	.0871	0.47	.6808	.3192	.1808
0.23	.5910	.4090	.0910	0.48	.6844	.3156	.1844
0.24	.5948	.4052	.0948	0.49	.6879	.3121	.1879

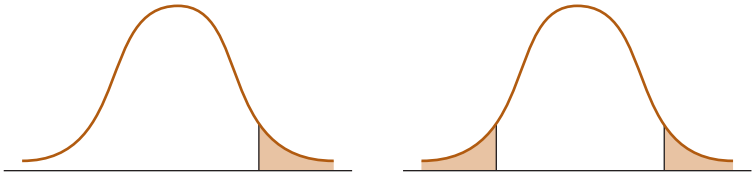
(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z	(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
0.50	.6915	.3085	.1915	1.00	.8413	.1587	.3413
0.51	.6950	.3050	.1950	1.01	.8438	.1562	.3438
0.52	.6985	.3015	.1985	1.02	.8461	.1539	.3461
0.53	.7019	.2981	.2019	1.03	.8485	.1515	.3485
0.54	.7054	.2946	.2054	1.04	.8508	.1492	.3508
0.55	.7088	.2912	.2088	1.05	.8531	.1469	.3531
0.56	.7123	.2877	.2123	1.06	.8554	.1446	.3554
0.57	.7157	.2843	.2157	1.07	.8577	.1423	.3577
0.58	.7190	.2810	.2190	1.08	.8599	.1401	.3599
0.59	.7224	.2776	.2224	1.09	.8621	.1379	.3621
0.60	.7257	.2743	.2257	1.10	.8643	.1357	.3643
0.61	.7291	.2709	.2291	1.11	.8665	.1335	.3665
0.62	.7324	.2676	.2324	1.12	.8686	.1314	.3686
0.63	.7357	.2643	.2357	1.13	.8708	.1292	.3708
0.64	.7389	.2611	.2389	1.14	.8729	.1271	.3729
0.65	.7422	.2578	.2422	1.15	.8749	.1251	.3749
0.66	.7454	.2546	.2454	1.16	.8770	.1230	.3770
0.67	.7486	.2514	.2486	1.17	.8790	.1210	.3790
0.68	.7517	.2483	.2517	1.18	.8810	.1190	.3810
0.69	.7549	.2451	.2549	1.19	.8830	.1170	.3830
0.70	.7580	.2420	.2580	1.20	.8849	.1151	.3849
0.71	.7611	.2389	.2611	1.21	.8869	.1131	.3869
0.72	.7642	.2358	.2642	1.22	.8888	.1112	.3888
0.73	.7673	.2327	.2673	1.23	.8907	.1093	.3907
0.74	.7704	.2296	.2704	1.24	.8925	.1075	.3925
0.75	.7734	.2266	.2734	1.25	.8944	.1056	.3944
0.76	.7764	.2236	.2764	1.26	.8962	.1038	.3962
0.77	.7794	.2206	.2794	1.27	.8980	.1020	.3980
0.78	.7823	.2177	.2823	1.28	.8997	.1003	.3997
0.79	.7852	.2148	.2852	1.29	.9015	.0985	.4015
0.80	.7881	.2119	.2881	1.30	.9032	.0968	.4032
0.81	.7910	.2090	.2910	1.31	.9049	.0951	.4049
0.82	.7939	.2061	.2939	1.32	.9066	.0934	.4066
0.83	.7967	.2033	.2967	1.33	.9082	.0918	.4082
0.84	.7995	.2005	.2995	1.34	.9099	.0901	.4099
0.85	.8023	.1977	.3023	1.35	.9115	.0885	.4115
0.86	.8051	.1949	.3051	1.36	.9131	.0869	.4131
0.87	.8078	.1922	.3078	1.37	.9147	.0853	.4147
0.88	.8106	.1894	.3106	1.38	.9162	.0838	.4162
0.89	.8133	.1867	.3133	1.39	.9177	.0823	.4177
0.90	.8159	.1841	.3159	1.40	.9192	.0808	.4192
0.91	.8186	.1814	.3186	1.41	.9207	.0793	.4207
0.92	.8212	.1788	.3212	1.42	.9222	.0778	.4222
0.93	.8238	.1762	.3238	1.43	.9236	.0764	.4236
0.94	.8264	.1736	.3264	1.44	.9251	.0749	.4251
0.95	.8289	.1711	.3289	1.45	.9265	.0735	.4265
0.96	.8315	.1685	.3315	1.46	.9279	.0721	.4279
0.97	.8340	.1660	.3340	1.47	.9292	.0708	.4292
0.98	.8365	.1635	.3365	1.48	.9306	.0694	.4306
0.99	.8389	.1611	.3389	1.49	.9319	.0681	.4319

(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*	(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*
1.50	.9332	.0668	.4332	2.00	.9772	.0228	.4772
1.51	.9345	.0655	.4345	2.01	.9778	.0222	.4778
1.52	.9357	.0643	.4357	2.02	.9783	.0217	.4783
1.53	.9370	.0630	.4370	2.03	.9788	.0212	.4788
1.54	.9382	.0618	.4382	2.04	.9793	.0207	.4793
1.55	.9394	.0606	.4394	2.05	.9798	.0202	.4798
1.56	.9406	.0594	.4406	2.06	.9803	.0197	.4803
1.57	.9418	.0582	.4418	2.07	.9808	.0192	.4808
1.58	.9429	.0571	.4429	2.08	.9812	.0188	.4812
1.59	.9441	.0559	.4441	2.09	.9817	.0183	.4817
1.60	.9452	.0548	.4452	2.10	.9821	.0179	.4821
1.61	.9463	.0537	.4463	2.11	.9826	.0174	.4826
1.62	.9474	.0526	.4474	2.12	.9830	.0170	.4830
1.63	.9484	.0516	.4484	2.13	.9834	.0166	.4834
1.64	.9495	.0505	.4495	2.14	.9838	.0162	.4838
1.65	.9505	.0495	.4505	2.15	.9842	.0158	.4842
1.66	.9515	.0485	.4515	2.16	.9846	.0154	.4846
1.67	.9525	.0475	.4525	2.17	.9850	.0150	.4850
1.68	.9535	.0465	.4535	2.18	.9854	.0146	.4854
1.69	.9545	.0455	.4545	2.19	.9857	.0143	.4857
1.70	.9554	.0446	.4554	2.20	.9861	.0139	.4861
1.71	.9564	.0436	.4564	2.21	.9864	.0136	.4864
1.72	.9573	.0427	.4573	2.22	.9868	.0132	.4868
1.73	.9582	.0418	.4582	2.23	.9871	.0129	.4871
1.74	.9591	.0409	.4591	2.24	.9875	.0125	.4875
1.75	.9599	.0401	.4599	2.25	.9878	.0122	.4878
1.76	.9608	.0392	.4608	2.26	.9881	.0119	.4881
1.77	.9616	.0384	.4616	2.27	.9884	.0116	.4884
1.78	.9625	.0375	.4625	2.28	.9887	.0113	.4887
1.79	.9633	.0367	.4633	2.29	.9890	.0110	.4890
1.80	.9641	.0359	.4641	2.30	.9893	.0107	.4893
1.81	.9649	.0351	.4649	2.31	.9896	.0104	.4896
1.82	.9656	.0344	.4656	2.32	.9898	.0102	.4898
1.83	.9664	.0336	.4664	2.33	.9901	.0099	.4901
1.84	.9671	.0329	.4671	2.34	.9904	.0096	.4904
1.85	.9678	.0322	.4678	2.35	.9906	.0094	.4906
1.86	.9686	.0314	.4686	2.36	.9909	.0091	.4909
1.87	.9693	.0307	.4693	2.37	.9911	.0089	.4911
1.88	.9699	.0301	.4699	2.38	.9913	.0087	.4913
1.89	.9706	.0294	.4706	2.39	.9916	.0084	.4916
1.90	.9713	.0287	.4713	2.40	.9918	.0082	.4918
1.91	.9719	.0281	.4719	2.41	.9920	.0080	.4920
1.92	.9726	.0274	.4726	2.42	.9922	.0078	.4922
1.93	.9732	.0268	.4732	2.43	.9925	.0075	.4925
1.94	.9738	.0262	.4738	2.44	.9927	.0073	.4927
1.95	.9744	.0256	.4744	2.45	.9929	.0071	.4929
1.96	.9750	.0250	.4750	2.46	.9931	.0069	.4931
1.97	.9756	.0244	.4756	2.47	.9932	.0068	.4932
1.98	.9761	.0239	.4761	2.48	.9934	.0066	.4934
1.99	.9767	.0233	.4767	2.49	.9936	.0064	.4936

(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*	(A) *z*	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and *z*
2.50	.9938	.0062	.4938	2.95	.9984	.0016	.4984
2.51	.9940	.0060	.4940	2.96	.9985	.0015	.4985
2.52	.9941	.0059	.4941	2.97	.9985	.0015	.4985
2.53	.9943	.0057	.4943	2.98	.9986	.0014	.4986
2.54	.9945	.0055	.4945	2.99	.9986	.0014	.4986
2.55	.9946	.0054	.4946	3.00	.9987	.0013	.4987
2.56	.9948	.0052	.4948	3.01	.9987	.0013	.4987
2.57	.9949	.0051	.4949	3.02	.9987	.0013	.4987
2.58	.9951	.0049	.4951	3.03	.9988	.0012	.4988
2.59	.9952	.0048	.4952	3.04	.9988	.0012	.4988
2.60	.9953	.0047	.4953	3.05	.9989	.0011	.4989
2.61	.9955	.0045	.4955	3.06	.9989	.0011	.4989
2.62	.9956	.0044	.4956	3.07	.9989	.0011	.4989
2.63	.9957	.0043	.4957	3.08	.9990	.0010	.4990
2.64	.9959	.0041	.4959	3.09	.9990	.0010	.4990
2.65	.9960	.0040	.4960	3.10	.9990	.0010	.4990
2.66	.9961	.0039	.4961	3.11	.9991	.0009	.4991
2.67	.9962	.0038	.4962	3.12	.9991	.0009	.4991
2.68	.9963	.0037	.4963	3.13	.9991	.0009	.4991
2.69	.9964	.0036	.4964	3.14	.9992	.0008	.4992
2.70	.9965	.0035	.4965	3.15	.9992	.0008	.4992
2.71	.9966	.0034	.4966	3.16	.9992	.0008	.4992
2.72	.9967	.0033	.4967	3.17	.9992	.0008	.4992
2.73	.9968	.0032	.4968	3.18	.9993	.0007	.4993
2.74	.9969	.0031	.4969	3.19	.9993	.0007	.4993
2.75	.9970	.0030	.4970	3.20	.9993	.0007	.4993
2.76	.9971	.0029	.4971	3.21	.9993	.0007	.4993
2.77	.9972	.0028	.4972	3.22	.9994	.0006	.4994
2.78	.9973	.0027	.4973	3.23	.9994	.0006	.4994
2.79	.9974	.0026	.4974	3.24	.9994	.0006	.4994
2.80	.9974	.0026	.4974	3.30	.9995	.0005	.4995
2.81	.9975	.0025	.4975	3.40	.9997	.0003	.4997
2.82	.9976	.0024	.4976	3.50	.9998	.0002	.4998
2.83	.9977	.0023	.4977	3.60	.9998	.0002	.4998
2.84	.9977	.0023	.4977	3.70	.9999	.0001	.4999
2.85	.9978	.0022	.4978	3.80	.99993	.00007	.49993
2.86	.9979	.0021	.4979	3.90	.99995	.00005	.49995
2.87	.9979	.0021	.4979	4.00	.99997	.00003	.49997
2.88	.9980	.0020	.4980				
2.89	.9981	.0019	.4981				
2.90	.9981	.0019	.4981				
2.91	.9982	.0018	.4982				
2.92	.9982	.0018	.4982				
2.93	.9983	.0017	.4983				
2.94	.9984	.0016	.4984				

TABLE B.2 THE *t* DISTRIBUTION

Table entries are values of *t* corresponding to proportions in one tail or in two tails combined.



	Proportion in One Tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in Two Tails Combined					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

TABLE B.3 CRITICAL VALUES FOR THE *F*-MAX STATISTIC*

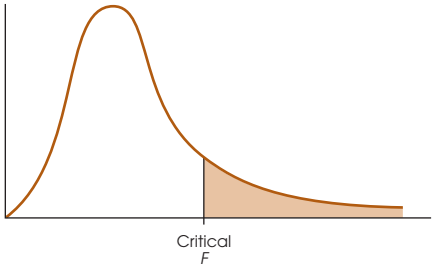
*The critical values for α = .05 are in lightface type, and for α = .01, they are in boldface type.

$n-1$	k = Number of Samples										
	2	3	4	5	6	7	8	9	10	11	12
4	9.60 **23.2**	15.5 **37.**	20.6 **49.**	25.2 **59.**	29.5 **69.**	33.6 **79.**	37.5 **89.**	41.4 **97.**	44.6 **106.**	48.0 **113.**	51.4 **120.**
5	7.15 **14.9**	10.8 **22.**	13.7 **28.**	16.3 **33.**	18.7 **38.**	20.8 **42.**	22.9 **46.**	24.7 **50.**	26.5 **54.**	28.2 **57.**	29.9 **60.**
6	5.82 **11.1**	8.38 **15.5**	10.4 **19.1**	12.1 **22.**	13.7 **25.**	15.0 **27.**	16.3 **30.**	17.5 **32.**	18.6 **34.**	19.7 **36.**	20.7 **37.**
7	4.99 **8.89**	6.94 **12.1**	8.44 **14.5**	9.70 **16.5**	10.8 **18.4**	11.8 **20.**	12.7 **22.**	13.5 **23.**	14.3 **24.**	15.1 **26.**	15.8 **27.**
8	4.43 **7.50**	6.00 **9.9**	7.18 **11.7**	8.12 **13.2**	9.03 **14.5**	9.78 **15.8**	10.5 **16.9**	11.1 **17.9**	11.7 **18.9**	12.2 **19.8**	12.7 **21.**
9	4.03 **6.54**	5.34 **8.5**	6.31 **9.9**	7.11 **11.1**	7.80 **12.1**	8.41 **13.1**	8.95 **13.9**	9.45 **14.7**	9.91 **15.3**	10.3 **16.0**	10.7 **16.6**
10	3.72 **5.85**	4.85 **7.4**	5.67 **8.6**	6.34 **9.6**	6.92 **10.4**	7.42 **11.1**	7.87 **11.8**	8.28 **12.4**	8.66 **12.9**	9.01 **13.4**	9.34 **13.9**
12	3.28 **4.91**	4.16 **6.1**	4.79 **6.9**	5.30 **7.6**	5.72 **8.2**	6.09 **8.7**	6.42 **9.1**	6.72 **9.5**	7.00 **9.9**	7.25 **10.2**	7.48 **10.6**
15	2.86 **4.07**	3.54 **4.9**	4.01 **5.5**	4.37 **6.0**	4.68 **6.4**	4.95 **6.7**	5.19 **7.1**	5.40 **7.3**	5.59 **7.5**	5.77 **7.8**	5.93 **8.0**
20	2.46 **3.32**	2.95 **3.8**	3.29 **4.3**	3.54 **4.6**	3.76 **4.9**	3.94 **5.1**	4.10 **5.3**	4.24 **5.5**	4.37 **5.6**	4.49 **5.8**	4.59 **5.9**
30	2.07 **2.63**	2.40 **3.0**	2.61 **3.3**	2.78 **3.5**	2.91 **3.6**	3.02 **3.7**	3.12 **3.8**	3.21 **3.9**	3.29 **4.0**	3.36 **4.1**	3.39 **4.2**
60	1.67 **1.96**	1.85 **2.2**	1.96 **2.3**	2.04 **2.4**	2.11 **2.4**	2.17 **2.5**	2.22 **2.5**	2.26 **2.6**	2.30 **2.6**	2.33 **2.7**	2.36 **2.7**

Table 31 of E. Pearson and H.O. Hartley, Biometrika Tables for Statisticians, 2nd ed. New York: Cambridge University Press, 1958. Adapted and reprinted with permission of the Biometrika trustees.

TABLE B.4 THE *F* DISTRIBUTION*

*Table entries in lightface type are critical values for the .05 level of significance.
Boldface type values are for the .01 level of significance.



Degrees of Freedom: Denominator	Degrees of Freedom: Numerator														
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
1	161 **4052**	200 **4999**	216 **5403**	225 **5625**	230 **5764**	234 **5859**	237 **5928**	239 **5981**	241 **6022**	242 **6056**	243 **6082**	244 **6106**	245 **6142**	246 **6169**	248 **6208**
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**
3	10.13 **34.12**	9.55 **30.92**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**	2.48 **3.70**	2.44 **3.62**	2.39 **3.51**
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.70 **4.14**	2.64 **4.00**	2.59 **3.89**	2.55 **3.80**	2.51 **3.73**	2.48 **3.67**	2.43 **3.56**	2.39 **3.48**	2.33 **3.36**
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.66 **4.03**	2.59 **3.89**	2.54 **3.78**	2.49 **3.69**	2.45 **3.61**	2.42 **3.55**	2.37 **3.45**	2.33 **3.37**	2.28 **3.25**

TABLE B.4 (continued)

Degrees of Freedom: Denominator	Degrees of Freedom: Numerator														
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**

TABLE B.4 (continued)

Degrees of Freedom: Denominator	Degrees of Freedom: Numerator														
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**

TABLE B.5 THE STUDENTIZED RANGE STATISTIC (q)*

*The critical values for q corresponding to $\alpha = .05$ (lightface type) and $\alpha = .01$ (boldface type).

	k = Number of Treatments										
df for Error Term	2	3	4	5	6	7	8	9	10	11	12
5	3.64 **5.70**	4.60 **6.98**	5.22 **7.80**	5.67 **8.42**	6.03 **8.91**	6.33 **9.32**	6.58 **9.67**	6.80 **9.97**	6.99 **10.24**	7.17 **10.48**	7.32 **10.70**
6	3.46 **5.24**	4.34 **6.33**	4.90 **7.03**	5.30 **7.56**	5.63 **7.97**	5.90 **8.32**	6.12 **8.61**	6.32 **8.87**	6.49 **9.10**	6.65 **9.30**	6.79 **9.48**
7	3.34 **4.95**	4.16 **5.92**	4.68 **6.54**	5.06 **7.01**	5.36 **7.37**	5.61 **7.68**	5.82 **7.94**	6.00 **8.17**	6.16 **8.37**	6.30 **8.55**	6.43 **8.71**
8	3.26 **4.75**	4.04 **5.64**	4.53 **6.20**	4.89 **6.62**	5.17 **6.96**	5.40 **7.24**	5.60 **7.47**	5.77 **7.68**	5.92 **7.86**	6.05 **8.03**	6.18 **8.18**
9	3.20 **4.60**	3.95 **5.43**	4.41 **5.96**	4.76 **6.35**	5.02 **6.66**	5.24 **6.91**	5.43 **7.13**	5.59 **7.33**	5.74 **7.49**	5.87 **7.65**	5.98 **7.78**
10	3.15 **4.48**	3.88 **5.27**	4.33 **5.77**	4.65 **6.14**	4.91 **6.43**	5.12 **6.67**	5.30 **6.87**	5.46 **7.05**	5.60 **7.21**	5.72 **7.36**	5.83 **7.49**
11	3.11 **4.39**	3.82 **5.15**	4.26 **5.62**	4.57 **5.97**	4.82 **6.25**	5.03 **6.48**	5.20 **6.67**	5.35 **6.84**	5.49 **6.99**	5.61 **7.13**	5.71 **7.25**
12	3.08 **4.32**	3.77 **5.05**	4.20 **5.50**	4.51 **5.84**	4.75 **6.10**	4.95 **6.32**	5.12 **6.51**	5.27 **6.67**	5.39 **6.81**	5.51 **6.94**	5.61 **7.06**
13	3.06 **4.26**	3.73 **4.96**	4.15 **5.40**	4.45 **5.73**	4.69 **5.98**	4.88 **6.19**	5.05 **6.37**	5.19 **6.53**	5.32 **6.67**	5.43 **6.79**	5.53 **6.90**
14	3.03 **4.21**	3.70 **4.89**	4.11 **5.32**	4.41 **5.63**	4.64 **5.88**	4.83 **6.08**	4.99 **6.26**	5.13 **6.41**	5.25 **6.54**	5.36 **6.66**	5.46 **6.77**
15	3.01 **4.17**	3.67 **4.84**	4.08 **5.25**	4.37 **5.56**	4.59 **5.80**	4.78 **5.99**	4.94 **6.16**	5.08 **6.31**	5.20 **6.44**	5.31 **6.55**	5.40 **6.66**
16	3.00 **4.13**	3.65 **4.79**	4.05 **5.19**	4.33 **5.49**	4.56 **5.72**	4.74 **5.92**	4.90 **6.08**	5.03 **6.22**	5.15 **6.35**	5.26 **6.46**	5.35 **6.56**
17	2.98 **4.10**	3.63 **4.74**	4.02 **5.14**	4.30 **5.43**	4.52 **5.66**	4.70 **5.85**	4.86 **6.01**	4.99 **6.15**	5.11 **6.27**	5.21 **6.38**	5.31 **6.48**
18	2.97 **4.07**	3.61 **4.70**	4.00 **5.09**	4.28 **5.38**	4.49 **5.60**	4.67 **5.79**	4.82 **5.94**	4.96 **6.08**	5.07 **6.20**	5.17 **6.31**	5.27 **6.41**
19	2.96 **4.05**	3.59 **4.67**	3.98 **5.05**	4.25 **5.33**	4.47 **5.55**	4.65 **5.73**	4.79 **5.89**	4.92 **6.02**	5.04 **6.14**	5.14 **6.25**	5.23 **6.34**
20	2.95 **4.02**	3.58 **4.64**	3.96 **5.02**	4.23 **5.29**	4.45 **5.51**	4.62 **5.69**	4.77 **5.84**	4.90 **5.97**	5.01 **6.09**	5.11 **6.19**	5.20 **6.28**
24	2.92 **3.96**	3.53 **4.55**	3.90 **4.91**	4.17 **5.17**	4.37 **5.37**	4.54 **5.54**	4.68 **5.69**	4.81 **5.81**	4.92 **5.92**	5.01 **6.02**	5.10 **6.11**
30	2.89 **3.89**	3.49 **4.45**	3.85 **4.80**	4.10 **5.05**	4.30 **5.24**	4.46 **5.40**	4.60 **5.54**	4.72 **5.65**	4.82 **5.76**	4.92 **5.85**	5.00 **5.93**
40	2.86 **3.82**	3.44 **4.37**	3.79 **4.70**	4.04 **4.93**	4.23 **5.11**	4.39 **5.26**	4.52 **5.39**	4.63 **5.50**	4.73 **5.60**	4.82 **5.69**	4.90 **5.76**
60	2.83 **3.76**	3.40 **4.28**	3.74 **4.59**	3.98 **4.82**	4.16 **4.99**	4.31 **5.13**	4.44 **5.25**	4.55 **5.36**	4.65 **5.45**	4.73 **5.53**	4.81 **5.60**
120	2.80 **3.70**	3.36 **4.20**	3.68 **4.50**	3.92 **4.71**	4.10 **4.87**	4.24 **5.01**	4.36 **5.12**	4.47 **5.21**	4.56 **5.30**	4.64 **5.37**	4.71 **5.44**
∞	2.77 **3.64**	3.31 **4.12**	3.63 **4.40**	3.86 **4.60**	4.03 **4.76**	4.17 **4.88**	4.28 **4.99**	4.39 **5.08**	4.47 **5.16**	4.55 **5.23**	4.62 **5.29**

Table 29 of E. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, 2nd ed. New York: Cambridge University Press, 1958. Adapted and reprinted with permission of the Biometrika trustees.

TABLE B.6 CRITICAL VALUES FOR THE PEARSON CORRELATION*

*To be significant, the sample correlation, r, must be greater than or equal to the critical value in the table.

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
$df = n - 2$	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.496
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254

TABLE B.7 CRITICAL VALUES FOR THE SPEARMAN CORRELATION*

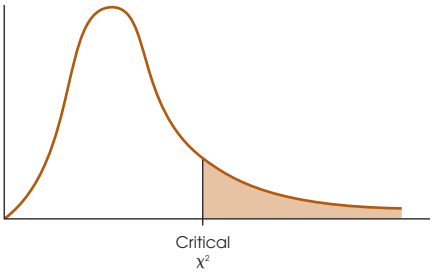
*To be significant, the sample correlation, r_s, must be greater than or equal to the critical value in the table.

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
n	.10	.05	.02	.01
4	1.000			
5	0.900	1.000	1.000	
6	0.829	0.886	0.943	1.000
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.755
12	0.503	0.587	0.671	0.727
13	0.484	0.560	0.648	0.703
14	0.464	0.538	0.622	0.675
15	0.443	0.521	0.604	0.654
16	0.429	0.503	0.582	0.635
17	0.414	0.485	0.566	0.615
18	0.401	0.472	0.550	0.600
19	0.391	0.460	0.535	0.584
20	0.380	0.447	0.520	0.570
21	0.370	0.435	0.508	0.556
22	0.361	0.425	0.496	0.544
23	0.353	0.415	0.486	0.532
24	0.344	0.406	0.476	0.521
25	0.337	0.398	0.466	0.511
26	0.331	0.390	0.457	0.501
27	0.324	0.382	0.448	0.491
28	0.317	0.375	0.440	0.483
29	0.312	0.368	0.433	0.475
30	0.306	0.362	0.425	0.467
35	0.283	0.335	0.394	0.433
40	0.264	0.313	0.368	0.405
45	0.248	0.294	0.347	0.382
50	0.235	0.279	0.329	0.363
60	0.214	0.255	0.300	0.331
70	0.190	0.235	0.278	0.307
80	0.185	0.220	0.260	0.287
90	0.174	0.207	0.245	0.271
100	0.165	0.197	0.233	0.257

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67, 578–580. Reprinted with permission from the *Journal of the American Statistical Association*.

TABLE B.8 THE CHI-SQUARE DISTRIBUTION*

*The table entries are critical values of χ^2.



df	*Proportion in Critical Region* 0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.64
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67
40	51.81	55.76	59.34	63.69	66.77
50	63.17	67.50	71.42	76.15	79.49
60	74.40	79.08	83.30	88.38	91.95
70	85.53	90.53	95.02	100.42	104.22
80	96.58	101.88	106.63	112.33	116.32
90	107.56	113.14	118.14	124.12	128.30
100	118.50	124.34	129.56	135.81	140.17

Table 8 of E. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, 3rd ed. New York: Cambridge University Press, 1966. Adapted and reprinted with permission of the Biometrika trustees.

TABLE B.9A CRITICAL VALUES OF THE MANN-WHITNEY *U* FOR $\alpha = .05$*

*Critical values are provided for a *one-tailed* test at $\alpha = .05$ (lightface type) and for a *two-tailed* test at $\alpha = .05$ (boldface type). To be significant for any given n_A and n_B, the obtained U must be *equal to* or *less than* the critical value in the table. Dashes (—) in the body of the table indicate that no decision is possible at the stated level of significance and values of n_A and n_B.

n_B \ n_A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0	0
	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
	—	—	—	—	—	—	—	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**
3	—	—	0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
	—	—	—	—	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**	**7**	**7**	**8**
4	—	—	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	—	—	—	**0**	**1**	**2**	**3**	**4**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**11**	**12**	**13**	**13**
5	—	0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	—	—	**0**	**1**	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**11**	**12**	**13**	**14**	**15**	**17**	**18**	**19**	**20**
6	—	0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	—	—	**1**	**2**	**3**	**5**	**6**	**8**	**10**	**11**	**13**	**14**	**16**	**17**	**19**	**21**	**22**	**24**	**25**	**27**
7	—	0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	—	—	**1**	**3**	**5**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**
8	—	1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	—	**0**	**2**	**4**	**6**	**8**	**10**	**13**	**15**	**17**	**19**	**22**	**24**	**26**	**29**	**31**	**34**	**36**	**38**	**41**
9	—	1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	—	**0**	**2**	**4**	**7**	**10**	**12**	**15**	**17**	**20**	**23**	**26**	**28**	**31**	**34**	**37**	**39**	**42**	**45**	**48**
10	—	1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	—	**0**	**3**	**5**	**8**	**11**	**14**	**17**	**20**	**23**	**26**	**29**	**33**	**36**	**39**	**42**	**45**	**48**	**52**	**55**
11	—	1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	—	**0**	**3**	**6**	**9**	**13**	**16**	**19**	**23**	**26**	**30**	**33**	**37**	**40**	**44**	**47**	**51**	**55**	**58**	**62**
12	—	2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	—	**1**	**4**	**7**	**11**	**14**	**18**	**22**	**26**	**29**	**33**	**37**	**41**	**45**	**49**	**53**	**57**	**61**	**65**	**69**
13	—	2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	79	75	80	84
	—	**1**	**4**	**8**	**12**	**16**	**20**	**24**	**28**	**33**	**37**	**41**	**45**	**50**	**54**	**59**	**63**	**67**	**72**	**76**
14	—	2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	—	**1**	**5**	**9**	**13**	**17**	**22**	**26**	**31**	**36**	**40**	**45**	**50**	**55**	**59**	**64**	**67**	**74**	**78**	**83**
15	—	3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	—	**1**	**5**	**10**	**14**	**19**	**24**	**29**	**34**	**39**	**44**	**49**	**54**	**59**	**64**	**70**	**75**	**80**	**85**	**90**
16	—	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	—	**1**	**6**	**11**	**15**	**21**	**26**	**31**	**37**	**42**	**47**	**53**	**59**	**64**	**70**	**75**	**81**	**86**	**92**	**98**
17	—	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
	—	**2**	**6**	**11**	**17**	**22**	**28**	**34**	**39**	**45**	**51**	**57**	**63**	**67**	**75**	**81**	**87**	**93**	**99**	**105**
18	—	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	—	**2**	**7**	**12**	**18**	**24**	**30**	**36**	**42**	**48**	**55**	**61**	**67**	**74**	**80**	**86**	**93**	**99**	**106**	**112**
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
	—	**2**	**7**	**13**	**19**	**25**	**32**	**38**	**45**	**52**	**58**	**65**	**72**	**78**	**85**	**92**	**99**	**106**	**113**	**119**
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
	—	**2**	**8**	**13**	**20**	**27**	**34**	**41**	**48**	**55**	**62**	**69**	**76**	**83**	**90**	**98**	**105**	**112**	**119**	**127**

TABLE B.9B CRITICAL VALUES OF THE MANN-WHITNEY *U* FOR α = .01*

*Critical values are provided for a *one-tailed* test at α = .01 (lightface type) and for a *two-tailed* test at α = .01 (boldface type). To be significant for any given n_A and n_B, the obtained U must be *equal to* or *less than* the critical value in the table. Dashes (—) in the body of the table indicate that no decision is possible at the stated level of significance and values of n_A and n_B.

n_B \ n_A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—	0	0	0	0	0	0	1	1
	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	**0**	**0**
3	—	—	—	—	—	—	0	0	1	1	1	2	2	2	3	3	4	4	4	5
	—	—	—	—	—	—	—	—	**0**	**0**	**0**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**3**	**3**
4	—	—	—	—	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
	—	—	—	—	—	**0**	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**5**	**5**	**6**	**6**	**7**	**8**
5	—	—	—	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	—	—	—	—	**0**	**1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
6	—	—	—	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
	—	—	—	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**10**	**11**	**12**	**13**	**15**	**16**	**17**	**18**
7	—	—	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
	—	—	—	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	**24**
8	—	—	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
	—	—	—	**1**	**2**	**4**	**6**	**7**	**9**	**11**	**13**	**15**	**17**	**18**	**20**	**22**	**24**	**26**	**28**	**30**
9	—	—	1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
	—	—	**0**	**1**	**3**	**5**	**7**	**9**	**11**	**13**	**16**	**18**	**20**	**22**	**24**	**27**	**29**	**31**	**33**	**36**
10	—	—	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
	—	—	**0**	**2**	**4**	**6**	**9**	**11**	**13**	**16**	**18**	**21**	**24**	**26**	**29**	**31**	**34**	**37**	**39**	**42**
11	—	—	1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
	—	—	**0**	**2**	**5**	**7**	**10**	**13**	**16**	**18**	**21**	**24**	**27**	**30**	**33**	**36**	**39**	**42**	**45**	**48**
12	—	—	2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
	—	—	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	—	0	2	5	9	12	16	2	23	27	31	35	39	43	47	51	55	59	63	67
	—	—	**1**	**3**	**7**	**10**	**13**	**17**	**20**	**24**	**27**	**31**	**34**	**38**	**42**	**45**	**49**	**53**	**56**	**60**
14	—	0	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
	—	—	**1**	**4**	**7**	**11**	**15**	**18**	**22**	**26**	**30**	**34**	**38**	**42**	**46**	**50**	**54**	**58**	**63**	**67**
15	—	0	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
	—	—	**2**	**5**	**8**	**12**	**16**	**20**	**24**	**29**	**33**	**37**	**42**	**46**	**51**	**55**	**60**	**64**	**69**	**73**
16	—	0	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
	—	—	**2**	**5**	**9**	**13**	**18**	**22**	**27**	**31**	**36**	**41**	**45**	**50**	**55**	**60**	**65**	**70**	**74**	**79**
17	—	0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
	—	—	**2**	**6**	**10**	**15**	**19**	**24**	**29**	**34**	**39**	**44**	**49**	**54**	**60**	**65**	**70**	**75**	**81**	**86**
18	—	0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
	—	—	**2**	**6**	**11**	**16**	**21**	**26**	**31**	**37**	**42**	**47**	**53**	**58**	**64**	**70**	**75**	**81**	**87**	**92**
19	—	1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
	—	**0**	**3**	**7**	**12**	**17**	**22**	**28**	**33**	**39**	**45**	**51**	**56**	**63**	**69**	**74**	**81**	**87**	**93**	**99**
20	—	1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114
	—	**0**	**3**	**8**	**13**	**18**	**24**	**30**	**36**	**42**	**48**	**54**	**60**	**67**	**73**	**79**	**86**	**92**	**99**	**105**

TABLE B.10 CRITICAL VALUES OF T FOR THE WILCOXON SIGNED-RANKS TEST*

*To be significant, the obtained T must be *equal to* or *less than* the critical value. Dashes (—) in the columns indicate that no decision is possible for the stated α and n.

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
n	.10	.05	.02	.01
5	0	—	—	—
6	2	0	—	—
7	3	2	0	—
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68
26	110	98	84	75
27	119	107	92	83
28	130	116	101	91
29	140	126	110	100

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
n	.10	.05	.02	.01
30	151	137	120	109
31	163	147	130	118
32	175	159	140	128
33	187	170	151	138
34	200	182	162	148
35	213	195	173	159
36	227	208	185	171
37	241	221	198	182
38	256	235	211	194
39	271	249	224	207
40	286	264	238	220
41	302	279	252	233
42	319	294	266	247
43	336	310	281	261
44	353	327	296	276
45	371	343	312	291
46	389	361	328	307
47	407	378	345	322
48	426	396	362	339
49	446	415	379	355
50	466	434	397	373

Adapted from F. Wilcoxon, S. K. Katti, and R. A. Wilcox, *Critical Values and Probability Levels of the Wilcoxon Rank-Sum Test and the Wilcoxon Signed-Ranks Test*. Wayne, NJ: American Cyanamid Company, 1963. Adapted and reprinted with permission of the American Cyanamid Company.

APPENDIX C Solutions for Odd-Numbered Problems in the Text

Note: Many of the problems in the text require several stages of computation. At each stage there is an opportunity for rounding answers. Depending on the exact sequence of operations used to solve a problem, different individuals will round their answers at different times and in different ways. As a result, you may obtain answers that are slightly different from those presented here. To help minimize this problem, we have tried to include the numerical values obtained at different stages of complex problems rather than presenting a single final answer.

CHAPTER 1: INTRODUCTION TO STATISTICS

1. a. The population is the entire set of adolescent boys who are taking medication for depression.
b. The sample is the group of 30 boys who were tested in the study.

3. Descriptive statistics are used to simplify and summarize data. Inferential statistics use sample data to make general conclusions about populations.

5. A correlational study has only one group of individuals and measures two (or more) different variables for each individual. Other research methods evaluating relationships between variables compare two (or more) different groups of scores.

7. The independent variable is holding a pen in your teeth versus holding the pen in your lips. The dependent variable is the rating given to each cartoon.

9. a. This is a nonexperimental study. The researcher is simply observing, not manipulating, two variables.
b. This is an experiment. The researcher is manipulating the type of drink and should control other variables by beginning with equivalent groups of participants.

11. This is not an experiment because there is no manipulation. Instead, the study is comparing two preexisting groups (American and Canadian students).

13. a. continuous. Time is infinitely divisible.
b. discrete. Family size consists of whole-number categories that cannot be divided.
c. discrete. There are two separate and distinct categories (analog and digital).
d. continuous. The variable is knowledge of statistics, which is measured with quiz scores. It could be a 5-point quiz, a 10-point quiz, or a 50-point quiz, which indicates that knowledge can be divided indefinitely.

15. a. The independent variable is humorous versus nonhumorous.
b. The independent variable is measured on a nominal scale.
c. The dependent variable is the number of sentences recalled.
d. The dependent variable is measured on a ratio scale.

17. a. The independent variable is whether or not the motivational signs were posted, and the dependent variable is amount of use of the stairs.
b. Posting versus not posting is measured on a nominal scale.

19. a. $\Sigma X = 15$
b. $\Sigma X^2 = 65$
c. $\Sigma(X + 1) = 20$
d. $\Sigma(X + 1)^2 = 100$

21. a. $\Sigma X = 11$
b. $\Sigma Y = 25$
c. $\Sigma XY = 54$

23. a. $\Sigma X^2 = 30$
b. $(\Sigma X)^2 = 64$
c. $\Sigma(X - 2) = 0$
d. $\Sigma(X - 2)^2 = 14$

CHAPTER 2: FREQUENCY DISTRIBUTIONS

1.

X	f
10	3
9	6
8	4
7	2
6	3
5	1
4	1

3. a. $n = 12$
b. $\Sigma X = 40$
c. $\Sigma X^2 = 148$

5. a.

X	f
28-29	1
26-27	4
24-25	7
22-23	4
20-21	2
18-19	2
16-17	1
14-15	0
12-13	1
10-11	1
8-9	1

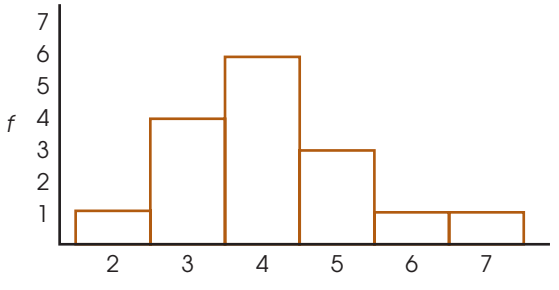
b.

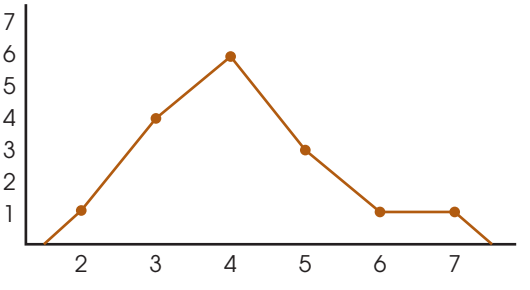
X	f
25-29	8
20-24	10
15-19	3
10-14	2
5-9	1

7. a. 2 points wide and around 8 intervals
b. 5 points wide and around 12 intervals or 10 points wide and around 6 intervals
c. 10 points wide and around 9 intervals

9. A bar graph leaves a space between adjacent bars and is used with data from nominal or ordinal scales. In a histogram, adjacent bars touch at the real limits. Histograms are used to display data from interval or ratio scales.

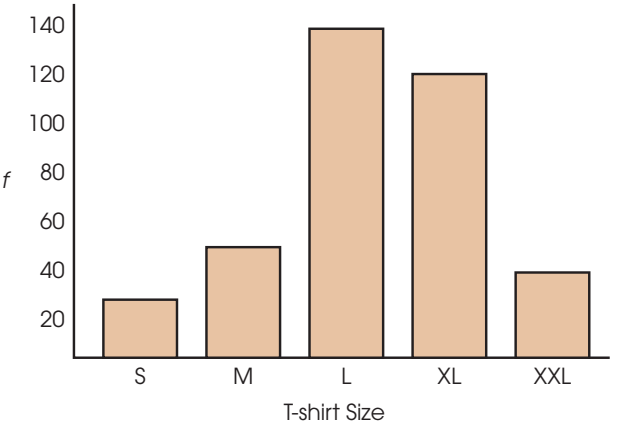
11.





13. a. A bar graph should be used for measurements from an ordinal scale.

b.



15. a.

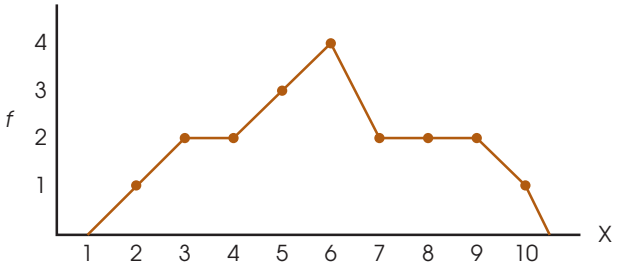
X	f
9	1
8	1
7	4
6	5
5	7
4	2

b. positively skewed

17. a.

X	f
10	1
9	2
8	2
7	2
6	4
5	3
4	2
3	2
2	1

b.



c. It is a fairly symmetrical distribution centered at $X = 6$. The scores are scattered across the scale.

19.

X	f	cf	c%
7	2	24	100
6	3	23	92
5	6	20	80
4	9	14	56
3	4	5	20
2	1	1	4

a. The percentile rank for $X = 2.5$ is 4%
b. The percentile rank for $X = 6.5$ is 92%
c. The 20th percentile is $X = 3.5$.
d. The 80th percentile is $X = 5.5$.

21.

X	f	cf	c%
10	2	50	100
9	5	48	96
8	8	43	86
7	15	35	70
6	10	20	40
5	6	10	20
4	4	4	8

a. The percentile rank for $X = 6$ is 30%.
b. The percentile rank for $X = 9$ is 91%.
c. The 25th percentile is $X = 5.75$.
d. The 90th percentile is $X = 8.9$.

23. a. The percentile rank for $X = 5$ is 8%.
b. The percentile rank for $X = 12$ is 85%.
c. The 25th percentile is $X = 7$.
d. The 70th percentile is $X = 10$.

25.

1	796
2	0841292035826
3	094862
4	543
5	3681
6	4

27.

2	80472
3	49069
4	543976
5	4319382
6	5505
7	24
8	1

CHAPTER 3: CENTRAL TENDENCY

1. The purpose of central tendency is to identify a single score that serves as the best representative for an entire distribution, usually a score from the center of the distribution.

3. The mean is $\frac{29}{10} = 2.9$, the median is 2.5, and the mode is 2.

5. The mean is $\frac{69}{12} = 5.75$, the median is 6, and the mode is 7.

7. a. Median = 2.83 (2.5 + 0.33)
b. Median = 3

9. $N = 25$

11. The original sample has $n = 5$ and $\Sigma X = 60$. The new sample has $n = 4$ and $\Sigma X = 52$. The new mean is $M = 13$.

13. After the score is removed, $n = 8$, $\Sigma X = 88$, and $M = 11$.

15. After the score is changed, $n = 7$, $\Sigma X = 49$, and $M = 7$.

17. The original sample has $n = 16$ and $\Sigma X = 320$. The new sample has $n = 15$ and $\Sigma X = 285$. The score that was removed must be $X = 35$.

19. a. The new mean is $\frac{75}{10} = 7.5$.

b. The new mean is (20 + 60)/10 = 8
c. The new mean is (30 + 40)/10 = 7

21. The median is used instead of the mean when there is a skewed distribution (a few extreme scores), an open-ended distribution, undetermined scores, or an ordinal scale.

23. a. Mode = 2
b. Median = 2
c. You cannot find the total number of fast-food visits (ΣX) for this sample.

25. a. For weekdays $M = 0.99$ inches and for weekend days is $M = 1.67$ inches.
b. There does appear to be more rain on weekend days than there is on weekdays.

CHAPTER 4: VARIABILITY

1. a. *SS* is the sum of squared deviation scores.
b. Variance is the mean squared deviation.
c. Standard deviation is the square root of the variance. It provides a measure of the standard distance from the mean.

3. Standard deviation and variance are measures of distance and are always greater than or equal to zero.

5. Without some correction, the sample variance underestimates the variance for the population. Changing the formula for sample variance (using $n - 1$ instead of N) is the necessary correction.

7. a. $s = 2$ is better (you are above the mean by 3 standard deviations).
b. $s = 10$ is better (you are below the mean by less than half a standard deviation).

9. a. The original mean is $M = 80$ and the standard deviation is $s = 8$.
b. The original mean is $M = 12$ and the standard deviation is $s = 3$.

11. a. The range is either 11 or 12, and the standard deviation is $\sigma = 4$.
b. After adding 2 points to each score, the range is still either 11 or 12, and the standard deviation is still $\sigma = 4$. Adding a constant to every score does not affect measures of variability.

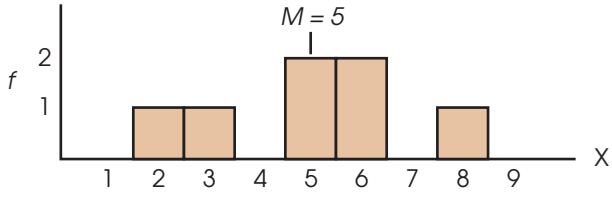
13. For sample A the mean is $M = 4.50$, so the computational formula would be easier. For this sample, $SS = 25$. For sample B the mean is $M = 4$ and the definitional formula would be easier. For this sample, $SS = 42$.

15. a. The mean is $M = 4$ and the standard deviation is $s = \sqrt{9} = 3$.
b. The new mean is $M = 6$ and the new standard deviation is $\sqrt{49} = 7$.
c. Changing one score changes both the mean and the standard deviation.

17. $SS = 32$, the population variance is 4, and the standard deviation is 2.

19. $SS = 36$, the sample variance is 9, and the standard deviation is 3.

21. a.



b. The mean is $\frac{35}{7} = 5$. The two scores of $X = 5$ are exactly equal to the mean. The scores $X = 2$ and $X = 8$ are farthest from the mean (3 points). The standard deviation should be between 0 and 3 points.
c. $SS = 24$, $s^2 = 4$, $s = 2$, which agrees with the estimate.

23. a. For the younger woman, the variance is $s^2 = 0.786$. For the older woman, the variance is $s^2 = 1.696$.
b. The variance for the younger woman is only half as large as for the older woman. The younger woman's scores are much more consistent.

SECTION I REVIEW

1. a. The goal for descriptive statistics is to simplify, organize, and summarize data so that it is easier for researchers to see patterns.
b. A frequency distribution provides an organized summary of the complete set of scores.
c. A measure of central tendency summarizes an entire set of scores with a single value that is representative of the whole set.
d. A measure of variability provides a single number that describes the differences that exist from one score to another.

2. a.

The original rats appear to make far more errors that the seventh-generation maze-bright rats

b. The original rats made an average of $M = 12.43$ errors compared to an average of only $M = 7.33$ for the maze-bright rats. On average, the original rats made far more errors.

c. For the original rats, $SS = 427.14$, the variance is $s^2 = 21.36$ and the standard deviation is $s = 4.62$. For the maze-bright rats, $SS = 54.67$, the variance is $s^2 = 2.73$ and the standard deviation is $s = 1.65$. The error scores for the original rats are much more spread out. The seventh generation rats are a much more homogeneous group.

CHAPTER 5: *z*-SCORES

1. The sign of the *z*-score tells whether the location is above (+) or below (−) the mean, and the magnitude tells the distance from the mean in terms of the number of standard deviations.

3. a. above the mean by 12 points
b. above the mean by 3 points
c. below the mean by 12 points
d. below the mean by 3 points

5.

X	*z*	*X*	*z*	*X*	*z*
45	0.71	51	1.57	41	0.14
30	−1.43	25	−2.14	38	−0.29

7. a.

X	*z*	*X*	*z*	*X*	*z*
44	0.50	50	1.25	52	1.50
34	−0.75	28	−1.50	64	3.00

b.

X	*z*	*X*	*z*	*X*	*z*
46	0.75	52	1.50	24	−2.00
38	−0.25	36	−0.50	50	1.25

9.

X	*z*	*X*	*z*	*X*	*z*
88	0.80	92	1.20	100	2.00
76	−0.40	74	−0.60	62	−1.80

11. a. $X = 41$
b. $X = 42$
c. $X = 43$
d. $X = 45$

13. $\sigma = 4$

15. $M = 50$

17. $\sigma = 4$

19. $\mu = 61$ and $\sigma = 3$. The distance between the two scores is 3 points which is equal to 1.0 standard deviation.

21. a. $\sigma = 4$
b. $\sigma = 8$

23. a. $X = 95$ ($z = -0.25$)
b. $X = 80$ ($z = -1.00$)
c. $X = 125$ ($z = 1.25$)
d. $X = 110$ ($z = 0.50$)

25. a. $\mu = 5$ and $\sigma = 4$
b. and **c.**

Original *X*	*z*-score	Transformed *X*
0	1.25	75
6	0.25	105
4	−0.25	95
3	−0.50	90
12	1.75	135

CHAPTER 6: PROBABILITY

1. a. $p = \frac{1}{50} = 0.02$
b. $p = \frac{10}{50} = 0.20$
c. $p = \frac{20}{50} = 0.40$

3. The two requirements for a random sample are: (1) each individual has an equal chance of being selected, and (2) if more than one individual is selected, the probabilities must stay constant for all selections.

5. a. tail to the right, $p = 0.0228$
b. tail to the right, $p = 0.2743$
c. tail to the left, $p = 0.0968$
d. tail to the left, $p = 0.3821$

7. a. $p(z > 0.25) = 0.4013$
b. $p(z > -0.75) = 0.7734$
c. $p(z < 1.20) = 0.8849$
d. $p(z < -1.20) = 0.1151$

9. a. $p = 0.1974$
b. $p = 0.9544$
c. $p = 0.4592$
d. $p = 0.4931$

11. a. $z = \pm 0.25$
b. $z = \pm 0.67$
c. $z = \pm 1.96$
d. $z = \pm 2.58$

13. a. tail to the right, $p = 0.4013$
b. tail to the left, $p = 0.3085$
c. tail to the right, $p = 0.0668$
d. tail to the left, $p = 0.1587$

15. a. $z = 2.00$, $p = 0.0228$
b. $z = 0.50$, $p = 0.3085$
c. $z = 1.28$, $X = 628$
d. $z = -0.25$, $X = 475$

17. a. $p(z > 1.50) = 0.0668$
b. $p(z < -2.00) = 0.0228$

19. a. $z = 0.60$, $p = 0.2743$
b. $z = -1.40$, $p = 0.0808$
c. $z = 0.84$, $X = \$206$ or more

21. $p(X > 36) = p(z > 2.17) = 0.0150$ or 1.50%

23. a. $p = \frac{1}{2}$
b. $\mu = 20$
c. $\mu = \sqrt{10} = 3.16$ and for $X = 25.5$, $z = 1.74$, and $p = 0.0409$
d. For $X = 24.5$, $z = 1.42$, and $p = 0.0778$

25. a. With five options, $p = \frac{1}{5}$ for each trial, $\mu = 20$, and $\sigma = 4$; for $X = 20$, $z = \pm 0.13$ and $p = 0.1034$.
b. For $X = 30.5$, $z = 2.63$ and $p = 0.0043$.
c. $\mu = 40$ and $\sigma = 5.66$; for $X = 49.5$, $z = 1.68$ and $p = 0.0465$

27. a. With $n = 50$ and $p = q = \frac{1}{2}$, you may use the normal approximation with $\mu = 25$ and $\sigma = 3.54$. Using the upper real limit of 30.5, $p(X > 30.5) = p(z > 1.55) = 0.0606$.
b. The normal approximation has $\mu = 50$ and $\sigma = 5$. Using the upper real limit of 60.5, $p(X > 60.5) = p(z > 2.10) = 0.0179$.
c. Getting 60% heads with a balanced coin is an unusual event for a large sample. Although you might get 60% heads with a small sample, you should get very close to a 50-50 distribution as the sample gets larger. With a larger sample, it becomes very unlikely to get 60% heads.

CHAPTER 7: THE DISTRIBUTION OF SAMPLE MEANS

1. a. The distribution of sample means consists of the sample means for all the possible random samples of a specific size (n) from a specific population.
b. The expected value of M is the mean of the distribution of sample means (μ).
c. The standard error of M is the standard deviation of the distribution of sample means ($\sigma_M = \frac{\sigma}{\sqrt{n}}$).

3. a. The expected value is $\mu = 40$ and $\sigma_M = \frac{8}{\sqrt{4}} = 4$.
b. The expected value is $\mu = 40$ and $\sigma_M = \frac{8}{\sqrt{16}} = 2$.

5. a. Standard error $= \frac{30}{\sqrt{4}} = 15$ points
b. Standard error $= \frac{30}{\sqrt{25}} = 6$ points
c. Standard error $= \frac{30}{\sqrt{100}} = 3$ points

7. a. $n \geq 16$
b. $n \geq 100$
c. $n \geq 400$

9. a. $\sigma = 50$
b. $\sigma = 25$
c. $\sigma = 10$

11. a. $\sigma_M = 5$ points and $z = -1.00$
b. $\sigma_M = 10$ points and $z = -0.50$
c. $\sigma_M = 20$ points and $z = -0.25$

13. a. With a standard error of 4, $M = 33$ corresponds to $z = 0.75$, which is not extreme.

b With a standard error of 1, $M = 33$ corresponds to $z = 3.00$, which is extreme.

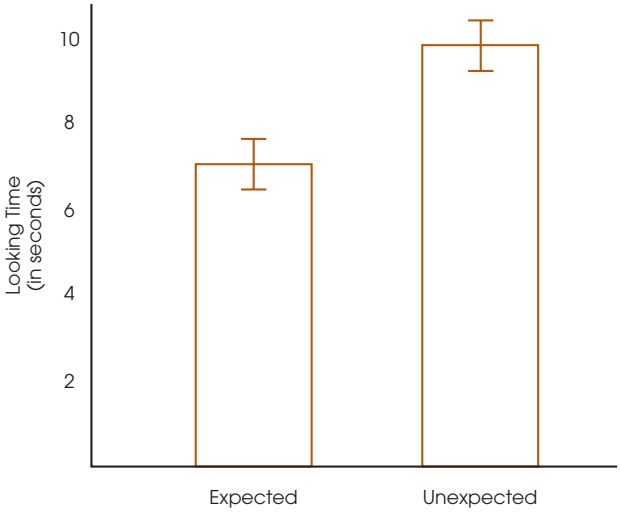
15. a. $z = 0.50$ and $p = 0.6915$
b. $\sigma_M = 5$, $z = 1.00$ and $p = 0.8413$
c. $\sigma_M = 2$, $z = 2.50$ and $p = 0.9938$

17. a. $z = \pm 0.50$ and $p = 0.3830$
b. $\sigma_M = 5$, $z = \pm 1.00$ and $p = 0.6826$
c. $\sigma_M = 2.5$, $z = \pm 2.00$ and $p = 0.9544$

19. a. $p(z < -0.50) = 0.3085$
b. $p(z < -1.00) = 0.1587$

21. a. With a standard error of 3.58 this sample mean corresponds to a z-score of $z = 1.28$. A z-score this large (or larger) has a probability of $p = 0.1003$.
b. A sample mean this large should occur only 1 out of 10 times. This is not a very representative sample.

23.



CHAPTER 8: INTRODUCTION TO HYPOTHESIS TESTING

1. a. $M - \mu$ measures the difference between the sample mean and the hypothesized population mean.
b. A sample mean is not expected to be identical to the population mean. The standard error measures how much difference, on average, is reasonable to expect between M and μ.

3. The alpha level is a small probability value that defines the concept of "very unlikely." The critical region consists of outcomes that are very unlikely to occur if the null hypothesis is true, where "very unlikely" is defined by the alpha level.

5. a. The null hypothesis states that the herb has no effect on memory scores.
b. H_0: $\mu = 80$ (even with the herbs, the mean is still 80). H_1: $\mu \neq 80$ (the mean has changed)
c. The critical region consists of z-scores beyond ± 1.96.
d. For these data, the standard error is 3 and $z = \frac{4}{3} = 1.33$.
e. Fail to reject the null hypothesis. The herbal supplements do not have a significant effect on memory scores.

7. a. H_0 $\mu = 80$. With $\sigma = 12$, the sample mean corresponds to $z = -\frac{4}{3} = -1.33$. This is not sufficient to reject the null hypothesis. You cannot conclude that the course has a significant effect.
b. H_0 $\mu = 80$. With $\sigma = 6$, the sample mean corresponds to $z = -\frac{4}{1.5} = -2.67$. This is sufficient to reject the null hypothesis and conclude that the course does have a significant effect.
c. There is a 4 point difference between the sample and the hypothesis. In part a, the standard error is 3 points and the 4-point difference is not significant. However, in part b, the standard error is only 1.5 points and the 4-point difference is now significantly more than is expected by chance. In general, a larger standard deviation produces a larger standard error, which reduces the likelihood of rejecting the null hypothesis.

9. a. With $\sigma = 18$, the standard error is 3, and $z = -\frac{8}{3} = -2.67$. Reject H_0.
b. With $\sigma = 30$, the standard error is 5, and $z = -\frac{8}{5} = -1.60$. Fail to reject H_0.
c. Larger variability reduces the likelihood of rejecting H_0.

11. a. With a 2-point treatment effect, for the z-score to be greater than 1.96, the standard error must be smaller than 1.02. The sample size must be greater than 96.12; a sample of $n = 97$ or larger is needed.
b. With a 1-point treatment effect, for the z-score to be greater than 1.96, the standard error must be smaller than 0.51. The sample size must be greater than 384.47; a sample of $n = 385$ or larger is needed.

13. a. H_0: $\mu = 4.9$ and the critical values are ± 1.96. The standard error is 0.21 and $z = -3.33$. Reject the null hypothesis.
b. Cohen's $d = \frac{0.7}{0.84} = 0.833$ or 83.3%
c. The results indicate that the presence of a tattoo has a significant effect on the judged attractiveness of a woman, $z = -3.33$, $p < .01$, $d = 0.833$.

15. a. H_0: $\mu \leq 50$ (endurance is not increased). The critical region consists of z-scores beyond $z = +1.65$.

For these data, $\sigma_M = 1.70$ and $z = 1.76$. Reject H_0 and conclude that endurance scores are significantly higher with the sports drink.

b. H_0: $\mu = 50$ (no change in endurance). The critical region consists of z-scores beyond $z = \pm 1.96$. Again, $\sigma_M = 1.70$ and $z = 1.76$. Fail to reject H_0 and conclude that the sports drink does not significantly affect endurance scores.

c. The two-tailed test requires a larger z-score for the sample to be in the critical region.

17. H_0: $\mu \leq 12$ (no increase during hot weather). H_1: $\mu > 12$ (there is an increase). The critical region consists of z-score values greater than $+1.65$. For these data, the standard error is 1.50, and $z = 2.33$ which is in the critical region so we reject the null hypothesis and conclude that there is a significant increase in the average number of hit players during hot weather.

19. a. With no treatment effect the distribution of sample means is centered at $\mu = 75$ with a standard error of 1.90 points. The critical boundary of $z = 1.96$ corresponds to a sample mean of $M = 78.72$. With a 4-point treatment effect, the distribution of sample means is centered at $\mu = 79$. In this distribution a mean of $M = 78.72$ corresponds to $z = -0.15$. The power for the test is the probability of obtaining a z-score greater than -0.15, which is $p = 0.5596$.

b. With a one-tailed test, critical boundary of $z = 1.65$ corresponds to a sample mean of $M = 78.14$. With a 4-point treatment effect, the distribution of sample means is centered at $\mu = 79$. In this distribution a mean of $M = 78.14$ corresponds to $z = -0.45$. The power for the test is the probability of obtaining a z-score greater than -0.45, which is $p = 0.6736$.

21. a. Increasing alpha increases power.

b. Changing from one- to two-tailed decreases power.

23. a. For a sample of $n = 16$ the standard error would be 5 points, and the critical boundary for $z = 1.96$ corresponds to a sample mean of $M = 89.8$. With a 12-point effect, the distribution of sample means would be centered at $\mu = 92$. In this distribution, the critical boundary of $M = 89.8$ corresponds to $z = -0.44$. The power for the test is $p(z > -0.44) = 0.6700$ or 67%.

b. For a sample of $n = 25$ the standard error would be 4 points, and the critical boundary for $z = 1.96$ corresponds to a sample mean of $M = 87.84$. With a 12-point effect, the distribution of sample means would be centered at $\mu = 92$. In this distribution, the critical boundary of $M = 87.84$ corresponds to $z = -1.04$. The power for the test is $p(z > -1.04) = 0.8508$ or 85.08%.

SECTION II **REVIEW**

1. a. $z = 1.50$

b. $X = 36$

c. If the entire population of X values is transformed into z-scores, the set of z-scores will have a mean of 0 and a standard deviation of 1.00.

d. The standard error is 4 points and $z = 0.50$.

e. The standard error is 2 points and $z = 1.00$.

2. a. $p(X > 40) = p(z > 0.36) = 0.3594$ or 35.94%.

b. $p(X < 10) = p(z < -1.79) = 0.0367$ or 3.67%.

c. The standard error is 2 points and $z = -2.50$. The probability is $p = 0.0062$.

3. a. The null hypothesis states that the overweight students are no different from the overall population, $\mu = 4.22$. The standard error is 0.10 and the z-score for this sample is $z = 2.60$. Reject the null hypothesis. The number of snacks eaten by overweight students is significantly different from the number for the general population.

b. The null hypothesis states that the healthy-weight students do not eat fewer snacks than the overall population, H_0: $\mu \geq 4.22$. The standard error is 0.12 and the z-score for this sample is $z = -1.75$. For a one-tailed test, the critical value is $z = -1.65$. Reject the null hypothesis. The number of snacks eaten by healthy-weight students is significantly less than the number for the general population.

CHAPTER 9: **INTRODUCTION TO THE *t* STATISTIC**

1. A z-score is used when the population standard deviation (or variance) is known. The t statistic is used when the population variance or standard deviation is unknown. The t statistic uses the sample variance or standard deviation in place of the unknown population values.

3. a. The sample variance is 16 and the estimated standard error is 2.

b. The sample variance is 54 and the estimated standard error is 3.

c. The sample variance is 12 and the estimated standard error is 1.

5. a. $t = \pm 2.571$

b. $t = \pm 2.201$

c. $t = \pm 2.069$

7. a. $M = 5$ and $s = \sqrt{20} = 4.47$

b. $s_M = 2$.

9. a. With $s = 9$, $s_M = 3$ and $t = -\frac{7}{3} = -2.33$. This is beyond the critical boundaries of ± 2.306, so we reject the null hypothesis and conclude that there is a significant treatment effect.

b. With $s = 15$, $s_M = 5$ and $t = -\frac{7}{5} = -1.40$. This value is not beyond the critical boundaries, so there is no significant effect.

c. As the sample variability increases, the likelihood of rejecting the null hypothesis decreases.

11. a. With a two tailed test, the critical boundaries are ± 2.306 and the obtained value of $t = \frac{3.3}{1.5} = 2.20$ is not sufficient to reject the null hypothesis.

b. For the one-tailed test the critical value is 1.860, so we reject the null hypothesis and conclude that participants significantly overestimated the number who noticed.

13. a. With $df = 15$, the critical values are ± 2.947. For these data, the sample variance is 16, the estimated standard error is 1, and $t = \frac{8.2}{1} = 8.20$. Reject the null hypothesis and conclude that there has been a significant change in the level of anxiety.

b. With $df = 15$, the t values for 90% confidence are ± 1.753, and the interval extends from 21.547 to 25.053.

c. The data indicate a significant change in the level of anxiety, $t(16) = 8.20$, $p < .01$, 95% CI [21.547, 25.053].

15. a. With $df = 63$, the critical values are ± 2.660 (using $df = 60$ in the table). For these data, the estimated standard error is 1.50, and $t = \frac{7}{1.50} = 4.67$. Reject the null hypothesis and conclude that there has been a significant change in the average IQ score.

b. Using $df = 60$, the t values for 80% confidence are ± 1.296, and the interval extends from 105.056 to 108.944.

17. a. The estimated standard error is 1.50, and $t = \frac{7.7}{1.50} = 5.13$. For a one-tailed test, the critical value is 2.602. Reject the null hypothesis, children with a history of day care have significantly more behavioral problems.

b. The percentage of variance accounted for is $r^2 = \frac{26.32}{41.32} = 0.637$ or 63.7%.

c. The results show that kindergarten children with a history of day care have significantly more behavioral problems than other kindergarten children, $t(15) = 5.13$, $p < .01$, $r^2 = 0.637$.

19. a. Cohen's $d = \frac{3}{6} = 0.50$. With $s = 6$, the estimated standard error is 1.2 and $t = \frac{3}{1.2} = 2.50$. $r^2 = \frac{6.25}{30.25} = 0.207$.

b. Cohen's $d = \frac{3}{15} = 0.20$. With $s = 15$, the estimated standard error is 3 and $t = \frac{3}{3} = 1.00$. $r^2 = \frac{1.00}{25.00} = 0.04$.

c. Measures of effect size tend to decrease as sample variance increases.

21. a. The estimated standard error is 0.20 and $t = \frac{2.2}{0.2} = 11.00$. The t value is well beyond the critical value of 2.492. Reject the null hypothesis.

b. Cohen's $d = \frac{2.2}{1} = 2.20$ and $r^2 = \frac{121}{145} = 0.8345$.

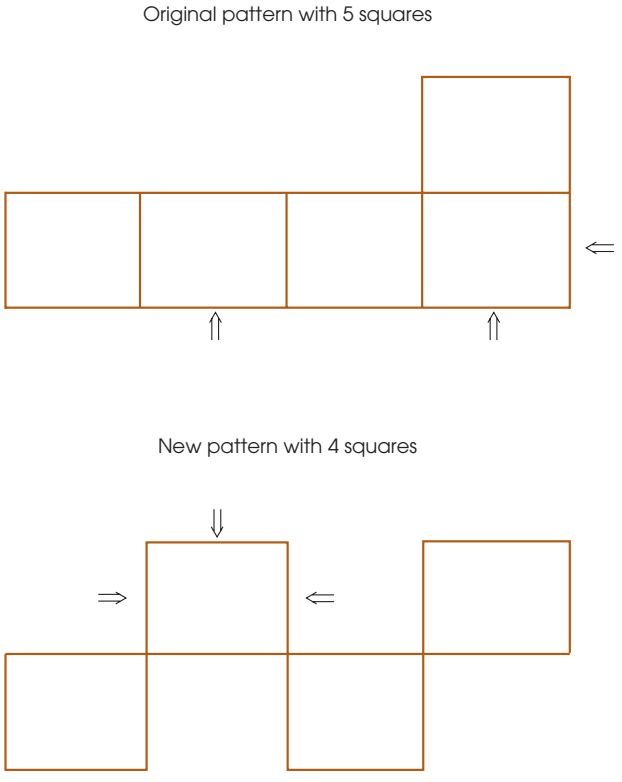
23. a. H_0: $\mu = 40$. With $df = 8$ the critical values are $t = \pm 2.306$. For these data, $M = 44$, $SS = 162$, $s^2 = 20.25$, the standard error is 1.50, and $t = 2.67$. Reject H_0 and conclude that depression for the elderly is significantly different from depression for the general population.

b. Cohen's $d = \frac{4}{4.5} = 0.889$.

c. The results indicate that depression scores for the elderly are significantly different from scores for the general population, $t(8) = 2.67$, $p < .05$, $d = 0.889$.

CHAPTER 10: THE *t* TEST FOR TWO INDEPENDENT SAMPLES

There are several possible solutions to the matchstick problem in the Chapter 10 Preview but all involve destroying two of the existing squares. One square is destroyed by removing two matchsticks from one of the corners and a second square is destroyed by removing one matchstick. The three removed matchsticks are then used to build a new square using a line that already exists in the figure as the fourth side. One solution is shown in the following figure. Arrows indicate the three matchsticks to be removed from the original pattern and their locations in the new pattern.



1. An independent-measures study uses a separate sample for each of the treatments or populations being compared.

3. a. The size of the two samples influences the magnitude of the estimated standard error in the denominator of the t statistic. As sample size increases, the value of t also increases (moves farther from zero), and the likelihood of rejecting H_0 also increases.

b. The variability of the scores influences the estimated standard error in the denominator. As the variability of the scores increases, the value of t decreases (becomes closer to zero), and the likelihood of rejecting H_0 decreases.

5. a. The first sample has $s^2 = 12$ and the second has $s^2 = 8$. The pooled variance is $\frac{80}{8} = 10$ (halfway between).

b. The first sample has $s^2 = 12$ and the second has $s^2 = 4$. The pooled variance is $\frac{80}{12} = 6.67$ (closer to the variance for the larger sample).

7. a. The pooled variance is 6 and the estimated standard error is 1.50.

b. The pooled variance is 24 and the estimated standard error is 3.

c. Larger variability produces a larger standard error.

9. a. The pooled variance is 90.

b. The estimated standard error is 5.

c. A mean difference of 10 points produces $t = 2.00$. With critical boundaries of ± 2.160, fail to reject H_0.

d. A mean difference of 13 points produces $t = 2.60$. With critical boundaries of ± 2.160, reject H_0

11. a. The pooled variance is 60 and the estimated standard error is 5.

b. The pooled variance is 240 and the estimated standard error is 10.

c. Increasing the sample variance produces an increase in the standard error.

13. a. Using $df = 30$, because 34 is not listed in the table, and $\alpha = .05$, the critical region consists of t values beyond ± 2.042. The pooled variance is 81, the estimated standard error is 3, and $t(34) = \frac{7.6}{3} = 2.53$. The t statistic is in the critical region. Reject H_0 and conclude that there is a significant difference.

b. For 90% confidence, the t values are ± 1.697 (using $df = 30$), and the interval extends from 2.509 to 12.691 points higher with the calming music.

c. Classroom performance was significantly better with background music, $t(34) = 2.53$, $p < .05$, 95% CI [2.509, 12.691].

15. a. For the offensive linemen, the standard error is 0.97 and $t = 4.54$. For a one-tailed test with $df = 16$, the critical value is 2.583. Reject the null hypothesis. The offensive linemen are significantly above the criterion for BMI.

b. For the defensive linemen, the standard error is 0.80 and $t = 2.375$. For a one-tailed test with $df = 18$, the critical value is 2.552. Fail to reject the null hypothesis. The defensive linemen are not significantly above the criterion for BMI.

c. For the independent-measures t, the pooled variance is 14.01, the estimated standard error is 1.25, and $t(34) = 2.00$. For a two-tailed test using $df = 30$ (because 34 is not listed), the critical value is 2.750. Fail to reject the null hypothesis. There is no significant difference between the two groups.

17. a. The research prediction is that participants who hear the verb "smashed into" will estimate higher speeds than those who hear the verb "hit." For these data, the pooled variance is 33, the estimated standard error is 2.10, and $t(28) = 3.24$. With $df = 28$ and $\alpha = .01$, the critical value is $t = 2.467$. The sample mean difference is in the right direction and is large enough to be significant. Reject H_0.

b. The estimated Cohen's $d = \frac{6.8}{\sqrt{33}} = 1.18$.

c. The results show that participants who heard the verb "smashed into" estimated significantly higher speeds than those who heard the verb "hit," $t(28) = 3.24$, $p < .01$, $d = 1.18$.

19. a. The null hypothesis states that there is no difference between the two sets of instructions, H_0: $\mu_1 - \mu_2 = 0$. With $df = 6$ and $\alpha = .05$, the critical region consists of t values beyond ± 2.447. For the first set, $M = 6$ and $SS = 16$. For the second set, $M = 10$ with $SS = 32$. For these data, the pooled variance is 8, the estimated standard error is 2, and $t(6) = 2.00$. Fail to reject H_0. The data are not sufficient to conclude that there is a significant difference between the two sets of instructions.

b. For these data, the estimated $d = \frac{4}{\sqrt{8}} = 1.41$ (a very large effect) and $r^2 = \frac{4}{10} = 0.40$ (40%).

21. The humorous sentences produced a mean of $M = 4.25$ with $SS = 35$, and the non-humorous sentences had $M = 4.00$ with $SS = 26$. The pooled variance is 2.03, the estimated standard error is 0.504, and $t = 0.496$. With $df = 30$, the critical value is 2.042. Fail to reject the null hypothesis and conclude that there is no significant difference in memory for the two types of sentences.

23. a. The null hypothesis states that the lighting in the room does not affect behavior. For the well-lit room the mean is $M = 7.55$ with $SS = 42.22$. For the dimly-lit room, $M = 11.33$ with $SS = 38$. The pooled variance is 5.01, the standard error is 1.06, and $t(16) = 3.57$. With $df = 16$ the critical values are ± 2.921. Reject the null hypothesis and conclude that the lighting did have an effect on behavior.

b. $d = \frac{3.78}{2.24} = 1.69$.

CHAPTER 11: THE *t* TEST FOR TWO RELATED SAMPLES

1. a. This is an independent-measures experiment with two separate samples.

b. This is repeated-measures. The same individuals are measured twice.

c. This is repeated-measures. The same individuals are measured twice.

3. For a repeated-measures design the same subjects are used in both treatment conditions. In a matched-subjects design, two different sets of subjects are used. However, in a matched-subjects design, each subject in one condition is matched with respect to a specific variable with a subject in the second condition so that the two separate samples are equivalent with respect to the matching variable.

5. a. The standard deviation is 5 points and measures the average distance between an individual score and the sample mean.

b. The estimated standard error is 1.67 points and measures the average distance between a sample mean and the population mean.

7. a. The estimated standard error is 2 points and $t(8) = 1.50$. With a critical boundary of ± 2.306, fail to reject the null hypothesis.

b. With $M_D = 12$, $t(8) = 6.00$. With a critical boundary of ± 2.306, reject the null hypothesis.

c. The larger the mean difference, the greater the likelihood of finding a significant difference.

9. The sample variance is 9, the estimated standard error is 0.75, and $t(15) = 4.33$. With critical boundaries of ± 2.131, reject H_0

11. a. The null hypothesis says that there is no difference in judgments for smiling versus frowning. For these data, the sample variance is 6.25, the estimated standard error is 0.5, and $t = \frac{1.6}{0.5} = 3.20$. For a one-tailed test with $df = 24$, the critical value is 2.492. Reject the null hypothesis.

b. $r^2 = \frac{10.24}{34.24} = 0.299$ (29.9%)

c. The cartoons were rated significantly funnier when people held a pen in their teeth compared to holding a pen in their lips, $t(24) = 3.20$, $p < .01$, one tailed, $r^2 = 0.299$.

13. The null hypothesis states that there is no difference in the perceived intelligence between attractive and unattractive photos. For these data, the estimated standard error is 0.4 and $t = \frac{2.7}{0.4} = 6.75$. With $df = 24$, the critical value is 2.064. Reject the null hypothesis.

15. a. The difference scores are 3, 7, 3, and 3. $M_D = 4$.

b. $SS = 12$, sample variance is 4, and the estimated standard error is 1.

c. With $df = 3$ and $\alpha = .05$, the critical values are $t = \pm 3.182$. For these data, $t = 4.00$. Reject H_0. There is a significant treatment effect.

17. The null hypothesis states that the images have no effect on performance. For these data, the sample variance is 12.6, the estimated standard error is 1.45, and $t(5) = 2.97$. With $df = 5$ and $\alpha = .05$, the critical values are $t = \pm 2.571$. Reject the null hypothesis, the images have a significant effect.

19. a. The pooled variance is 6.4 and the estimated standard error is 1.46.

b. For the difference scores the variance is 24, the estimated standard error is 2.

21. a. The null hypothesis says that changing answers has no effect, H_0: $\mu_D = 0$. With $df = 8$ and $\alpha = .05$, the

critical values are $t = \pm 2.306$. For these data, $M_D = 7$, $SS = 288$, the standard error is 2, and $t(8) = 3.50$. Reject H_0 and conclude that changing answers has a significant effect on exam performance.

b. For 95% confidence use $t = \pm 2.306$. The interval extends from 2.388 to 11.612.

c. Changing answers resulted in significantly higher exam scores, $t(8) = 3.50$, $p < .05$, 95% CI [2.388, 11.612].

23. The null hypothesis says that there is no difference between shots fired during versus between heart beats, H_0: $\mu_D = 0$. With $\alpha = .05$, the critical region consists of t values beyond ± 2.365. For these data, $M_D = 3$, $SS = 36$, $s^2 = 5.14$, the standard error is 0.80, and $t(7) = 3.75$. Reject H_0 and conclude that the timing of the shot has a significant effect on the marksmen's scores.

SECTION III **REVIEW**

1. a. For these data, the mean is $M = 23$ and the standard deviation is $s = 3$.

b. H_0: $\mu \leq 20$. With $df = 8$, the critical region consists of t values greater than 1.860. For these data, the standard error is 1, and $t(8) = 3.00$. Reject H_0 and conclude that participation in the interview significantly increases life satisfaction.

c. Cohen's $d = \frac{3}{3} = 1.00$.

d. The 90% confidence interval is $\mu = 23 \pm 1.86$ and extends from 21.14 to 24.86.

2. a. The pooled variance is 1.2, the standard error is 0.40, and $t(28) = \frac{0.7}{0.4} = 1.75$. With a critical value of 2.048, the decision is to fail to reject the null hypothesis.

b. For these data, $r^2 = \frac{3.06}{31.06} = 0.099$ or 9.9%.

c. The presence of a tattoo did not have a significant effect on the attractiveness ratings, $t(28) = 1.75$, $p > .05$, $r^2 = 0.099$.

3. a. The estimated standard error is 2.5 and $t(19) = 1.92$. With a critical value of 1.729, reject the null hypothesis and conclude that attention span is significantly longer with the medication.

b. The 80% confidence interval is $\mu_D = 4.8 \pm 1.328(2.5)$ and extends from 1.48 to 8.12.

CHAPTER 12: **INTRODUCTION TO ANALYSIS OF VARIANCE**

1. When there is no treatment effect, the numerator and the denominator of the F-ratio are both measuring the same sources of variability (random, unsystematic differences from sampling error). In this case, the F-ratio is balanced and should have a value near 1.00.

3. a. As the differences between sample means increase, $MS_{between}$ also increases, and the F-ratio increases.

b. Increases in sample variability cause MS_{within} to increase and, thereby, decrease the F-ratio.

5. a. Posttests are used to determine exactly which treatment conditions are significantly different.

b. If there are only two treatments, then there is no question as to which two treatments are different.

c. If the decision is to fail to reject H_0, then there are no significant differences.

7. a.

Source	SS	df	MS	
Between treatments	84	2	42	$F(2, 15) = 6.00$
Within treatments	105	15	7	
Total	189	17		

With $\alpha = .05$, the critical value is $F = 3.68$. Reject the null hypothesis and conclude that there are significant differences among the three treatments.

b. $\eta^2 = \frac{84}{189} = 0.444$.

c. Analysis of variance showed significant mean differences among the three treatments, $F(2, 15) = 6.00$, $p < .05$, $\eta^2 = 0.444$.

9. a. The sample variances are 4, 5, and 6.

b.

Source	SS	df	MS	
Between treatments	90	2	45	$F(2, 12) = 9.00$
Within treatments	60	12	5	
Total	150	14		

With $\alpha = .05$, the critical value is $F = 3.68$. Reject the null hypothesis and conclude that there are significant differences among the three treatments.

11. a.

Source	SS	df	MS	
Between treatments	70	2	35	$F(2, 12) = 17.50$
Within treatments	24	12	2	
Total	94	14		

With $\alpha = .05$, the critical value is $F = 3.68$. Reject the null hypothesis and conclude that there are significant differences among the three treatments.

b. $\eta^2 = \frac{70}{94} = 0.745$.

c. Analysis of variance showed significant mean differences in perfectionism related to parental criticism among the three groups of students, $F(2, 15) = 6.00$, $p < .05$, $\eta^2 = 0.745$.

13. a. $k = 3$ treatment conditions.

b. The study used a total of $N = 57$ participants.

15.

Source	SS	df	MS	
Between treatments	30	2	15	$F = 5$
Within treatments	63	21	3	
Total	93	23		

17.

Source	SS	df	MS	
Between treatments	20	2	10	$F = 2.50$
Within treatments	180	45	4	
Total	200	47		

19. a. The pooled variance is 6, the estimated standard error is 1.50 and $t(10) = 4.00$. With $df = 10$, the critical value is 2.228. Reject the null hypothesis.

b.

Source	SS	df	MS	
Between treatments	96	1	96	$F(1, 10) = 16$
Within treatments	60	10	6	
Total	156	11		

With $df = 1, 10$, the critical value is 4.96. Reject the null hypothesis. Note that $F = t^2$.

21. a.

Source	SS	df	MS	
Between treatments	252	2	126	$F(2, 15) = 19.30$
Within treatments	98	15	6.53	
Total	350	17		

With $df = 2, 15$ the critical value is 3.68. Reject the null hypothesis.

b. The percentage of variance explained by the mean differences is $\eta^2 = 0.72$ or 72%.

c. The analysis of variance shows significant differences in average brain size among the three groups of birds, $F(2, 15) = 19.30$, $p < .01$, $\eta^2 = 0.72$.

d. With $k = 3$ groups and $df = 15$, $q = 3.67$. The HSD = 3.83. The non-migrating birds are significantly different from either other group, but there is no significant difference between the short- and long-distance migrants.

23. a. The means and standard deviations are

Little	Moderate	Substantial
$M = 4.00$	$M = 5.00$	$M = 6.50$
$s = 2.11$	$s = 2.00$	$s = 1.51$

Source	SS	df	MS	
Between treatments	31.67	2	15.83	$F(2, 27) = 4.25$
Within treatments	100.50	27	3.72	
Total	132.17	29		

With $df = 2, 27$ the critical value is 3.35. Reject the null hypothesis.

b. $\eta^2 = \frac{31.67}{132.17} = 0.240$.

c. Tukey's HSD = 3.49(0.610) = 2.13 (using $df = 30$). The only significant mean difference is between those who watch little or no TV and those whose viewing is substantial.

CHAPTER 13: REPEATED-MEASURES ANOVA

1. For an independent measures design, the variability within treatments is the appropriate error term. For repeated measures, however, you must subtract out variability due to individual differences from the variability within treatments to obtain a measure of error.

3. a. A total of 30 participants is needed; three separate samples, each with $n = 10$. The F-ratio has $df = 2, 27$.

b. One sample of $n = 10$ is needed. The F-ratio has $df = 2, 18$.

5. a. 3 treatments
b. 16 participants

7.

Source	SS	df	MS	
Between treatments	28	2	14	$F(2, 10) = 7.78$
Within treatments	28	15		
Between subjects	10	5		
Error	18	10	1.8	
Total	56	17		

With $df = 2, 10$, the critical value is 4.10. Reject H_0. There are significant differences among the three treatments.

9. a. The null hypothesis states that there are no differences among the three treatments. With $df = 2, 8$, the critical value is 4.46.

Source	SS	df	MS	
Between treatments	70	2	35	$F(2, 8) = 35$
Within treatments	26	12		
Between subjects	18	4		
Error	8	8	1	
Total	96	14		

Reject H_0. There are significant differences among the three treatments.

b. For these data, $\eta^2 = \frac{70}{78} = 0.897$.

c. The analysis of variance shows significant mean differences among the three treatments, $F(2, 8) = 35.00$, $p < .05$, $\eta^2 = 0.897$.

11. The null hypothesis states that there are no differences among the three treatments, H_0: $\mu_1 = \mu_2 = \mu_3$. With $df = 2, 6$, the critical value is 5.14.

Source	SS	df	MS	
Between treatments	8	2	4	$F(2, 6) = 6.00$
Within treatments	94	9		
Between subjects	90	3		
Error	4	6	0.67	
Total	102	11		

Reject H_0. There are significant differences among the three treatments.

13. a. For the independent-measures ANOVA, we obtain:

Source	SS	df	MS	
Between treatments	48	2	24	$F(2, 15) = 3.46$
Within treatments	104	15	6.93	
Total	152	17		

With a critical value of 3.68 for $\alpha = .05$, fail to reject the null hypothesis.

b. For the repeated-measures ANOVA,

Source	SS	df	MS	
Between treatments	48	2	24	$F(2, 10) = 12.00$
Within treatments	104	15		
Between subjects	84	5		
Error	20	10	2	
Total	152	17		

With a critical value of 4.10 for $\alpha = .05$, reject the null hypothesis.

c. The repeated-measures ANOVA reduces the error variance by removing individual differences. This increases the likelihood that the ANOVA will find significant differences.

15.

Source	SS	df	MS	
Between treatments	2	1	2	$F(1, 24) = 4.00$
Within treatment	21	48		
Between subjects	9	24		
Error	12	24	0.5	
Total	23	49		

17.

Source	SS	df	MS	
Between treatments	54	3	18	$F(3, 33) = 6.00$
Within treatments	140	44		
Between subjects	41	11		
Error	99	33	3	
Total	194	47		

19. a. The null hypothesis states that there is no difference between the two treatments, H_0: $\mu_D = 0$. The critical region consists of t values beyond ± 3.182. The mean difference is $M_D = +4$. SS for the difference scores is 48, and $t(3) = 2.00$. Fail to reject H_0.

b. The null hypothesis states that there is no mean difference between treatments, H_0: $\mu_1 = \mu_2$. The critical value is $F = 10.13$.

Source	SS	df	MS	
Between treatments	32	1	32	$F(1, 3) = 4.00$
Within treatments	36	6		
Between subjects	12	3		
Error	24	3	8	
Total	68	7		

Fail to reject H_0. Note that $F = t^2$.

21. The means and standard deviations for the five delay periods are as follows:

1 month	6 months	1 year	2 years	5 years
$M = 866.67$	$M = 816.67$	$M = 766.67$	$M = 700.00$	$M = 583.33$
$S = 81.65$	$s = 81.65$	$s = 81.65$	$s = 70.71$	$s = 60.55$

Source	*SS*	*df*	*MS*	
Between treatments	291,333.3	4	72,833.3	$F(4, 20) = 56.75$
Within treatments	143,333.3	25		
Between subjects	117,666.7	5		
Error	25,666.7	20	1,283.3	
Total	434,666.7	29		

With $\alpha = .01$, the critical value is 4.43. There are significant differences.

CHAPTER 14: TWO-FACTOR ANOVA

1. a. In analysis of variance, an independent variable (or a quasi-independent variable) is called a *factor*.

b. The values of a factor that are used to create the different groups or treatment conditions are called the *levels* of the factor.

c. A research study with two independent (or quasi-independent) variables is called a *two-factor study*.

3. During the second stage of the two-factor ANOVA the mean differences between treatments are analyzed into differences from each of the two main effects and differences from the interaction.

5. a. $M = 10$

b. $M = 30$

c. $M = 50$

7. a. The scores in treatment 1 are consistently higher than the scores in treatment 2. There is a main effect for treatment.

b. The overall mean is around $M = 15$ for all three age groups. There is no main effect for age.

c. The two lines are not parallel. Instead, the difference between the treatments increases as the participants get older. Yes, there is an interaction.

9. a.

Source	*SS*	*df*	*MS*	
Between treatments	100	3		
A	10	1	10	$F(1,36) = 2$
B	90	1	90	$F(1,36) = 18.00$
A × B	0	1	0	$F(1,36) = 0$
Within treatments	180	36	5	
Total	280	39		

All F-ratios have $df = 1, 36$ and the critical value is $F = 4.11$. The main effect for factor B is significant, but factor A and the interaction are not.

b. For factor A, $\eta^2 = \frac{10}{190} = 0.053$, for factor B, $\eta^2 = \frac{90}{270} = 0.333$, and for the interaction, $\eta^2 = 0$.

11. a. $df = 1, 66$

b. $df = 2, 66$

c. $df = 2, 66$

13. a.

Source	*SS*	*df*	*MS*	
Between treatments	340	5		
Pouring	60	1	60	$F(1,54) = 10.00$
Temperature	280	2	140	$F(2,54) = 23.33$
Interaction	0	2	0	$F(2,42) = 0$
Within treatments	324	54	6	
Total	664	59		

b. Temperature and pouring method both have significant effects on the bubbles in the wine. However, the effects are independent, there is no interaction.

15.

Source	*SS*	*df*	*MS*	
Between treatments	144	8		
A	36	2	18	$F(2,72) = 6.00$
B	24	2	12	$F(2,72) = 4.00$
A × B	84	4	21	$F(4,72) = 7.00$
Within treatments	216	72	3	
Total	360	80		

17.

Source	SS	df	MS	
Between treatments	116	5		
A	28	1	28	$F(1,24) = 7.00$
B	64	2	32	$F(1,24) = 8.00$
A × B	24	2	12	$F(1,24) = 3.00$
Within treatments	240	60	4	
Total	356	65		

19. a.

Source	SS	df	MS	
Between treatments	360	5		
Gender	72	1	72	$F(1, 12) = 9.00$
Treatments	252	2	126	$F(1, 12) = 15.75$
Gender × treatment	36	2	18	$F(1, 12) = 2.25$
Within treatments	96	12	8	
Total	456	17		

With $df = 1, 12$, the critical value for the gender main effect is 4.75. The main effect for gender is significant. With $df = 2, 12$, the critical value for the treatment main effect and the interaction is 3.88. The main effect for treatments is significant but the interaction is not.

b. For treatment I, $F = 0$; for treatment II, $F = \frac{54}{8} = 6.75$; and for treatment III, $F = \frac{54}{8} = 6.75$. With $df = 1, 12$, the critical value for all three tests is 4.75. The results indicate a significant difference between males and females in treatments II and III, but not in treatment I.

21. a. The means for the six groups are as follows:

	Middle School	High School	College
Non-User	4.00	4.00	4.00
User	3.00	2.00	1.00

Source	SS	df	MS	
Between treatments	32	5		
Use	24	1	24	$F(1, 18) = 14.4$
School level	4	2	2	$F(2, 18) = 1.2$
Interaction	4	2	2	$F(2, 18) = 1.2$
Within treatments	30	18	1.67	
Total	62	23		

For $df = 1, 18$ the critical value is 4.41 and for $df = 2, 18$ it is 3.55. The main effect for Facebook use is significant but the other main effect and the interaction are not.

b. Grades are significantly lower for Facebook users. A difference exists for all three grade levels but appears to increase as the students get older although there is no significant interaction.

23. a.

Source	SS	df	MS	
Between treatments	216	3		
Self-esteem	96	1	96	$F(1, 20) = 22.33$
Audience	96	1	96	$F(1, 20) = 22.33$
Interaction	24	1	24	$F(1, 20) = 5.58$
Within treatments	86	20	4.3	
Total	300	23		

With $df = 1, 20$ the critical value is 4.35 for all three tests. Both main effects and the interaction are significant. Overall, there are fewer errors for the high self-esteem participants and for those working alone. The audience condition has very little effect on the high self-esteem participants but a very large effect on those with low self-esteem.

b. For both main effects, $\eta^2 = \frac{96}{182} = 0.527$. For the interaction, $\eta^2 = \frac{24}{110} = 0.218$.

SECTION IV **REVIEW**

1. a.

Source	SS	df	MS	
Between treatments	40	3	13.33	$F(3, 12) = 7.98$
Within treatments	20	12	1.67	
Total	60	15		

With $\alpha = .05$, the critical value is $F = 3.49$. Reject the null hypothesis.

b. $\eta^2 = \frac{40}{60} = 0.67$

c. The results show significant differences among the four levels of severity, $F(3, 12) = 7.98$, $p < .05$, $\eta^2 = 0.67$.

2. a. The null hypothesis states that there are no differences in quality of life among the three time periods.

Source	SS	df	MS	
Between treatments	56	2	28	$F(2, 6) = 10.49$
Within treatments	28	9		
Between subjects	12	3		
Error	16	6	2.67	
Total	84	11		

With $df = 2, 6$, the critical value is 5.14. Reject H_0.

b. For these data, $\eta^2 = \frac{56}{72} = 0.778$.

c. The results indicate significant changes in life satisfaction across the three time periods, $F(2, 6) = 10.49$, $p < .05$, $\eta^2 = 0.778$.

3. An interaction indicates that the effect of one factor depends on the levels of the other factor. Alternatively, it indicates that the main effects for one factor are not consistent across the levels of the other factor.

4.

Source	SS	df	MS	
Between treatments	148	3		
A (safety)	98	1	98	$F(1,28) = 13.52$
B (skill)	32	1	32	$F(1,28) = 4.41$
A × B	18	1	18	$F(1,28) = 2.48$
Within treatments	203	28	7.25	
Total	351	31		

All F-ratios have $df = 1, 28$ and the critical value is $F = 4.20$. With $\alpha = .05$ both main effects are significant but the interaction is not. Overall driving risk was significantly higher for those drivers who rated themselves as highly skilled compared to those with low ratings. Also, drivers rated low in safety had significantly more risk than those rated high.

CHAPTER 15: CORRELATION

1. A positive correlation indicates that X and Y change in the same direction: As X increases, Y also increases. A negative correlation indicates that X and Y tend to change in opposite directions: As X increases, Y decreases.

3. $SP = 15$

5. a. The scatter plot shows points widely scattered around a line sloping up to the right.

b. The correlation is small but positive; around 0.4 to 0.6.

c. For these scores, $SS_X = 32$, $SS_Y = 8$, and $SP = 8$. The correlation is $r = \frac{8}{16} = 0.50$.

7. a. The scatter plot shows points moderately scattered around a line sloping down to the right.

b. $SS_X = 10$, $SS_Y = 40$, and $SP = -13$. The correlation is $r = -\frac{13}{20} = -0.65$.

9. a. The scatter plot shows points clustered around a line sloping up to the right.

b. $SS_X = 18$, $SS_Y = 18$, and $SP = 5$. The correlation is $r = \frac{5}{18} = 0.278$.

11. a. For the children, $SS = 32$ and for the birth parents, $SS = 14$. $SP = 15$. The correlation is $r = 0.709$.

b. For the children, $SS = 32$ and for the adoptive parents $SS = 16$. $SP = 3$. The correlation is $r = 0.133$.

c. The children's behavior is strongly related to their birth parents and only weakly related to their adoptive parents. The data suggest that the behavior is inherited rather than learned.

13. a. For the men's weights, $SS = 18$ and for their incomes, $SS = 13{,}060$. $SP = 281$. The correlation is $r = 0.580$.

b. With $n = 8$, $df = 6$ and the critical value is 0.707. The correlation is not significant.

15. a. $r = 0.707$

b. $r = 0.468$

c. $r = 0.374$

17. a. $r_{XY\text{-}Z} = \frac{0.38}{0.57} = 0.667$

b. $r_{XZ\text{-}Y} = \frac{0.04}{0.428} = 0.093$

19. a. $r_S = +0.907$

b. With $n = 11$, the critical value is 0.618. The correlation is significant.

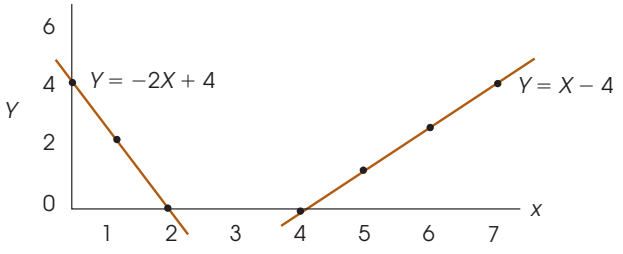
21. a. $r_S = -0.985$

b. For $n = 10$, the critical values are 0.648 and 0.794 for 5 and 0.1, respectively. The correlation is significant at either level.

23. Using the eating concern scores as the X variable and coding males as 1 and females as 0 for the Y variable produces $SS_X = 1875.6$, $SS_Y = 3.6$, and $SP = -50.4$. The point-biserial correlation is $r = -0.613$. (Reversing the codes for males and females will change the sign of the correlation.)

CHAPTER 16: INTRODUCTION TO REGRESSION

1.



3. a. $r = 0.80$
b. $\hat{Y} = 2X + 8$

5. The standard error of estimate is a measure of the average distance between the predicted Y points from the regression equation and the actual Y points in the data.

7. $SS_X = 32$, $SS_Y = 8$, $SP = 8$. The regression equation is $\hat{Y} = X + 3$

9. $SS_{regression} = r^2SS_Y = 90.02$ with $df = 1$. $MS_{residual} = \frac{18}{1} = 4.5$. $F = \frac{90.02}{4.5} = 20.00$. With $df = 1, 4$, the F-ratio is significant with $\alpha = .05$.

11. a. $SS_{weight} = 20$, $SS_{income} = 7430$, $SP = -359$. $\hat{Y} = -17.95X + 119.85$
b. $r = -0.931$ and $r^2 = 0.867$
c. $F = 52.15$ with $df = 1, 8$. The regression equation is significant with $\alpha = .05$ or $\alpha = .01$.

13. a. $\hat{Y} = 1.38X + 7.34$
b. $r^2 = 0.743$ or 74.3%
c. $F = 20.23$ with $df = 1, 7$. The equation accounts for a significant portion of the variance.

15. a. The standard error of estimate is $\sqrt{36/16} = 1.50$.
b. The standard error of estimate is $\sqrt{36/36} = 1.00$.

17. a. $df = 1, 23$
b. $n = 20$ pairs of scores

19. a. $F = \frac{2.2}{1.04} = 2.11$. With $df = 2, 15$, the critical value is 3.68. The equation does not account for a significant portion of the variance.
b. $F = \frac{2.2}{3.12} = 0.705$. The equation is not significant.

21. a. $SS_{churches} = 390$, $SS_{population} = 10$, and $SS_{crime} = 390$. SP for churches and population is 60, SP for churches and crime is 363, and SP for population and crime is 60. The regression equation is. $\hat{Y} = 0.1X_1 + 5.4X_2 - 2.7$.
b. $R^2 = 0.924$ or 92.4%
c. Population by itself predicts 92.4% of the variance. Nothing is gained by adding churches as a second variable.

23. Using the biological parents as a single predictor accounts for $r^2 = 0.503$ or 50.3% of the variance. The multiple regression equation accounts for $R^2 = 58.7\%$ (see problem 22). The extra variance predicted by adding the adoptive parents as a second predictor is $58.7 - 50.3 = 8.4\%$ and has $df = 1$. The residual from the multiple regression is $1 - R^2 = 41.3\%$ and has $df = 6$. The F-ratio is 8.4/(41.3/6) = 1.22. With $df = 1, 6$ the F-ratio is not significant.

25. $n = 39$

CHAPTER 17: CHI-SQUARE TESTS

1. Nonparametric tests make few if any assumptions about the populations from which the data are obtained. For example, the populations do not need to form normal distributions, nor is it required that different populations in the same study have equal variances (homogeneity of variance assumption). Parametric tests require data measured on an interval or ratio scale. For nonparametric tests, any scale of measurement is acceptable.

3. a. The null hypothesis states that there is no preference among the four colors; $p = \frac{1}{4}$ for all categories. The expected frequencies are $f_e = 15$ for all categories, and chi-square = 4.53. With $df = 3$, the critical value is 7.81. Fail to reject H_0 and conclude that there are no significant preferences.
b. The results indicate that there are no significant preferences among the four colors, $\chi^2(3, N = 60) = 4.53$, $p > .05$.

5. The null hypothesis states that wins and loses are equally likely. With 64 games, the expected frequencies are 32 wins and 32 losses. With $df = 1$ the critical value is 3.84, and the data produce a chi-square of 6.25. Reject the null hypothesis and conclude that home team wins are significantly more common than would be expected by chance.

7. a. The null hypothesis states that couples with the same initial do not occur more often than would be expected by chance. For a sample of 400, the expected frequencies are 26 with the same initial and 374 with different initials. With $df = 1$ the

critical value is 3.84, and the data produce a chi-square of 5.92. Reject the null hypothesis.

b. A larger sample should be more representative of the population. If the sample continues to be different from the hypothesis as the sample size increases, eventually the difference will be significant.

9. a. H_0 states that the distribution of automobile accidents is the same as the distribution of registered drivers: 16% under age 20, 28% age 20 to 29, and 56% age 30 or older. With $df = 2$, the critical value is 5.99. The expected frequencies for these three categories are 48, 84, and 168. Chi-square = 13.76. Reject H_0 and conclude that the distribution of automobile accidents is not identical to the distribution of registered drivers.

b. The chi-square test shows that the age distribution for people in automobile accidents is significantly different from the age distribution of licensed drivers, $\chi^2(3, N = 180) = 13.76$, $p < .05$.

11. The null hypothesis states that there are no preferences among the three designs; $p = \frac{1}{3}$ for all categories. With $df = 2$, the critical value is 5.99. The expected frequencies are $f_e = 40$ for all categories, and chi-square = 8.60. Reject H_0 and conclude that there are significant preferences.

13. The null hypothesis states that there is no relationship between the type of music and whether the women give their phone numbers. With $df = 1$, the critical value is 3.84. The expected frequencies are:

	Phone Number	No Number	
Romantic Music	15	25	40
Neutral Music	15	25	40
	30	50	

Chi-square = 7.68. Reject H_0.

15. a. The null hypothesis states that the distribution of opinions is the same for those who live in the city and those who live in the suburbs. For $df = 1$ and $\alpha = .05$, the critical value for chi-square is 3.84. The expected frequencies are:

	Favor	Oppose
City	30	20
Suburb	60	40

For these data, chi-square = 3.12. Fail to reject H_0 and conclude that opinions in the city are not different from those in the suburbs.

b. The phi coefficient is 0.144.

17. a. The null hypothesis states that the proportion who falsely recall seeing broken glass should be the same for all three groups. The expected frequency of saying yes is 9.67 for all groups, and the expected frequency for saying no is 40.33 for all groups. With $df = 2$, the critical value is 5.99. For these data, chi-square = 7.78. Reject the null hypothesis and conclude that the likelihood of recalling broken glass depends on the question that the participants were asked.

b. Cramérs $V = 0.228$.

c. Participants who were asked about the speed of the cars that "smashed into" each other were more than two times more likely to falsely recall seeing broken glass.

d. The results of the chi-square test indicate that the phrasing of the question had a significant effect on the participants' recall of the accident, $\chi^2(2, N = 150) = 7.78$, $p < .05$, $V = 0.228$.

19. The null hypothesis states that IQ and gender are independent. The distribution of IQ scores for boys should be the same as the distribution for girls. With $df = 2$ and and $\alpha = .05$, the critical value is 5.99. The expected frequencies are 15 low IQ, 48 medium, and 17 high for both boys and girls. For these data, chi-square is 3.76. Fail to reject the null hypothesis. These data do not provide evidence for a significant relationship between IQ and gender.

21. The null hypothesis states that there is no difference between the distribution of preferences predicted by women and the actual distribution for men. With $df = 3$ and $\alpha = .05$, the critical value is 7.81. The expected frequencies are:

	Somewhat Thin	Slightly Thin	Slightly Heavy	Somewhat Heavy
Women	22.9	22.9	22.9	11.4
Men	17.1	17.1	17.1	8.6

Chi-square = 9.13. Reject H_0 and conclude that there is a significant difference in the preferences predicted by women and the actual preferences expressed by men.

23. a. The null hypothesis states that there is no relationship between IQ and volunteering. With $df = 2$ and $\alpha = .05$, the critical value is 5.99. The expected frequencies are:

	High IQ	Medium IQ	Low IQ
Volunteer	37.5	75	37.5
Not Volunteer	12.5	25	12.5

The chi-square statistic is 4.75. Fail to reject H_0 with $\alpha = .05$ and $df = 2$.

25. The null hypothesis states that there is no relationship between the season of birth and schizophrenia. With $df = 3$ and $\alpha = .05$, the critical value is 7.81. The expected frequencies are:

	Summer	Fall	Winter	Spring
No Disorder	23.33	23.33	26.67	26.67
Schizophrenia	11.67	11.67	13.33	13.33

Chi-square = 3.62. Fail to reject H_0 and conclude that these data do not provide enough evidence to conclude that there is a significant relationship between the season of birth and schizophrenia.

CHAPTER 18: THE BINOMIAL TEST

1. H_0: p(home team win) = .50 (no preference). The critical boundaries are $z = \pm 1.96$. With $X = 42$, $\mu = 32$, and $\sigma = 4$, we obtain $z = 2.50$. Reject H_0 and conclude that there is a significant difference. Home teams win significantly more than would be expected by chance.

3. H_0: $p = q = \frac{1}{2}$ (right and left are equally common). The critical boundaries are $z = \pm 2.58$. With $X = 104$, $\mu = 72$, and $\sigma = 6$, we obtain $z = 5.33$. Reject H_0 and conclude that right- and left-handed rats are not equally common.

5. H_0: $p = 0.065$ (just chance). The critical boundaries are $z = \pm 1.96$. With $X = 38$, $\mu = 26$, and $\sigma = 4.93$, we obtain $z = 2.43$. Reject H_0. The initials of spouses are significantly different from what would be expected by chance.

7. H_0: $p = .08$ (still 8% learning disabled). The critical boundaries are $z = \pm 1.96$. With $X = 42$, $\mu = 24$, and $\sigma = 4.70$, we obtain $z = 3.83$. Reject H_0 and conclude that there has been a significant change in the proportion of students classified as learning disabled.

9. H_0: $p = .25$ (the general population has the same proportion of belief as the psychotherapists). The critical boundaries are $z = \pm 1.96$. With $X = 65$, $\mu = 48$, and $\sigma = 6$, we obtain $z = 2.83$. Reject H_0 and conclude that the proportion of belief is significantly different for the general population and for psychotherapists.

11. H_0: p(accident) = 0.12 (no change). The critical boundaries are $z = \pm 1.96$. With $X = 44$, $\mu = 60$, and $\sigma = 7.27$, we obtain $z = -2.20$. Reject H_0, there has been a significant change in the accident rate.

13. H_0: $p = \frac{1}{4} = p$(guessing correctly). The critical boundaries are $z = \pm 1.96$. With $X = 32$, $\mu = 25$, and $\sigma = 4.33$, we obtain $z = 1.62$. Fail to reject H_0 and conclude that this level of performance is not significantly different from chance.

15. a. H_0: $p = q = \frac{1}{2}$ (positive and negative correlations are equally likely). The critical boundaries are $z = \pm 1.96$. With $X = 25$, $\mu = 13.5$, and $\sigma = 2.60$, we obtain $z = 4.42$. Reject H_0, positive and negative correlations are not equally likely.

b. The critical boundaries are $z = \pm 1.96$. With $X = 20$, $\mu = 13.5$, and $\sigma = 2.60$, we obtain $z = 2.50$. Reject H_0, positive and negative correlations are not equally likely.

17. H_0: $p = .30$ and $q = .70$ (proportions for the special program are the same as in the population). The critical boundaries are $z = \pm 1.96$. The binomial distribution has $\mu = 27$ and $\sigma = 4.35$. With $X = 43$ we obtain $z = 3.68$. Reject H_0 and conclude that there is a significant difference between special program students and the general population.

19. H_0: $p = \frac{1}{2} = p$(reduced reactions). The critical boundaries are $z = \pm 1.96$. The binomial distribution has $\mu = 32$ and $\sigma = 4$. With $X = 47$ we obtain $z = 3.75$. Reject H_0 and conclude that there is evidence of significantly reduced allergic reactions.

21. H_0: $p = q = \frac{1}{2}$(higher and lower grades are equally likely). The critical boundaries are $z = \pm 2.58$. The binomial distribution has $\mu = 20$ and $\sigma = 3.16$. With $X = 29$ we obtain $z = 2.85$. Reject H_0 and conclude that there are significantly more higher grades than would be expected by chance.

23. a. H_0: $p = \frac{1}{3}$ and $q = \frac{2}{3}$ (just guessing). The critical boundaries are $z = \pm 1.96$. The binomial distribution has $\mu = 12$ and $\sigma = 2.83$. With $X = 25$ we obtain $z = 4.59$. Reject H_0 and conclude that the children with autism are performing significantly better than chance.

b. H_0: $p = \frac{1}{3}$ and $q = \frac{2}{3}$ (just guessing). The critical boundaries are $z = \pm 1.96$. The binomial distribution has $\mu = 12$ and $\sigma = 2.83$. With

$X = 16$ we obtain $z = 1.41$. Fail to reject H_0 and conclude that the children with SLI are not performing significantly better than chance.

25. H_0: $p = q = \frac{1}{2}$ (any change in grade point average is due to chance). The critical boundaries are $z = \pm 2.58$. The binomial distribution has $\mu = 22.5$ and $\sigma = 3.35$. With $X = 31$ we obtain $z = 2.54$. Fail to reject H_0 and conclude that there is no significant change in grade point average after the workshop.

27. a. H_0: $p = q = \frac{1}{2}$ (the training has no effect). The critical boundaries are $z = \pm 1.96$. Discarding the 11 people who showed no change, the binomial distribution has $\mu = 19.5$ and $\sigma = 3.12$. With $X = 29$ we obtain $z = 3.05$. Reject H_0 and conclude that biofeedback training has a significant effect.

b. Discarding only 1 participant and dividing the other 10 equally, the binomial distribution has $\mu = 24.5$ and $\sigma = 3.50$. With $X = 34$ we obtain $z = 2.71$. Reject H_0.

SECTION V **REVIEW**

1. a. $SS_X = 40$, $SS_Y = 5984$, $SP = 480$, and the Pearson correlation is $r = 0.981$.

b. The Spearman correlation is $r_S = 1.00$.

2. a. For these data, $SS_{\text{wife}} = 172$, $SS_{\text{husband}} = 106$, and $SP = 122$. The Pearson correlation is $r = 0.904$.

b. $b = \frac{122}{172} = 0.709$ and $a = 9 - 0.709(7) = 4.037$. $\hat{Y} = 0.709X + 4.037$.

3. The null hypothesis states that there is no preference among the three photographs; $p = \frac{1}{3}$ for all categories. The expected frequencies are $f_e = 50$ for all categories, and chi-square = 20.28. With $df = 2$, the critical value is 5.99. Reject H_0 and conclude that there are significant preferences

4. a. The null hypothesis states that there is no relationship between personality and heart disease. For $df = 1$ and $\alpha = .05$, the critical value for chi-square is 3.84. The expected frequencies are:

	No Disease	Heart Disease
Type A	40	10
Type B	120	30

For these data, chi-square = 10.67. Reject H_0 and conclude that there is a significant relationship between personality and heart disease.

b. $\varphi = 0.231$

5. H_0: $p = 0.20$ (correct predictions are just chance). The critical boundaries are $z = \pm 1.96$. With $X = 27$, $\mu = 20$, and $\sigma = 4$, we obtain $z = 1.75$. Fail to reject H_0 and conclude that this level of performance is not significantly better than chance.

CHAPTER 19: **CHOOSING THE RIGHT STATISTICS**

1. The mean and standard deviation could be used to describe the set of scores before treatment and the set of scores after treatment. Or a difference score could be computed for each participant and the results could be described with the mean and standard deviation for the set of difference scores. A repeated-measures t test would evaluate the significance of the mean difference and effect size would be measured by Cohen's d or r^2.

3. The data would form a 2×2 frequency distribution matrix and the proportion in each cell would describe the result. A chi-square test for independence would determine whether the proportions for the peer mentor group are significantly different from the proportions for other freshmen. Effect size would be measured with a phi-coefficient.

5. The mean and standard deviation could be used to describe the set of scores with texting and the set of scores without texting. Or a difference score could be computed for each participant and the results could be described with the mean and standard deviation for the set of difference scores. A repeated-measures t test would evaluate the significance of the mean difference and effect size would be measured by Cohen's d or r^2.

7. The mean and standard deviation could be used to describe the set of scores for each condition. An independent-measures t test would evaluate the significance of the mean difference, and effect size would be measured by Cohen's d or r^2.

9. The mean and standard deviation could be used to describe the set of scores for each condition. An independent-measures ANOVA would evaluate the significance of the mean differences and effect size would be measured by η^2.

11. The mean and standard deviation could be used to describe the set of scores for each of the four birth order groups. An independent-measures ANOVA would evaluate the significance of the mean differences and effect size would be measured by η^2.

13. The mean and standard deviation could be used to describe the set of scores for each of the four weight groups. An independent-measures ANOVA would evaluate the significance of the mean differences and effect size would be measured by η^2.

15. The mean and standard deviation rating could be used to describe the group. If the rating scale has a neutral point, a single-sample *t* test could be used to determine whether the mean optimism level is significantly different from neutral.

17. a. The data could be described by how the higher and lower ranks are clustered in the two groups. A Mann-Whitney test could determine whether there is a significant difference between the groups.

b. With two ordinal scores for each child, a Spearman correlation could measure and describe the relationship between variables. The significance of the correlation could be determined by comparing the sample value with the critical values listed in Table B7.

19. The mean and standard deviation could be used to describe the set of scores for each group. An independent-measures *t* test would evaluate the significance of the mean difference, and effect size would be measured by Cohen's *d* or r^2.

21. The proportion or percentage showing decreased pain could be used to describe the results. A binomial sign test would evaluate the significance of the treatment.

23. The mean and standard deviation could be used to describe the set of scores for each group. An independent-measures *t* test would evaluate the significance of the mean difference, and effect size would be measured by Cohen's *d* or r^2.

25. a. The mean and standard deviation could be used to describe the set of scores for each carrier. A repeated-measures ANOVA would evaluate the significance of the mean differences, and effect size would be measured by η^2.

b. The data would form a 3×3 frequency distribution matrix and the proportion in each cell would describe the results. A chi-square test for independence would determine whether the proportions of 1st-, 2nd-, and 3rd-place ratings are significantly different from one carrier to another. Effect size would be measured with Cramér's *V*. Alternatively, the data form three sets of scores, one for each phone. The scores are the rankings given by the participants. A Friedman test could evaluate the significance of the difference between carriers.

c. The three proportions would describe the relative preference for the carriers. A chi-square test for goodness of fit would determine whether there are significant preferences among the three carriers.

27. The mean and standard deviation could be used to describe the three deprivation conditions. A repeated-measures ANOVA would evaluate the significance of the mean differences, and effect size would be measured by η^2.

29. The data would form a 2×2 frequency distribution matrix, and the proportion in each cell would describe the results. A chi-square test for independence would determine whether the proportions using the stairs and elevators are significantly different from one condition to the other. Effect size would be measured with a phi-coefficient.

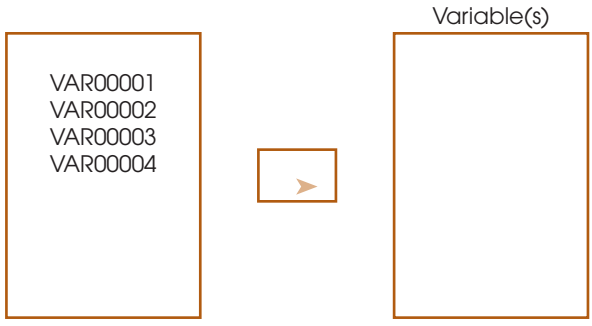
APPENDIX D

General Instructions for Using SPSS

The Statistical Package for the Social Sciences, commonly known as SPSS, is a computer program that performs statistical calculations, and is widely available on college campuses. Detailed instructions for using SPSS for specific statistical calculations (such as computing sample variance or performing an independent-measures *t* test) are presented at the end of the appropriate chapter in the text. Look for the SPSS logo in the Resources section at the end of each chapter. In this appendix, we provide a general overview of the SPSS program.

SPSS consists of two basic components: A data editor and a set of statistical commands. The **data editor** is a huge matrix of numbered rows and columns. To begin any analysis, you must type your data into the data editor. Typically, the scores are entered into columns of the editor. Before scores are entered, each of the columns is labeled "var." After scores are entered, the first column becomes VAR00001, the second column becomes VAR00002, and so on. To enter data into the editor, the **Data View** tab must be set at the bottom left of the screen. If you want to name a column (instead of using VAR00001), click on the **Variable View** tab at the bottom of the data editor. You will get a description of each variable in the editor, including a box for the name. You may type in a new name using up to 8 lowercase characters (no spaces, no hyphens). Click the **Data View** tab to go back to the data editor.

The **statistical commands** are listed in menus that are made available by clicking on **Analyze** in the tool bar at the top of the screen. When you select a statistical command, SPSS typically asks you to identify exactly where the scores are located and exactly what other options you want to use. This is accomplished by identifying the column(s) in the data editor that contain the needed information. Typically, you are presented with a display similar to the following figure. On the left is a box that lists all of the columns in the data editor that contain information. In this example, we have typed values into columns 1, 2, 3, and 4. On the right is an empty box that is waiting for you to identify the correct column. For example, suppose that you wanted to do a statistical calculation using the scores in column 3. You should highlight VAR00003 by clicking on it in the left-hand box, then click the arrow to move the column label into the right hand box. (If you make a mistake, you can highlight the variable in the right-hand box, which will reverse the arrow so that you can move the variable back to the left-hand box.)



SPSS DATA FORMATS

The SPSS program uses two basic formats for entering scores into the data matrix. Each is described and demonstrated as follows:

1. The first format is used when the data consist of several scores (more than one) for each individual. This includes data from a repeated-measures study, in which each person is measured in all of the different treatment conditions, and data from a correlational study where there are two scores, *X* and *Y,* for each individual. Table D1 illustrates this kind of data and shows how the scores would appear in the SPSS data matrix. Note that the scores in the data matrix have exactly the same structure as the scores in the original data. Specifically, each row of the data matrix contains the scores for an individual participant, and each column contains the scores for one treatment condition.

TABLE D1

Data for a repeated-measures or correlational study with several scores for each individual. The left half of the table (a) shows the original data, with three scores for each person; and the right half (b) shows the scores as they would be entered into the SPSS data matrix. Note: SPSS automatically adds the two decimal points for each score. For example, you type in 10 and it appears as 10.00 in the matrix.

(a) Original data

	Treatments		
Person	I	II	III
A	10	14	19
B	9	11	15
C	12	15	22
D	7	10	18
E	13	18	20

(b) Data as entered into the SPSS data matrix

	VAR0001	VAR0002	VAR0003	var
1	10.00	14.00	19.00	
2	9.00	11.00	15.00	
3	12.00	15.00	22.00	
4	7.00	10.00	18.00	
5	13.00	18.00	20.00	

2. The second format is used for data from an independent-measures study using a separate group of participants for each treatment condition. This kind of data is entered into the data matrix in a *stacked* format. Instead of having the scores from different treatments in different columns, all of the scores from all of the treatment conditions are entered into a single column so that the scores from one treatment condition are literally stacked on top of the scores from another treatment condition. A code number is then entered into a second column beside each score to tell the computer which treatment condition corresponds to each score. For example, you could enter a value of 1 beside each score from treatment #1, enter a 2 beside each score from treatment #2, and so on. Table D2 illustrates this kind of data and shows how the scores would be entered into the SPSS data matrix.

TABLE D2

Data for an independent-measures study with a different group of participants in each treatment condition. The left half of the table shows the original data, with three separate groups, each with five participants, and the right half shows the scores as they would be entered into the SPSS data matrix. Note that the data matrix lists all 15 scores in the same column, then uses code numbers in a second column to indicate the treatment condition corresponding to each score.

(a) Original data

Treatments		
I	II	III
10	14	19
9	11	15
12	15	22
7	10	18
13	18	20

(b) Data as entered into the SPSS data matrix

	VAR0001	VAR0002	var
1	10.00	1.00	
2	9.00	1.00	
3	12.00	1.00	
4	7.00	1.00	
5	13.00	1.00	
6	14.00	2.00	
7	11.00	2.00	
8	15.00	2.00	
9	10.00	2.00	
10	18.00	2.00	
11	19.00	3.00	
12	15.00	3.00	
13	22.00	3.00	
14	18.00	3.00	
15	20.00	3.00	

This page intentionally left blank

APPENDIX E

Hypothesis Tests for Ordinal Data: Mann-Whitney, Wilcoxon, Kruskal-Wallis, and Friedman Tests

E.1 DATA FROM AN ORDINAL SCALE

Occasionally, a research study generates data that consist of measurements on an ordinal scale. Recall from Chapter 1 that an ordinal scale simply produces a *rank ordering* for the individuals being measured. For example, a kindergarten teacher may rank children in terms of their maturity, or a business manager may classify job applicants as outstanding, good, and average.

RANKING NUMERICAL SCORES

In addition to obtaining measurements from an ordinal scale, a researcher may begin with a set of numerical measurements and convert these scores into ranks. For example, if you had a listing of the actual heights for a group of individuals, you could arrange the numbers in order from greatest to least. This process converts data from an interval or a ratio scale into ordinal measurements. In Chapter 17 (page 593), we identify several reasons for converting numerical scores into nominal categories. These same reasons also provide justification for transforming scores into ranks. The following list should give you an idea of why there can be an advantage to using ranks instead of scores.

1. Ranks are simpler. If someone asks you how tall your sister is, you could reply with a specific numerical value, such as 5 feet $7\frac{3}{4}$ inches tall. Or you could answer, "She is a little taller than I am." For many situations, the relative answer would be better.
2. The original scores may violate some of the basic assumptions that underlie certain statistical procedures. For example, the *t* tests and ANOVA assume that the data come from normal distributions. Also, the independent-measures tests assume that the different populations all have the same variance (the homogeneity-of-variance assumption). If a researcher suspects that the data do not satisfy these assumptions, it may be safer to convert the scores to ranks and use a statistical technique designed for ranks.
3. The original scores may have unusually high variance. Variance is a major component of the standard error in the denominator of *t* statistics and the error term in the denominator of *F*-ratios. Thus, large variance can greatly reduce the likelihood that these parametric tests will find significant differences. Converting the scores to ranks essentially eliminates the variance. For example, 10 scores have ranks from 1 to 10 no matter how variable the original scores are.
4. Occasionally, an experiment produces an undetermined, or infinite, score. For example, a rat may show no sign of solving a particular maze after hundreds of trials. This animal has an infinite, or undetermined, score. Although there is no absolute score that can be assigned, you can say that this rat has the highest score for the sample and then rank the rest of the scores by their numerical values.

RANKING TIED SCORES

Whenever you are transforming numerical scores into ranks, you may find two or more scores that have exactly the same value. Because the scores are tied, the transformation process should produce ranks that are also tied. The procedure for converting tied scores into tied ranks was presented in Chapter 15 (page 539) when we introduced the Spearman correlation, and is repeated briefly here. First, you list the scores in order, including tied values. Second, you assign each position in the list a rank (1st, 2nd, and

so on). Finally, for any scores that are tied, you compute the mean of the tied ranks, and use the mean value as the final rank. The following set of scores demonstrates this process for a set of $n = 8$ scores.

Original scores:	3	4	4	7	9	9	9	12
Position ranks:	1	2	3	4	5	6	7	8
Final ranks:	1	2.5	2.5	4	6	6	6	8

STATISTICS FOR ORDINAL DATA

You should recall from Chapter 1 that ordinal values tell you only the direction from one score to another, but provide no information about the distance between scores. Thus, you know that first place is better than second or third, but you do not know how much better. Because the concept of *distance* is not well defined with ordinal data, it generally is considered unwise to use traditional statistics such as *t* tests and analysis of variance with scores consisting of ranks or ordered categories. Therefore, statisticians have developed special techniques that are designed specifically for use with ordinal data.

In this chapter we introduce four hypothesis-testing procedures that are used with ordinal data. Each of the new tests can be viewed as an alternative for a commonly used parametric test. The four tests and the situations in which they are used are as follows:

1. The Mann-Whitney test uses data from two separate samples to evaluate the difference between two treatment conditions or two populations. The Mann-Whitney test can be viewed as an alternative to the independent-measures *t* hypothesis test introduced in Chapter 10.
2. The Wilcoxon test uses data from a repeated-measures design to evaluate the difference between two treatment conditions. This test is an alternative to the repeated-measures *t* test from Chapter 11.
3. The Kruskal-Wallis test uses data from three or more separate samples to evaluate the differences between three or more treatment conditions (or populations). The Kruskal-Wallis test is an alternative to the single-factor, independent-measures ANOVA introduced in Chapter 12.
4. The Friedman test uses data from a repeated-measures design to compare the differences between three or more treatment conditions. This test is an alternative to the repeated-measures ANOVA from Chapter 13.

In each case, you should realize that the new, ordinal-data tests are back-up procedures that are available in situations in which the standard, parametric tests cannot be used. In general, if the data are appropriate for an ANOVA or one of the *t* tests, then the standard test is preferred to its ordinal-data alternative.

E.2 THE MANN-WHITNEY *U*-TEST: AN ALTERNATIVE TO THE INDEPENDENT-MEASURES *t* TEST

Recall that a study using two separate samples is called an *independent-measures* study or a *between-subjects* study. The Mann-Whitney test is designed to use the data from two separate samples to evaluate the difference between two treatments (or two populations). The calculations for this test require that the individual scores in the

two samples be rank-ordered. The mathematics of the Mann-Whitney test are based on the following simple observation:

> A real difference between the two treatments should cause the scores in one sample to be generally larger than the scores in the other sample. If the two samples are combined and all the scores are ranked, then the larger ranks should be concentrated in one sample and the smaller ranks should be concentrated in the other sample.

THE NULL HYPOTHESIS FOR THE MANN-WHITNEY TEST

Because the Mann-Whitney test compares two distributions (rather than two means), the hypotheses tend to be somewhat vague. We state the hypotheses in terms of a consistent, systematic difference between the two treatments being compared.

H_0: There is no difference between the two treatments. Therefore, there is no tendency for the ranks in one treatment condition to be systematically higher (or lower) than the ranks in the other treatment condition.

H_1: There is a difference between the two treatments. Therefore, the ranks in one treatment condition are systematically higher (or lower) than the ranks in the other treatment condition.

CALCULATING THE MANN-WHITNEY U

Again, we begin by combining all the individuals from the two samples and then rank ordering the entire set. If you have a numerical score for each individual, combine the two sets of scores and rank order them. The Mann-Whitney U is then computed as if the two samples were two teams of athletes competing in a sports event. Each individual in sample A (the A team) gets one point whenever he or she is ranked ahead of an individual from sample B. The total number of points accumulated for sample A is called U_A. In the same way, a U value, or team total, is computed for sample B. The final Mann-Whitney U is the smaller of these two values. This process is demonstrated in the following example.

EXAMPLE E.1

We begin with two separate samples with $n = 6$ scores in each.

Sample A (treatment 1):	27	2	9	48	6	15
Sample B (treatment 2):	71	63	18	68	94	8

Next, the two samples are combined, and all 12 scores are ranked.

Ranks for Sample A	7	1	4	8	2	5
Ranks for Sample B	11	9	6	10	12	3

Each individual in sample A is assigned 1 point for every score in sample B that has a higher rank. $U_A = 4 + 6 + 5 + 4 + 6 + 5 = 30$. Similarly, $U_B = 0 + 0 + 2 + 0 + 0 + 4 = 6$.

Thus, the Mann-Whitney U is 6. As a simple check on your arithmetic, note that $U_A + U_B = n_A n_B$. For these data, $30 + 6 = 6(6)$.

COMPUTING *U* FOR LARGE SAMPLES

Because the process of counting points to determine the Mann-Whitney U can be tedious, especially with large samples, there is a formula that will generate the U value for each sample. To use this formula, you combine the samples and rank-order all the individuals as before. Then you must find ΣR_A, which is the sum of the ranks for

individuals in sample A, and the corresponding ΣR_B for sample B. The U value for each sample is then computed as follows: For sample A,

$$U_A = n_A n_B + \frac{n_A(n_A + 1)}{2} - \Sigma R_A$$

and for sample B,

$$U_B = n_A n_B + \frac{n_B(n_B + 1)}{2} - \Sigma R_B$$

These formulas are demonstrated using the data from Example E.1. For sample A, the sum of the ranks is

$$\Sigma R_A = 1 + 2 + 4 + 5 + 7 + 8 = 27$$

For sample B, the sum of the ranks is

$$\Sigma R_B = 3 + 6 + 9 + 10 + 11 + 12 = 51$$

By using the special formula, for sample A,

$$\begin{aligned} U_A &= n_A n_B + \frac{n_A(n_A + 1)}{2} - \Sigma R_A \\ &= 6(6) + \frac{6(7)}{2} - 27 \\ &= 36 + 21 - 27 \\ &= 30 \end{aligned}$$

For sample B,

$$\begin{aligned} U_B &= n_A n_B + \frac{n_B(n_B + 1)}{2} - \Sigma R_B \\ &= 6(6) + \frac{6(7)}{2} - 51 \\ &= 36 + 21 - 51 \\ &= 6 \end{aligned}$$

Notice that these are the same U values we obtained in Example E.1 using the counting method. The Mann-Whitney U value is the smaller of these two, $U = 6$

EVALUATING THE SIGNIFICANCE OF THE MANN-WHITNEY *U*

Table B9 in Appendix B lists critical value of U for $\alpha = .05$ and $\alpha = .01$. The null hypothesis is rejected when the sample data produce a U that is *less than or equal to* the table value.

In Example E.1 both samples have $n = 6$ scores, and the table shows a critical value of $U = 5$ for a two-tailed test with $\alpha = .05$. This means that a value of $U = 5$ or smaller is very unlikely to occur (probability less than .05) if the null hypothesis is true. The data actually produced $U = 6$, which is not in the critical region. Therefore, we fail to reject H_0 because the data do not provide enough evidence to conclude that there is a significant difference between the two treatments.

There are no strict rules for reporting the outcome of a Mann-Whitney U-test. However, APA guidelines suggest that the report include a summary of the data (including such information as the sample size and the sum of the ranks) and the

obtained statistic and p value. For the study presented in Example E.1, the results could be reported as follows:

> The original scores were ranked ordered and a Mann-Whitney U-test was used to compare the ranks for the $n = 6$ participants in treatment A and the $n = 6$ participants in treatment B. The results indicate no significant difference between treatments, $U = 6$, $p > .05$, with the sum of the ranks equal to 27 for treatment A and 51 for treatment B.

NORMAL APPROXIMATION FOR THE MANN-WHITNEY *U*

When the two samples are both large (about $n = 20$) and the null hypothesis is true, the distribution of the Mann-Whitney U statistic tends to approximate a normal shape. In this case, the Mann-Whitney hypotheses can be evaluated using a z-score statistic and the unit normal distribution. The procedure for this normal approximation is as follows:

1. Find the U values for sample A and sample B as before. The Mann-Whitney U is the smaller of these two values.
2. When both samples are relatively large (around $n = 20$ or more), the distribution of the Mann-Whitney U statistic tends to form a normal distribution with

$$\mu = \frac{n_A n_B}{2} \qquad \text{and} \qquad \sigma = \sqrt{\frac{n_A n_B (n_A + n_B + 1)}{12}}$$

   The Mann-Whitney U obtained from the sample data can be located in this distribution using a z-score:

$$z = \frac{X - \mu}{\sigma} = \frac{U - \dfrac{n_A n_B}{2}}{\sqrt{\dfrac{n_A n_B (n_A + n_B + 1)}{12}}}$$

3. Use the unit normal table to establish the critical region for this z-score. For example, with $\alpha = .05$, the critical values would be ± 1.96.

Usually the normal approximation is used with samples of $n = 20$ or larger; however, we will demonstrate the formulas with the data that were used in Example E.1. This study compared two treatments, A and B, using a separate sample of $n = 6$ for each treatment. The data produced a value of $U = 6$. The z-score corresponding to $U = 6$ is

$$z = \frac{U - \dfrac{n_A n_B}{2}}{\sqrt{\dfrac{n_A n_B (n_A + n_B + 1)}{12}}}$$

$$= \frac{6 - \dfrac{6(6)}{2}}{\sqrt{\dfrac{6(6)(6 + 6 + 1)}{12}}}$$

$$= \frac{-12}{\sqrt{\dfrac{468}{12}}}$$

$$= -1.92$$

With $\alpha = .05$, the critical value is $z = \pm 1.96$. Our computed z-score, $z = -1.92$, is not in the critical region, so the decision is to fail to reject H_0. Note that we reached the same conclusion for the original test using the critical values in the Mann-Whitney U table.

E.3 THE WILCOXON SIGNED-RANKS TEST: AN ALTERNATIVE TO THE REPEATED-MEASURES *t* TEST

The Wilcoxon test is designed to evaluate the difference between two treatments, using the data from a repeated-measures experiment. Recall that a repeated-measures study involves only one sample, with each individual in the sample being measured twice. The difference between the two measurements for each individual is recorded as the score for that individual. The Wilcoxon test requires that the differences be rank-ordered from smallest to largest in terms of their absolute magnitude, without regard for sign or direction. For example, Table E.1 shows differences scores and ranks for a sample of $n = 8$ participants.

TABLE E.1

Ranking difference scores. Note that the differences are ranked by magnitude, independent of direction.

Participant	Difference from Treatment 1 to Treatment 2	Rank
A	+4	1
B	−14	5
C	+5	2
D	−20	7
E	−6	3
F	−16	6
G	−8	4
H	−24	8

ZERO DIFFERENCES AND TIED SCORES

For the Wilcoxon test, there are two possibilities for tied scores:

1. A participant may have the same score in treatment 1 and in treatment 2, resulting in a difference score of zero.
2. Two (or more) participants may have identical difference scores (ignoring the sign of the difference).

When the data include individuals with difference scores of zero, one strategy is to discard these individuals from the analysis and reduce the sample size (n). However, this procedure ignores the fact that a difference score of zero is evidence for retaining the null hypothesis. A better procedure is to divide the zero differences evenly between the positives and negatives. (With an odd number of zero differences, discard one and divide the rest evenly.) When there are ties among the difference scores, each of the tied scores should be assigned the average of the tied ranks. This procedure was presented in detail in an earlier section of this appendix (see page 742).

HYPOTHESES FOR THE WILCOXON TEST

The null hypothesis for the Wilcoxon test simply states that there is no consistent, systematic difference between the two treatments.

If the null hypothesis is true, any differences that exist in the sample data must be due to chance. Therefore, we would expect positive and negative differences to be intermixed evenly. On the other hand, a consistent difference between the two treatments should cause the scores in one treatment to be consistently larger than the scores in

the other. This should produce difference scores that tend to be consistently positive or consistently negative. The Wilcoxon test uses the signs and the ranks of the difference scores to decide whether there is a significant difference between the two treatments.

H_0: There is no difference between the two treatments. Therefore, in the general population there is no tendency for the difference scores to be either systematically positive or systematically negative.

H_1: There is a difference between the two treatments. Therefore, in the general population the difference scores are systematically positive or systematically negative.

CALCULATION AND INTERPRETATION OF THE WILCOXON *T*

After ranking the absolute values of the difference scores, the ranks are separated into two groups: those associated with positive differences (increases) and those associated with negative differences (decreases). Next, the sum of the ranks is computed for each group. The smaller of the two sums is the test statistic for the Wilcoxon test and is identified by the letter T. For the difference scores in Table E.1, the positive differences have ranks of 1 and 2, which add to $\Sigma R = 3$ and the negative difference scores have ranks of 5, 7, 3, 6, 4, and 8, which add up to $\Sigma R = 33$. For these scores, $T = 3$.

Table B10 in Appendix B lists critical values of T for $\alpha = .05$ and $\alpha = .01$. The null hypothesis is rejected when the sample data produce a T that is *less than or equal to* the table value. With $n = 8$ and $\alpha = .05$ for a two-tailed test, the table lists a critical value of 3. For the data in Table E.1, we obtained $T = 3$ so we would reject the null hypothesis and conclude that there is a significant difference between the two treatments.

As with the Mann-Whitney *U*-test, there is no specified format for reporting the results of a Wilcoxon *T*-test. It is suggested, however, that the report include a summary of the data and the value obtained for the test statistic as well as the p value. If there are zero-difference scores in the data, it is also recommended that the report describe how they were treated. For the data in Table E.1, the report could be as follows:

> The eight participants were rank ordered by the magnitude of their difference scores and a Wilcoxon T was used to evaluate the significance of the difference between treatments. The results showed a significant difference, $T = 3$, $p < .05$, with the positive ranks totaling 3 and the negative ranks totaling 33.

NORMAL APPROXIMATION FOR THE WILCOXON *T*-TEST

When a sample is relatively large, the values for the Wilcoxon T statistic tend to form a normal distribution. In this situation, it is possible to perform the test using a z-score statistic and the normal distribution rather than looking up a T value in the Wilcoxon table. When the sample size is greater than 20, the normal approximation is very accurate and can be used. For samples larger than $n = 50$, the Wilcoxon table typically does not provide any critical values, so the normal approximation must be used. The procedure for the *normal approximation for the Wilcoxon* T is as follows:

1. Find the total for the positive ranks and the total for the negative ranks as before. The Wilcoxon T is the smaller of the two values.
2. With n greater than 20, the Wilcoxon T values form a normal distribution with a mean of

$$\mu = \frac{n(n + 1)}{4}$$

and a standard deviation of

$$\sigma = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$$

The Wilcoxon T from the sample data corresponds to a z-score in this distribution defined by

$$z = \frac{X - \mu}{\sigma} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n + 1)(2n + 1}{24}}}$$

3. The unit normal table is used to determine the critical region for the z-score. For example, the critical values are ±1.96 for $\alpha = .05$.

Although the normal approximation is intended for samples with at least $n = 20$ individuals, we will demonstrate the calculations with the data in Table E.1. These data have $n = 8$ and produced $T = 3$. Using the normal approximation, these values produce

$$\mu = \frac{n(n + 1)}{4} = \frac{8(9)}{4} = 18$$

$$\sigma = \sqrt{\frac{n(n +1)(2n + 1)}{24}} = \sqrt{\frac{8(9)(17)}{24}} = \sqrt{51} = 7.14$$

With these values, the obtained $T = 3$ corresponds to a z-score of

$$z = \frac{T - \mu}{\sigma} = \frac{3-18}{7.14} = \frac{-15}{7.14} = 2.10$$

With critical boundaries of ±1.96, the obtained z-score is close to the boundary, but it is enough to be significant at the .05 level. Note that this is exactly the same conclusion we reached using the Wilcoxon T table.

E.4 THE KRUSKAL-WALLIS TEST: AN ALTERNATIVE TO THE INDEPENDENT-MEASURES ANOVA

The *Kruskal-Wallis test* is used to evaluate differences among three or more treatment conditions (or populations) using ordinal data from an independent-measures design. You should recognize that this test is an alternative to the single-factor ANOVA introduced in Chapter 12. However, the ANOVA requires numerical scores that can be used to calculate means and variances. The Kruskal-Wallis test, on the other hand, simply requires that you are able to rank-order the individuals for the variable being measured. You also should recognize that the Kruskal-Wallis test is similar to the Mann-Whitney test introduced earlier in this chapter. However, the Mann-Whitney test is limited to

comparing only two treatments, whereas the Kruskal-Wallis test is used to compare three or more treatments.

THE DATA FOR A KRUSKAL-WALLIS TEST

The Kruskal-Wallis test requires three or more separate samples. The samples can represent different treatment conditions or they can represent different preexisting populations. For example, a researcher may want to examine how social input affects creativity. Children are asked to draw pictures under three different conditions: (1) working alone without supervision, (2) working in groups where the children are encouraged to examine and criticize each other's work, and (3) working alone but with frequent supervision and comments from a teacher. Three separate samples are used to represent the three treatment conditions with $n = 6$ children in each condition. At the end of the study, the researcher collects the drawings from all 18 children and rank-orders the complete set of drawings in terms of creativity. The purpose for the study is to determine whether one treatment condition produces drawings that are ranked consistently higher (or lower) than another condition. Notice that the researcher does not need to determine an absolute creativity score for each painting. Instead, the data consist of relative measures; that is, the researcher must decide which painting shows the most creativity, which shows the second most creativity, and so on.

The creativity study that was just described is an example of a research study comparing different treatment conditions. It also is possible that the three groups could be defined by a subject variable so that the three samples represent different populations. For example, a researcher could obtain drawings from a sample of 5-year-old children, a sample of 6-year-old children, and a third sample of 7-year-old children. Again, the Kruskal-Wallis test would begin by rank-ordering all of the drawings to determine whether one age group showed significantly more (or less) creativity than another.

Finally, the Kruskal-Wallis test can be used if the original, numerical data are converted into ordinal values. The following example demonstrates how a set of numerical scores is transformed into ranks to be used in a Kruskal-Wallis analysis.

EXAMPLE E.2

Table E.2(a) shows the original data from an independent-measures study comparing three treatment conditions. To prepare the data for a Kruskal-Wallis test, the complete set of original scores is rank-ordered using the standard procedure for ranking tied scores. Each of the original scores is then replaced by its rank to create the transformed data in Table E.2(b) that are used for the Kruskal-Wallis test.

THE NULL HYPOTHESIS FOR THE KRUSKAL-WALLIS TEST

As with the other tests for ordinal data, the null hypothesis for the Kruskal-Wallis test tends to be somewhat vague. In general, the null hypothesis states that there are no differences among the treatments being compared. Somewhat more specifically, H_0 states that there is no tendency for the ranks in one treatment condition to be systematically higher (or lower) than the ranks in any other condition. Generally, we use the concept of "systematic differences" to phrase the statement of H_0 and H_1. Thus, the hypotheses for the Kruskal-Wallis test are phrased as follows:

H_0: There is no tendency for the ranks in any treatment condition to be systematically higher or lower than the ranks in any other treatment condition. There are no differences between treatments.

H_1: The ranks in at least one treatment condition are systematically higher (or lower) than the ranks in another treatment condition. There are differences between treatments.

TABLE E.2

Preparing a set of data for analysis using the Kruskal-Wallis test. The original data consisting of numerical scores are shown in table (a). The original scores are combined into one group and rank ordered using the standard procedure for ranking tied scores. The ranks are then substituted for the original scores to create the set of ordinal data shown in table (b).

(a) Original Numerical Scores

I	II	III	
14	2	26	$N = 15$
3	14	8	
21	9	14	
5	12	19	
16	5	20	
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	

(b) Ordinal Data (Ranks)

I	II	III	
9	1	15	$N = 15$
2	9	5	
14	6	9	
3.5	7	12	
11	3.5	13	
$T_1 = 39.5$	$T_2 = 26.5$	$T_3 = 54$	
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	

Table E.2(b) presents the notation that is used in the Kruskal-Wallis formula along with the ranks. The notation is relatively simple and involves the following values:

1. The ranks in each treatment are added to obtain a total or T value for that treatment condition. The T values are used in the Kruskal-Wallis formula.
2. The number of subjects in each treatment condition is identified by a lowercase n.
3. The total number of subjects in the entire study is identified by an uppercase N.

The Kruskal-Wallis formula produces a statistic that is usually identified with the letter H and has approximately the same distribution as chi-square, with degrees of freedom defined by the number of treatment conditions minus one. For the data in Table E.2(b), there are 3 treatment conditions, so the formula produces a chi-square value with $df = 2$. The formula for the Kruskal-Wallis statistic is

$$H = \frac{12}{N(N+1)}\left(\Sigma\frac{T^2}{n}\right) - 3(N+1)$$

Using the data in Table E.2(b), the Kruskal-Wallis formula produces a chi-square value of

$$H = \frac{12}{15(16)}\left(\frac{39.5^2}{5} + \frac{26.5^2}{5} + \frac{54^2}{5}\right) - 3(16)$$

$$= 0.05(312.05 + 140.45 + 583.2) - 48$$

$$= 0.05(1035.7) - 48$$

$$= 51.785 - 48$$

$$= 3.785$$

With $df = 2$, the chi-square table lists a critical value of 5.99 for $\alpha = .05$. Because the obtained chi-square value (3.785) is not greater than the critical value, our statistical decision is to fail to reject H_0. The data do not provide sufficient evidence to conclude that there are significant differences among the three treatments.

As with the Mann-Whitney and the Wilcoxon tests, there is no standard format for reporting the outcome of a Kruskal-Wallis test. However, the report should provide a summary of the data, the value obtained for the chi-square statistic, as well as the value of *df, N,* and the *p* value. For the Kruskal-Wallis test that we just completed, the results could be reported as follows:

> After ranking the individual scores, a Kruskal-Wallis test was used to evaluate differences among the three treatments. The outcome of the test indicated no significant differences among the treatment conditions, $H = 3.785$ (2, $N = 15$), $p > .05$.

There is one assumption for the Kruskal-Wallis test that is necessary to justify using the chi-square distribution to identify critical values for *H*. Specifically, each of the treatment conditions must contain at least five scores.

E.5 THE FRIEDMAN TEST: AN ALTERNATIVE TO THE REPEATED-MEASURES ANOVA

The *Friedman test* is used to evaluate the differences between three or more treatment conditions using data from a repeated-measures design. This test is an alternative to the repeated-measures ANOVA that was introduced in Chapter 13. However, the ANOVA requires numerical scores that can be used to compute means and variances. The Friedman test simply requires ordinal data. The Friedman test is also similar to the Wilcoxon test that was introduced earlier in this chapter. However, the Wilcoxon test is limited to comparing only two treatments, whereas the Friedman test is used to compare three or more treatments.

THE DATA FOR A FRIEDMAN TEST

The Friedman test requires only one sample, with each individual participating in all of the different treatment conditions. The treatment conditions must be rank-ordered for each individual participant. For example, a researcher could observe a group of children diagnosed with ADHD in three different environments: at home, at school, and during unstructured play time. For each child, the researcher observes the degree to which the disorder interferes with normal activity in each environment, and then ranks the three environments from most disruptive to least disruptive. In this case, the ranks are obtained by comparing the individual's behavior across the three conditions. It is also possible for each individual to produce his or her own rankings. For example, each individual could be asked to evaluate three different designs for a new smart phone. Each person practices with each phone and then ranks them, 1st, 2nd, and 3rd in terms of ease of use.

Finally, the Friedman test can be used if the original data consist of numerical scores. However, the scores must be converted to ranks before the Friedman test is used. The following example demonstrates how a set of numerical scores is transformed into ranks for the Friedman test.

EXAMPLE E.3

To demonstrate the Friedman test, we will use the same data that were used to introduce the repeated-measures ANOVA in Chapter 13 (see page 440–441). The data are reproduced in Table E.3(a) and consist of ratings of four television-viewing distances for a sample of $n = 5$ participants. To convert the data for the Friedman test, the four scores for each participant are replaced with ranks 1, 2, 3, and 4, corresponding to the size of the original scores. As usual, tied scores are assigned the mean of the tied ranks. The complete set of ranks is shown in part (b) of the table.

THE HYPOTHESES FOR THE FRIEDMAN TEST

In general terms, the null hypothesis for the Friedman test states that there are no differences between the treatment conditions being compared so the ranks in one treatment should not be systematically higher (or lower) than the ranks in any other treatment condition. Thus, the hypotheses for the Friedman test can be phrased as follows:

H_0: There is no difference between treatments. Thus, the ranks in one treatment condition are not systematically higher or lower than the ranks in any other treatment condition.

H_1: There are differences between treatments. Thus, the ranks in at least one treatment condition are systematically higher or lower than the ranks in another treatment condition.

NOTATION AND CALCULATION FOR THE FRIEDMAN TEST

The first step in the Friedman test is to compute the sum of the ranks for each treatment condition. The ΣR values are shown in Table E.3(b). In addition to the ΣR values, the calculations for the Friedman test require the number of individuals in the sample (n) and the number of treatment conditions (k). For the data in Table E.3(b), $n = 5$ and

TABLE E.3

Results from a repeated-measures study comparing four television-viewing distances. Part (a) shows the original rating score for each of the distances. In part (b), the four distances are rank-ordered according to the preferences for each participant. The original data appear in Table 13.2 (page 441) and were used to demonstrate the repeated-measures ANOVA.

(a) The original rating scores.

Person	9 Feet	12 Feet	15 Feet	18 Feet
A	3	4	7	6
B	0	3	6	3
C	2	1	5	4
D	0	1	4	3
E	0	1	3	4

(b) The ranks of the treatment conditions for each participant

Person	9 Feet	9 Feet	9 Feet	9 Feet
A	1	2	4	3
B	1	2.5	4	2.5
C	2	1	4	3
D	1	2	4	3
E	1	2	3	4
	$\Sigma R_1 = 6$	$\Sigma R_2 = 9.5$	$\Sigma R_3 = 19$	$\Sigma R_4 = 15.5$

$k = 4$. The Friedman test evaluates the differences between treatments by computing the following test statistic:

$$\chi_r^2 = \frac{12}{nk(k+1)} \Sigma R^2 - 3n(k+1)$$

Note that the statistic is identified as chi-square (χ^2) with a subscript *r,* and corresponds to a chi-square statistic for ranks. This chi-square statistic has degrees of freedom determined by $df = k - 1$, and is evaluated using the critical values in the chi-square distribution shown in Table B8 in Appendix B.

For the data in Table E.3(b), the statistic is

$$\begin{aligned}\chi_r^2 &= \frac{12}{5(4)(5)} (6^2 + 9.5^2 + 15.5^2 + 19^2) - 3(5)(5) \\ &= \frac{12}{100} (36 + 90.25 + 240.25 + 361) - 75 \\ &= 0.12(727.5) - 75 \\ &= 12.3\end{aligned}$$

With $df = k - 1 = 3$, the critical value of chi-square is 9.35. Therefore, the decision is to reject the null hypothesis and conclude that there are significant differences among the four treatment conditions.

As with most of the tests for ordinal data, there is no standard format for reporting the results from a Friedman test. However, the report should provide the value obtained for the chi-square statistic as well as the values for *df, n,* and *p*. For the data that were used to demonstrate the Friedman test, the results would be reported as follows:

> After ranking the original scores, a Friedman test was used to evaluate the differences among the four treatment conditions. The outcome indicated that there are significant differences, $\chi_r^2 = 12.3$ (3, $n = 5$), $p < .05$.

References

Note: Numbers in **bold** type indicate the textbook page(s) on which each reference is mentioned.

Ackerman, P. L., & Beier, M. E. (2007). Further explorations of perceptual speed abilities in the context of assessment methods, cognitive abilities, and individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 13,* 249–272. **(p. 138)**

Albentosa, M. J., & Cooper, J. J. (2005). Testing resource value in group-housed animals: An investigation of cage height preference in laying hens. *Behavioural Processes, 70,* 113–121. **(p. 652)**

American Psychological Association (APA). (2010). *Publication manual of the American Psychological Association* (6th ed.) Washington, DC: Author. **(pp. 82, 93, 466)**

Anderson, D. R., Huston, A. C., Wright, J. C., & Collins, P. A. (1998). Initial findings on the long term impact of Sesame Street and educational television for children: The recontact study. In R. Noll and M. Price (Eds.), *A communication cornucopia: Markle Foundation essays on information policy* (pp. 279–296). Washington, DC: Brookings Institution. **(p. 326)**

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences, 41,* 39–48. **(pp. 131, 630)**

Athos, E. A., Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., & Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. *Proceedings of the National Academy of Science of the United States of America, 104,* 14795–14800. **(p. 88)**

Bartus, R. T. (1990). Drugs to treat age-related neurodegenerative problems: The final frontier of medical science? *Journal of the American Geriatrics Society, 38,* 680–695. **(pp. 235, 376)**

Bélisle, J., & Bodur, H. O. (2010). Avatars as information: Perception of consumers based on the avatars in virtual worlds. *Psychology & Marketing, 27,* 741–765. **(p. 630)**

Belsky, J., Weinraub, M., Owen, M., & Kelly, J. (2001). Quality of child care and problem behavior. In J. Belsky (Chair), *Early childcare and children's development prior to school entry*. Symposium conducted at the 2001 Biennial Meetings of the Society for Research in Child Development, Minneapolis, MN. **(pp. 313, 674)**

Blest, A. D. (1957). The functions of eyespot patterns in the Lepidoptera. *Behaviour, 11,* 209–255. **(p. 284)**

Blum, J. (1978). *Pseudoscience and mental ability: The origins and fallacies of the IQ controversy*. New York: Monthly Review Press. **(p. 521)**

Boogert, N. J., Reader, S. M., & Laland, K. N. (2006). The relation between social rank, neophobia and individual learning in starlings. *Behavioural Biology, 72,* 1229–1239. **(p. 555)**

Bradbury, T. N., & Miller, G. A. (1985). Season of birth in schizophrenia: A review of evidence, methodology, and etiology. *Psychological Bulletin, 98,* 569–594. **(p. 631)**

Broberg, A. G., Wessels, H., Lamb, M. E., & Hwang, C. P. (1997). Effects of day care on the development of cognitive abilities in 8-year-olds: A longitudinal study. *Development Psychology, 33*, 62–69. **(p. 313)**

Brunt, A., Rhee, Y., & Zhong, L. (2008). Differences in dietary patterns among college students according to body mass index. *Journal of American College Health. 56,* 629–634. **(pp. 278, 674)**

Byrne, D. (1971). *The attraction paradigm*. New York: Academic Press. **(p. 654)**

Camera, W. J., & Echternacht, G. (2000). *The SAT 1 and high school grades: Utility in predicting success in college* (College Board Report No. RN-10). New York: College Entrance Examination Board. **(p. 558)**

Candappa, R. (2000). *The little book of wrong shui*. Kansas City: Andrews McMeel Publishing. **(p. 4)**

Cerveny, R. S., & Balling, Jr., R. C. (1998). Weekly cycles of air pollutants, precipitation and tropical cyclones in the coastal NW Atlantic region. *Nature, 394*, 561–563. **(p. 102)**

Chandra, A., & Minkovitz, C. S. (2006). Stigma starts early: Gender differences in teen willingness to use mental health services. *Journal of Adolescent Health, 38*, 754.el–754.e8. **(p. 610)**

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*, 1015–1026. **(p. 631)**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates. **(pp. 262, 263, 299)**

Cohn, E. J., & Rotton, J. (2000). Weather, disorderly conduct, and assaults: From social contact to social avoidance. *Environment and Behavior, 32,* 651–673. **(p. 535)**

Collins, R. L., Elliott, M. N., Berry, S. H., Kanouse, D. E., Kunkel, D., Hunter, S. B., & Miu, A. (2004). Watching sex on television predicts adolescent initiation of sexual behavior. *Pediatrics, 114*, e280–e289. **(pp. 326, 534)**

Cook, M. (1977). Gaze and mutual gaze in social encounters. *American Scientist, 65,* 328–333. **(p. 284)**

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37,* 553–558. **(p. 246)**

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684. **(p. 386)**

Danner F., & Phillips B. (2008). Adolescent sleep, school start times, and teen motor vehicle crashes. *Journal of Clinical Sleep Medicine, 4,* 533–535. **(p. 651)**

Downs, D. S., & Abwender, D. (2002). Neuropsychological impairment in soccer athletes. *Journal of Sports Medicine and Physical Fitness, 42*, 103–107. **(pp. 349, 500)**

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good but . . . : meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin,110,* 109–128. **(pp. 376, 675)**

Elliot, A. J., Niesta, K., Greitemeyer, T., Lichtenfeld, S., Gramzow, R., Maier, M. A., & Liu, H. (2010). Red, rank, and romance in women viewing men. *Journal of Experimental Psychology: General, 139,* 399–417. **(p. 503)**

Elliot, A. J., & Niesta, D. (2008). Romantic red: Red enhances men's attraction to women. *Journal of Personality and Social Psychology, 95*, 1150–1164. **(pp. 359, 503)**

Fallon, A. E., & Rozin, P. (1985). Sex differences in perceptions of desirable body shape. *Journal of Abnormal Psychology*, *94,* 102–105. **(p. 630)**

Flett, G. L., Goldstein, A., Wall, A., Hewitt, P. L., Wekerle, C., and Azzi, N. (2008). Perfectionism and binge drinking in Canadian students making the transition to university. *Journal of American College Health*, *57,* 249–253. **(p. 429)**

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin. 95,* 29–51. **(p. 312)**

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20. **(pp. 312, 675)**

Ford, A. M., & Torok, D. (2008). Motivational signage increases physical activity on a college campus. *Journal of American College Health*, *57,* 242–244. **(pp. 35, 675)**

Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. *British Medical Journal (Clinical Research ed.), 337,* pp a2338 (electronic publication). **(p. 68)**

Friedman, M., & Rosenman, R. H. (1974). *Type A behavior and your heart*. New York: Knopf. **(p. 655)**

Frieswijk, N., Buunk, B. P., Steverink, N., & Slaets, J. P. J. (2004). The effect of social comparison information on the life satisfaction of frail older persons. *Psychology and Aging, 19,* 183–190. **(p. 380)**

Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., and Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology, 100,* 491–509. **(p. 502)**

Fung, C. H., Elliott, M. N., Hays, R. D., Kahn, K. L., Kanouse, D. E., McGlynn, E. A., Spranca, M. D., & Shekelle, P. G. (2005). Patient's preferences for technical versus interpersonal quality when selecting a primary care physician. *Health Services Research, 40,* 957–977. **(p. 651)**

Gibson, E. J., & Walk, R. D. (1960). The "visual cliff." *Scientific American, 202,* 64–71. **(p. 634)**

Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology, 78,* 211–222. **(p. 311)**

Gintzler, A. R. (1980). Endorphin-mediated increases in pain threshold during pregnancy. *Science, 210,* 193–195. **(p. 463)**

Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A lifespan comparison. *Psychological Science, 5,* 33–36. **(p. 434)**

International Journal of Neuroscience, 113, 1675–1689.

Guéguen, N., & Jacob, C. (2010). 'Love is in the air': Effects of songs with romantic lyrics on compliance with a courtship request. *Psychology of Music, 38,* 303–307. **(p. 628)**

Güven, M., Elaimis, D. D., Binokay, S., & Tan, O. (2003). Population-level right-paw preference in rats assessed by a new computerized food-reaching test. **(p. 650)**

Hallam, S., Price, J., & Katsarou, G. (2002). The effects of background music on primary school pupils, task

performance. *Educational Studies, 28,* 111–122. **(pp. 347, 674)**

Harlow, H. F. (1959). Love in infant monkeys. *Scientific American, 200,* 68–86. **(p. 312)**

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test (WCST) manual: Revised and expanded.* Odessa, FL: Psychological Assessment Resources. **(p. 235)**

Hill, R. A., & Barton, R. A. (2005). Red enhances human performance in contests. *Nature, 435,* 293. **(pp. 628, 650)**

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7. **(p. 259)**

Igou, E. R. (2008). 'How long will I suffer?' versus 'How long will you suffer?' A self-other effect in affective forecasting. *Journal of Personality and Social Psychology, 95,* 899–917. **(p. 376)**

Ijuin, M., Homma, A., Mimura, M., Kitamura, S., Kawai, Y., Imai, Y., & Gondo, Y. (2008). Validation of the 7-minute screen for the detection of early-stage Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 25,* 248–255. **(p. 554)**

Jackson, E. M., & Howton, A. (2008). Increasing walking in college students using a pedometer intervention: Differences according to body mass index. *Journal of American College Health*, *57*, 159–164. **(p. 463)**

Johnston, J. J. (1975). Sticking with first responses on multiple-choice exams: For better or worse? *Teaching of Psychology, 2,* 178–179. **(pp. 378, 675)**

Jones, B. T, Jones, B. C. Thomas, A. P., & Piper, J. (2003). Alcohol consumption increases attractiveness ratings of opposite sex faces: a possible third route to risky sex. *Addiction, 98,* 1069–1075. Doi: 10.1046/j.1360-0443.2003.00426.x **(p. 72)**

Jones, J. T, Pelham, B. W., Carvallo, M., & Mirenberg, M. C. (2004). How do I love thee, let me count the Js: Implicit egotism and interpersonal attraction. *Journal of Personality and Social Behavior, 87,* 665–683. **(p. 627, 650)**

Joseph. J. A., Shukitt-Hale. B., Denisova. N. A., Bielinuski, D., Martin, A., McEwen. J. J., & Bickford, P. C. (1999). Reversals of age-related declines in neuronal signal transduction, cognitive, and motor behavioral deficits with blueberry, spinach, or strawberry dietary supplementation. *Journal of Neuroscience, 19,* 8114–8121. **(pp. 235, 376, 673)**

Judge, T. A., & Cable, D. M. (2010). When it comes to pay, do the thin win? The effect of weight on pay for men and women. *Journal of Applied Psychology, 96,* 95–112. doi: 10.1037/a0020860 **(pp. 34, 553)**

Kasparek, D. G., Corwin, S. J., Valois, R. F., Sargent, R. G., & Morris, R. L. (2008). Selected health behaviors that influence college freshman weight change. *Journal of American College Health, 56,* 437–444. **(p. 629)**

Katona, G. (1940). *Organizing and memorizing*. New York: Columbia University Press. **(p. 316)**

Keppel, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall. **(p. 447)**

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs*. New York: W. H. Freeman. **(p. 447)**

Khan, A., Brodhead. A. E., Kolts. R. L., & Brown, W. A. (2005). Severity of depressive symptoms and response to antidepressants and placebo in antidepressant trials. *Journal of Psychiatric Research, 39,* 145–150. **(p. 505)**

Killeen. P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353. **(p. 259)**

Kirschner, P. A., & Karpinski, A. C. (2010). Facebook and academic performance. *Computers in Human Behavior, 26,* 1237–1245. Doi: 10.1016/j.chb.2010.03.024 **(pp. 432, 503, 675)**

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125,* 470–500. **(p. 345, 675)**

Kolodinsky, J., Labrecque, J., Doyon, M., Reynolds, T., Oble, F., Bellavance, F., & Marquis, M. (2008). Sex and cultural differences in the acceptance of functional foods: A comparison of American, Canadian, and French college students. *Journal of American College Health*, *57,* 143–149. **(p. 348)**

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischblacher, U., & Fehr, R. (2005). Oxytocin increases trust in humans. *Nature, 435,* 673–676. **(p. 34)**

Kramer, S. E., Allessie, G. H. M., Dondorp. A. W., Zekveld, A. A., & Kapteyn, T. S. (2005). A home education program for older adults with hearing impairment and their significant others: A randomized evaluating short- and long-term effects. *International Journal of Audiology, 44,* 255–264. **(pp. 506, 630)**

Kuo, M., Adlaf. E. M., Lee, H., Gliksman, L., Demers, A., & Wechsler, H. (2002). More Canadian students drink but American students drink more: Comparing college alcohol use in two countries. *Addiction, 97,* 1583–1592. **(p. 34)**

Langewitz, W., Izakovic, J., & Wyler, J. (2005). Effect of self-hypnosis on hay fever symptoms—a randomized controlled intervention. *Psychotherapy and Psychosomatics, 74,* 165–172. **(p. 652)**

Liger-Belair, G., Bourget, M., Villaume, S., Jeandet, P., Pron, H., & and Polidori, G. (2010). On the losses of dissolved CO_2 during Champagne serving. *Journal of Agricultural and Food Chemistry, 58,* 8768–8775. DOI: 10.1021/jf101239w

Linde, L., & Bergstroem, M, (1992). The effect of one night without sleep on problem-solving and immediate recall. *Psychological Research, 54,* 127–136. **(p. 376)**

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning & Verbal Behavior, 13,* 585–589. **(pp. 20, 348, 592, 605, 629, 650)**

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171. **(p. 259)**

Mathews, E. M., & Wagner, D. R. (2008). Prevalence of overweight and obesity in collegiate American football players, by position. *Journal of American College Health*, *57,* 33–38. **(p. 347)**

McGee, E., & Shevliln, M. (2009). Effect of humor on interpersonal attraction and mate selection. *Journal of Psychology, 143,* 67–77. **(p. 313, 675)**

McGee, R., Williams, S., Howden-Chapman, P., Martin, J., & Kawachi. I. (2006). Participation in clubs and groups from childhood to adolescence and its effects on attachment and self-esteem. *Journal of Adolescence, 29,* 1–17. **(p. 275, 673)**

Montarello, S., & Martens, B. K. (2005). Effects of interspersed brief problems on students' endurance at completing math work. *Journal of Behavioral Education, 14,* 249–266. **(p. 276)**

Morse, C. K. (1993). Does variability increase with age? An archival study of cognitive measures. *Psychology and Aging, 8,* 156–164. **(p. 104)**

Miller, K. E. (2008). Wired: Energy drinks, jock identity, masculine norms, and risk taking. *Journal of American College Health*, *56,* 481–490. **(p. 275)**

Mulvihill, B. A., Obuseh, F. A., & Caldwell, C. (2008). Healthcare providers' satisfaction with a State Children's Health Insurance Program (SCHIP). *Maternal & Child Health Journal, 12,* 260–265. **(p. 628)**

Murdock. T. B., Miller, M., & Kohlhardt, J. (2004). Effects of classroom context variables on high school students' judgments of the acceptability and likelihood of cheating. *Journal of Educational Psychology, 96,* 765–777. **(p. 430)**

Pelton, T. (1983). The shootists. *Science83, 4*(4), 84–86. **(p. 378)**

Persson, J., Bringlov, E., Nilsson, L., & Nyberg. L. (2004). The memory-enhancing effects of Ginseng and Ginkgo biloba in health volunteers. *Psychopharmacology, 172,* 430–434. **(p. 275)**

Plomin, R., Corley, R., DeFries, J. C., & Fulker, D. W. (1990). Individual differences in television viewing in early childhood: Nature as well as nurture. *Psychological Science, 1,* 371–377. **(p. 553)**

Reed, E. T., Vernon, P. A., & Johnson, A. M. (2004). Confirmation of correlation between brain nerve conduction velocity and intelligence level in normal adults. *Intelligence, 32,* 563–572. **(p. 651)**

Reifman, A. S., Larrick, R. P., & Fein, S. (1991). Temper and temperature on the diamond: The heat–aggression relationship in major league baseball. *Personality and Social Psychology Bulletin, 17,* 580–585. **(p. 277)**

Resenhoeft, A., Villa, J., & Wiseman, D. (2008). Tattoos can harm perceptions: A study and suggestions. *Journal of American College Health*, *56,* 593–596. **(pp. 276, 380, 674)**

Rogers, T. B., Kuiper, N. A., & Kirker, W. S, (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology, 35,* 677–688. **(p. 386)**

Rohwedder, S., & Willis, R. J. (2010). Mental retirement. *Journal of Economic Perspectives, 24,* 119–138. Doi: 10.1257/jep.24.1.119 **(p. 587)**

Rozin, P., Bauer, R., & Cantanese, D. (2003). Food and life, pleasure and worry, among American college students: Gender differences and regional similarities. *Journal of Personality and Social Psychology, 85,* 132–141. **(pp. 347, 555)**

Ryan, C. S., & Hemmes, N. S. (2005). Effects of the contingency for homework submission on homework submission and quiz performance in a college course. *Journal of Applied Behavior Analysis, 38,* 79–88. Doi: 10.1901/jaba.2005.123–03 **(p. 676)**

Scaife, M. (1976). The response to eye-like shapes by birds. I. The effect of context: A predator and a strange bird. *Animal Behaviour, 24,* 195–199. **(pp. 284, 312)**

Schachter. S. (1968). Obesity and eating. *Science, 161,* 751–756. **(p. 494)**

Schmidt, S. R. (1994). Effects of humor on sentence memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20,* 953–967. **(pp. 35, 38, 81, 349, 375)**

Segal, S. J., & Fusella, V. (1970). Influence of imaged pictures and sounds on detection of visual and auditory signals. *Journal of Experimental Psychology, 83,* 458–464. **(p. 377)**

Shrauger, J. S. (1972). Self-esteem and reactions to being observed by others. *Journal of Personality and Social Psychology, 23,* 192–200. **(p. 466)**

Shulman, C., & Guberman, A. (2007). Acquisition of verb meaning through syntactic cues: A comparison of children with autism, children with specific language impairment (SLI) and children with typical language development (TLD). *Journal of Child Language, 34,* 411–423. **(p. 652)**

Slater, A., Von der Schulenburg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S., & Samuels, C. (1998). Newborn infants prefer attractive faces. *Infant Behavior and Development, 21,* 345–354. **(p. 292)**

Smith, C., & Lapp, L. (1991). Increases in number of REMs and REM density in humans following an intensive learning period. *Sleep: Journal of Sleep Research & Sleep Medicine, 14,* 325–330. **(p. 276)**

Smyth, J. M., Stone, A. A., Hurewitz, A., & Kaell, A. (1999). Effects of writing about stressful experiences on symptom reduction in patients with asthma or rheumatoid arthritis: A randomized trial. *Journal of the American Medical Association, 281,* 1304–1309. **(p. 652)**

Sol, D., Lefebvre, L., & Rodriguez-Teijeiro, J. D. (2005). Brain size, innovative propensity and migratory behavior in temperate Palaearctic birds. *Proceedings [Proc Biol Sci], 272,* 1433–441. **(p. 431)**

Stephens, R., Atkins, J., & Kingston, A. (2009). Swearing as a response to pain. *NeuroReport: For Rapid Communication of Neuroscience Research, 20*, 1056–1060. Doi: 10.1097/WNR.0b013e32832e64b1 **(pp. 352, 379, 651)**

Stickgold, R., Whidbee, D., Schirmer B., Patel, V., & Hobson, J. A. (2000). Visual discrimination task improvement: A multi-step process occurring during sleep. *Journal of Cognitive Neuroscience, 12,* 246–254. **(p. 349)**

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A non-obtrusive test of the facial feedback hypothesis. *Journal of Personality & Social Psychology, 54,* 768–777. **(pp. 34, 376)**

Sümer, N., Özkan, T., & Lajunen, T. (2006). Asymetric relationship between driving and safety skills. *Accident Analysis and Prevention, 38,* 703–711. **(p. 506)**

Trockel, M. T., Barnes, M. D., & Egget, D. L. (2000). Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors. *Journal of American College Health, 49,* 125–131. **(p. 12)**

Tryon, R. C. (1940). Genetic differences in maze-learning ability in rats. *Yearbook of the National Society for the Study of Education, 39,* 111–119. **(p. 133)**

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley. **(pp. 60, 432, 669)**

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232. **(p. 164)**

Twenge, J. M. (2000). The age of anxiety? Birth cohort change in anxiety and neuroticism, 1952-1993. *Journal of Personality and Social Psychology, 79,* 1007–1021. **(p. 312)**

U.S. Census Bureau. (2005). *Americans spend more than 100 hours commuting to work each year, Census Bureau reports.* Retrieved January 14, 2009, from www.census.gov/Press-Release/www/releases/archives/american_community_survey_acs/004489.html **(p. 182)**

von Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education, 13.* **(p. 95)**

Wegesin, D. J., & Stern, Y. (2004). Inter- and intra-individual variability in recognition memory: Effects of aging and estrogen use. *Neuropsychology, 18,* 646–657. **(p. 104)**

Welsh, R. S., Davis, M. J., Burke, J. R., & Williams, H. G. (2002). Carbohydrates and physical/mental performance during intermittent exercise to fatigue. *Medicine & Science in Sports & Exercise, 34,* 723–731. **(p. 229)**

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594–604. **(p. 262)**

Winget, C., & Kramer, M. (1979). *Dimensions of dreams*. Gainesville: University Press of Florida. **(p. 630)**

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 5012–5015. **(p. 200)**

Yapko, M. D. (1994). Suggestibility and repressed memories of abuse: A survey of psychotherapists' beliefs. *American Journal of Clinical Hypnosis, 36,* 163–171. **(p. 651)**

Ye, Y. Beyond materialism: The role of health-related beliefs in the relationship between television viewing and life satisfaction among college students. *Mass Communication and Society, 13,* 458–478. Doi: 10.1080/15205430903296069 **(p. 432)**

Zhong, C., Bohns, V. K., & Gino, F. (2010). Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological Science, 21,* 311–314. **(pp. 349, 675)**

Zhou, X., Vohs, K. D., & Baumeister, R. F. (2009). The symbolic power of money: Reminders of money after social distress and physical pain. *Psychological Science, 20,* 700–706. **(p. 14)**

This page intentionally left blank

Index