Chapter 10. Experimental Design: Statistical Analysis of Data

Purpose of Statistical Analysis

In previous chapters, we have discussed the basic principles of good experimental design. Before examining specific experimental designs and the way that their data are analyzed, we thought that it would be a good idea to review some basic principles of statistics. We assume that most of you reading this book have taken a course in statistics. However, our experience is that statistical knowledge has a mysterious quality that inhibits long-term retention. Actually, there are several reasons why students tend to forget what they learned in a statistics course, but we won't dwell on those here. Suffice it to say, a chapter to refresh that information will be useful.

When we conduct a study and measure the dependent variable, we are left with sets of numbers. Those numbers inevitably are not the same. That is, there is variability in the numbers. As we have already discussed, that variability can be, and usually is, the result of multiple variables. These variables include extraneous variables such as individual differences, experimental error, and confounds, but may also include an effect of the independent variable. The challenge is to extract from the numbers a meaningful summary of the behavior observed and a meaningful conclusion regarding the influence of the experimental treatment (independent variable) on participant behavior. Statistics provide us with an objective approach to doing this.

Descriptive Statistics

Central Tendency and Variability

In the course of doing research, we are called on to summarize our observations, to estimate their reliability, to make comparisons, and to draw inferences. **Measures of central tendency** such as the mean, median, and mode summarize the performance level of a group of scores, and **measures of variability** describe the spread of scores among participants. Both are important. One provides information on the level of performance, and the other reveals the consistency of that performance.

Let's illustrate the two key concepts of central tendency and variability by considering a scenario that is repeated many times, with variations, every weekend in the fall and early winter in the high school, college, and professional ranks of our nation. It is the crucial moment in the football game. Your team is losing by four points. Time is running out, it is fourth down with two yards to go, and you need a first down to keep from losing possession of the ball. The quarterback must make a decision: run for two or pass. He calls a timeout to confer with the offensive coach, who has kept a record of the outcome of each offensive play in the game. His report is summarized in Table 10.1.

Table 10.1 **Yards Gained (or Lost) on 20 Pass and 20 Running Plays, Ordered from Largest Gain (Left Column) to Smallest Gain or Largest Loss (Right Column)**

Pass Plays				Running Plays			
31	9	0	−1	10	5	3	2
29	8	0	−5	10	4	3	1
20	5	0	−10	8	4	3	1
15	1	0	−15	8	3	3	1
10	0	0	−17	6	3	2	0

To make the comparison more visual, the statistician had prepared a chart of these data (Figure 10.1).

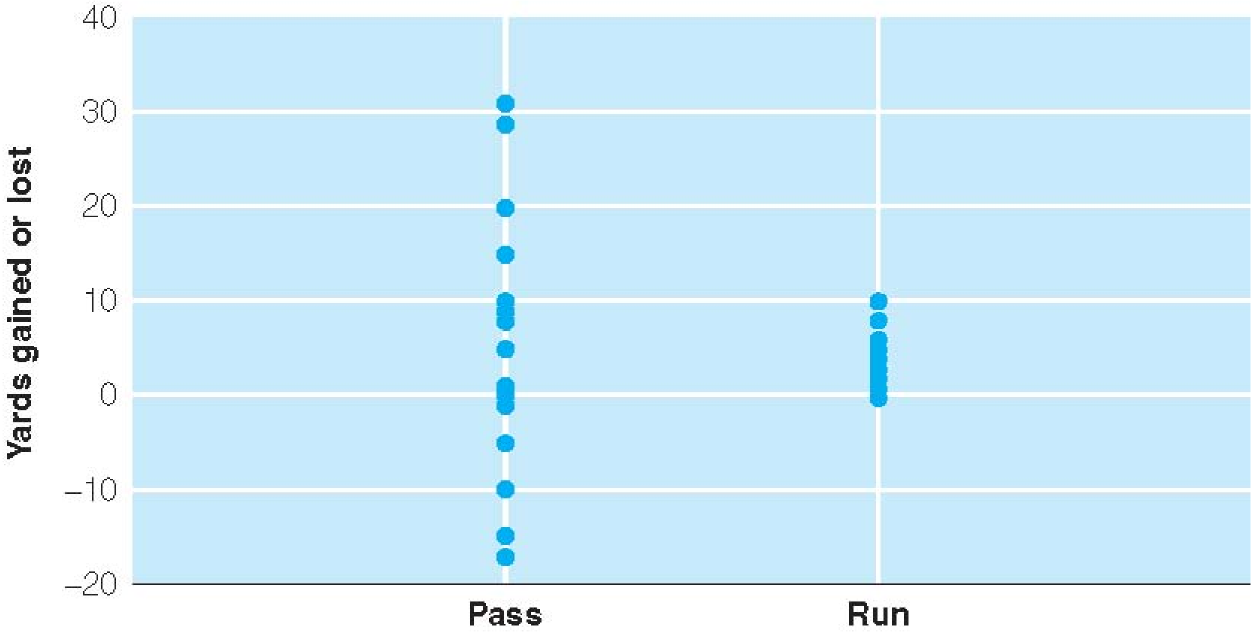

Figure 10.1 Yards gained or lost by passing and running plays. The mean gain per play, +4 yards, is identical for both running and passing plays.

What we have in Figure 10.1 are two frequency distributions of yards per play. A **frequency distribution** shows the number of times each score (in this case, the number of yards) is obtained. We can tell at a glance that these two distributions are markedly different. A pass play is a study in contrasts; it leads to extremely variable outcomes. Indeed, throwing a pass is somewhat like playing Russian roulette. Large gains, big losses, and incomplete passes (0 gain) are intermingled. A pass doubtless carries with it considerable excitement and apprehension. You never really know what to expect. On the other hand, a running play is a model of consistency. If it is not exciting, it is at least dependable. In no case did a run gain more than ten yards, but neither were there any losses. These two distributions exhibit extremes of variability. In this example, a coach and quarterback would probably pay little attention to measures of central tendency. As we shall see, the fact that the mean gain per pass and per run is the same would be of little relevance. What is relevant is the fact that the variability of running plays is less. It is a more dependable play in a short yardage situation. Seventeen of 20 running plays netted two yards or more. In contrast, only 8 of 20 passing plays gained as much as two yards. Had the situation been different, of course, the decision about what play to call might also have been different. If it were the last play in the ball game and 15 yards were needed for a touchdown, the pass would be the play of choice. Four times out of 20 a pass gained 15 yards or more, whereas a run never came close. Thus, in the strategy of football, variability is fundamental consideration. This is, of course, true of many life situations.

Some investors looking for a chance of a big gain will engage in speculative ventures where the risk is large but so, too, is the potential payoff. Others pursue a strategy of investments in blue chip stocks, where the proceeds do not fluctuate like a yo-yo. Many other real-life decisions are based on the consideration of extremes. A bridge is designed to handle a maximum rather than an average load; transportation systems and public utilities (such as gas, electric, water) must be prepared to meet peak rather than average demand in order to avoid shortages and outages.

Researchers are also concerned about variability. By and large, from a researcher's point of view, variability is undesirable. Like static on an AM radio, it frequently obscures the signal we are trying to detect. Often the signal of interest in psychological research is a measure of central tendency, such as the mean, median, or mode.

Measures of Central Tendency

The Mean. Two of the most frequently used and most valuable measures of central tendency in psychological research are the mean and median. Both tell us something about the central values or typical measure in a distribution of scores. However, because they are defined differently, these measures often take on different values. The **mean,** commonly known as the arithmetic average, consists of the sum of all scores divided by the number of scores. Symbolically, this is shown as $\overline{X} = \frac{\sum X}{n}$ in which $\overline{X}$ is the mean; the sign $\sum$ directs us to sum the values of the variable *X*. (Note: When the mean is abbreviated in text, it is symbolized *M*). Returning to Table 10.1, we find that the sum of all yards gained (or lost) by pass plays is 80. Dividing this sum by *n* (20) yields *M* =

4. Since the sum of yards gained on the ground is also 80 and *n* is 20, the mean yards gained per carry is also 4. If we had information only about the mean, our choice between a pass or a run would be up for grabs. But note how much knowledge of variability adds to the decision-making process. When considering the pass play, where the variability is high, the mean is hardly a precise indicator of the typical gain (or loss). The signal (the mean) is lost in a welter of static (the variability). This is not the case for the running play. Here, where variability is low, we see that more of the individual measures are near the mean. With this distribution, then, the mean is a better indicator of the typical gain.

It should be noted that each score contributes to the determination of the mean. Extreme values draw the mean in their direction. Thus, if we had one running play that gained 88 yards, the sum of gains would be 160, *n* would equal 21, and the mean would be 8. In other words, the mean would be doubled by the addition of one very large gain.

The Median. The median does not use the value of each score in its determination. To find the **median,** you arrange the values of the variable in order—either ascending or descending—and then count down (*n* + 1) / 2 scores. This score is the median. If *n* is an even number, the median is halfway between the two middle scores. Returning to Table 10.1, we find the median gain on a pass play by counting down to the 10.5th case [(20 + 1) / 2 = 10.5)]. This is halfway between the 10th and 11th scores. Because both are 0, the median gain is 0. Similarly, the median gain on a running play is 3.

The median is a particularly useful measure of central tendency when there are extreme scores at one end of a distribution. Such distributions are said to be skewed in the direction of the extreme scores. The median, unlike the mean, is unaffected by these scores; thus, it is more likely than the mean to be representative of central tendency in a skewed distribution. Variables that have restrictions at one end of a distribution but not at the other are prime candidates for the median as a measure of central tendency. A few examples are time scores (0 is the theoretical lower limit and there is no limit at the upper end), income (no one earns less than 0 but some earn in the millions), and number of children in a family (many have 0 but only one is known to have achieved the record of 69 by the same mother).

The Mode. A rarely used measure of central tendency, the **mode** simply represents the most frequent score in a distribution. Thus, the mode for pass plays is 0, and the mode for running plays is 3. The mode does not consider the values of any scores other than the most frequent score. The mode is most useful when summarizing data measured on a nominal scale of measurement. It can also be valuable to describe a multimodal distribution, one in which the scores tend to occur most frequently around 2 or 3 points in the distribution.

Measures of Variability

We have already seen that a measure of central tendency by itself provides only a limited amount of information about a distribution. To complete the description, it is necessary to have some idea of how the scores are distributed about the central value. If they are widely dispersed, as with the pass plays, we say that variability is high. If they are distributed compactly about the central value, as with the running plays, we refer to the variability as low. But *high* and *low* are descriptive words without precise quantitative meaning. Just as we needed a quantitative measure of centrality, so also do we require a quantitative index of variability.

The Range. One simple measure of variability is the **range,** defined as the difference between the highest and lowest scores in a distribution. Thus, referring to Table 10.1, we see that the range for pass plays is 31 – (–17) = 48; for running plays, it is 10 – 0 = 10. As you can see, the range provides a quick estimate of the variability of the two distributions. However, the range is determined by only the two most extreme scores. At times this may convey misleading impressions of total variability, particularly if one or both of these extreme scores are rare or unusual occurrences. For this and other reasons, the range finds limited use as a measure of variability.

The Variance and the Standard Deviation. Two closely related measures of variability overcome these disadvantages of the range: **variance** and **standard deviation.** Unlike the range, they both make use of all the scores in their computation. Indeed, both are based on the squared deviations of the scores in the distribution from the mean of the distribution.

Table 10.2 illustrates the number of aggressive behaviors during a one-week observation period for two different groups of children. The table includes measures of central tendency and measures of variability. Note that the symbols and formulas for variance and standard deviation are those that use sample data to provide estimates of variability in the population.

Table 10.2 Measures of Central Tendency and Measures of Variability for Aggression Scores From Two Different Groups of Children

	Group 1			Group 2		
	Score (X)	$X-\overline{X}$	$(X-\overline{X})^2$	Score (X)	$X-\overline{X}$	$(X-\overline{X})^2$
	14	2	4	18	6	36
	13	1	1	15	3	9
	12	0	0	12	0	0
	12	0	0	12	0	0
	11	−1	1	9	−3	9
	10	−2	4	6	−6	36
	$\Sigma = 72$	$\Sigma = 0$	$\Sigma = 10$	$\Sigma = 72$	$\Sigma = 0$	$\Sigma = 90$
Mean ($\overline{X}$)	12			12		
Median	12			12		
Range	4			12		
Variance (s^2) $\frac{\Sigma(X-\overline{X})^2}{n-1}$	2			18		
Standard deviation (s) $\sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$	1.41			1.97		

Notice that although the measures of central tendency are identical for both groups of scores, the measures of variability are not and reflect the greater spread of scores in Group 2. This is apparent in all three measures of variability (range, variance, standard deviation). Also notice that the variance is based on the squared deviations of scores from the mean and that the standard deviation is simply the square root of the variance. For most sets of scores that are measured on an interval or ratio scale of measurement, the standard deviation is the preferred measure of variability. Conceptually, you should think of standard deviation as “on average, how far scores are from the mean.”

Now, if the variable is distributed in a bell-shaped fashion known as the normal curve, the relationships can be stated with far more precision. Approximately 68% of the scores lie between the mean and ±1 standard deviation, approximately 95% of the scores lie between ±2 standard deviations, and approximately 98% of the scores lie between ±3 standard deviations. These features of normally distributed variables are summarized in Figure 10.2.

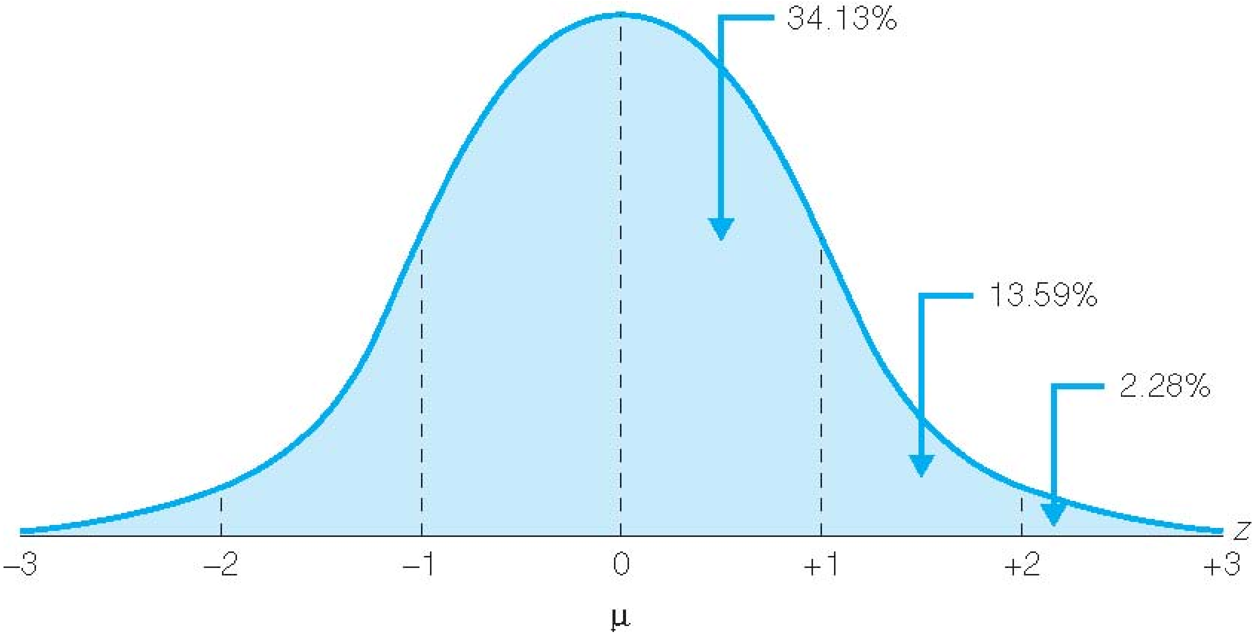

Figure 10.2 Areas between the mean and selected numbers of standard deviations above and below the mean for a normally distributed variable.

Note that these areas under the normal curve can be translated into probability statements. Probability and proportion are simply percentage divided by 100. The proportion of area found between any two points in Figure 10.2 represents the probability that a score, drawn at random from that population, will assume one of the values found between these two points. Thus, the probability of selecting a score that falls between 1 and 2 standard deviations above the mean is 0.1359. Similarly, the probability of selecting a score 2 or more standard deviations below the mean is 0.0228 (0.0215 + .0013).

Many of the variables with which psychologists concern themselves are normally distributed, such as standardized test scores. What is perhaps of greater significance for the researcher is the fact that distributions of sample statistics tend toward normality as sample size increases. This is true even if the population distribution is not normal. Thus, if you were to select a large number of samples of fixed sample size, say n = 30, from a nonnormal distribution, you would find that separate plots of their means, medians, standard deviations, and variances would be approximately normal.

The Importance of Variability

Why is variability such an important concept? In research, it represents the noisy background out of which we are trying to detect a coherent signal. Look again at Figure 10.1. Is it not clear that the mean is a more coherent representation of the typical results of a running play than is the mean of a pass play? When variability is large, it is simply more difficult to regard a measure of central tendency as a dependable guide to representative performance.

This also applies to detecting the effects of an experimental treatment. This task is very much like distinguishing two or more radio signals in the presence of static. In this analogy, the effects of the experimental variable (treatment) represent the radio signals, and the variability is the static (noise). If the radio signal is strong, relative to the static, it is easily detected; but if the radio signal is weak, relative to the static, the signal may be lost in a barrage of noise.

In short, two factors are commonly involved in assessing the effects of an experimental variable: a measure of centrality, such as the mean, median, or proportion; and a measure of variability, such as the standard deviation. Broadly speaking, the investigator exercises little control over the measure of centrality. If the effect of the treatment is large, the differences in measures of central tendency will generally be large. In contrast, control over variability is possible. Indeed, much of this text focuses, directly or indirectly, on procedures for reducing variability—for example, selecting a reliable dependent variable, providing uniform instructions and standardized experimental procedures, and controlling obtrusive and extraneous experimental stimuli. We wish to limit the extent of this unsystematic variability for much the same reasons that a radio operator wishes to limit static or noise—to permit better detection of a treatment effect in the one case and a radio signal in the other. The lower the unsystematic variability (random error), the more sensitive is our statistical test to treatment effects.

Tables and Graphs

Raw scores, measures of central tendency, and measures of variability are often presented in tables or graphs. Tables and graphs provide a user-friendly way of summarizing information and revealing patterns in the data. Let’s take a hypothetical set of data and play with it.

One group of 30 children was observed on the playground after watching a TV program without violence, and another group of 30 children was observed on the playground after watching a TV program with violence. In both cases, observers counted the number of aggressive behaviors. The data were as follows:

Program with no violence: 5, 2, 0, 4, 0, 1, 2, 1, 3, 6, 5, 1, 4, 2, 3, 2, 2, 2, 5, 3, 4, 2, 2, 3, 4, 3, 7, 3, 6, 3, 3

Program with violence: 5, 3, 1, 4, 2, 0, 5, 3, 4, 2, 6, 1, 4, 1, 5, 3, 7, 2, 4, 2, 3, 5, 4, 6, 3, 4, 4, 5, 6, 5

Take a look at the raw scores. Do you see any difference in number of aggressive behaviors between the groups? If you are like us, you find it difficult to tell.

One of the first ways we might aid interpretation is to place the raw scores in a table called a frequency distribution and then translate that same information into a graph called a frequency histogram (see Figure 10.3).

No Violence Group	
Score (X)	*Frequency (f)*
7	1
6	2
5	3
4	4
3	7
2	8
1	3
0	2

Violence Group	
Score (X)	*Frequency (f)*
7	1
6	3
5	6
4	7
3	5
2	4
1	3
0	1

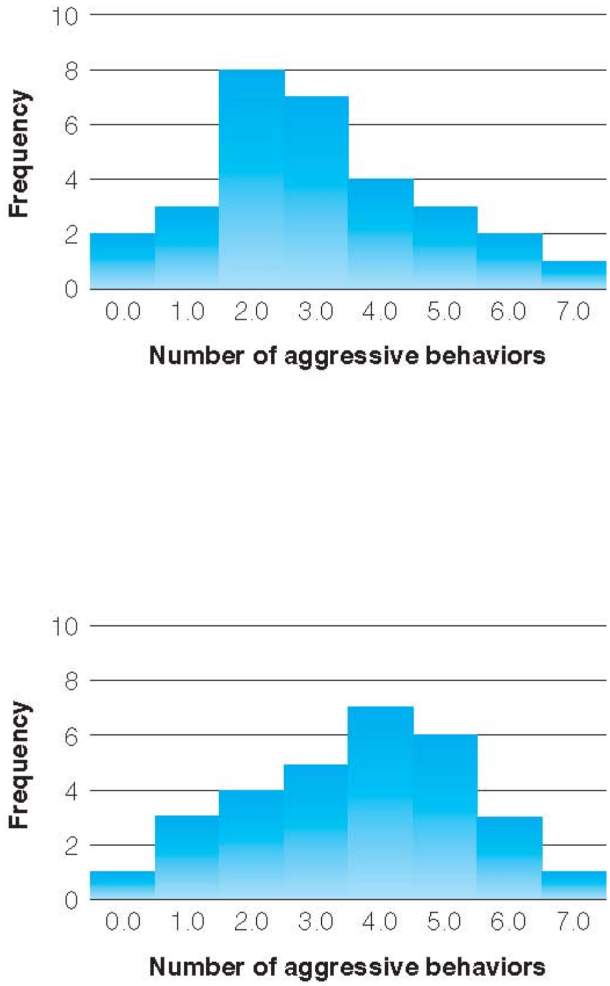

Figure 10.3 Number of aggressive behaviors illustrated in both frequency distributions and frequency histograms.

Both the frequency distribution and the frequency histogram in Figure 10.3 make it easy to detect the range of scores, the most frequent score (mode), and the shape of the distribution. A quick glance at the graphs now suggests that the children tended to exhibit fewer aggressive behaviors after the TV program with no violence. We can further summarize the data by calculating the mean and standard deviation for each group and presenting these values in both a table and a figure (see Table 10.3 and Figure 10.4).

Table 10.3 Descriptive Statistics for the No Violence and Violence Groups

	No Violence	Violence
M	3.00	3.60
SD	1.74	1.73
n	30	30

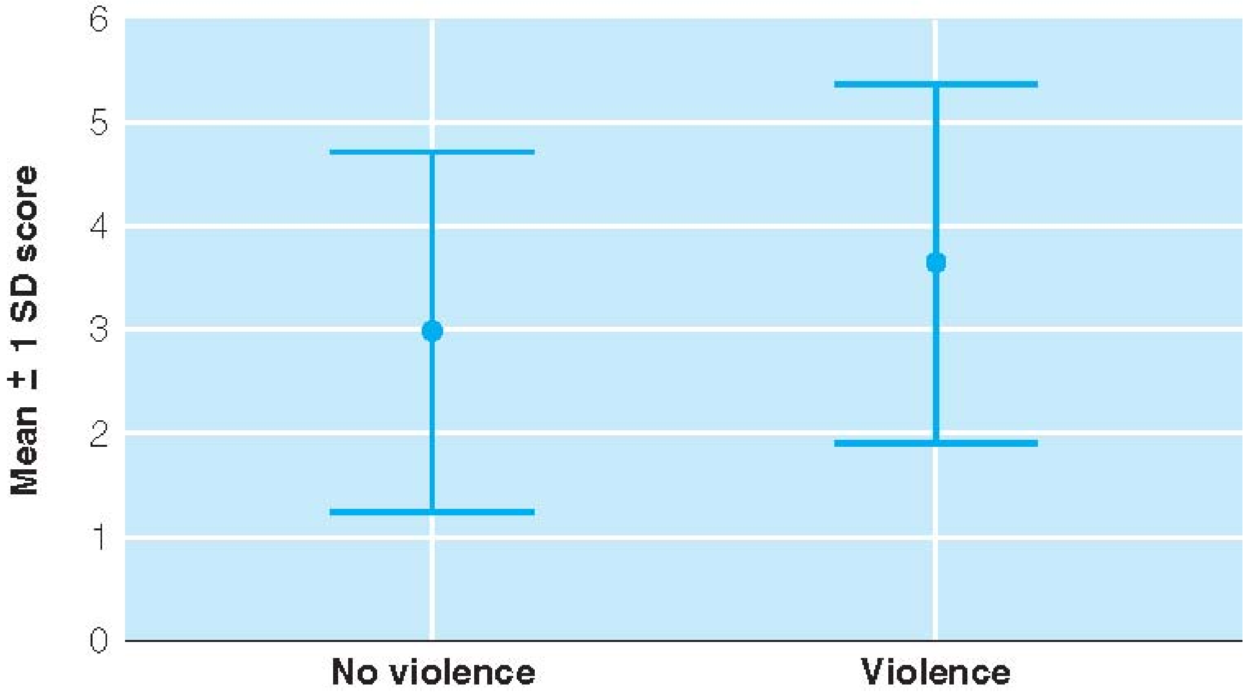

Figure 10.4 Graphical depiction of the mean and standard deviation for the No Violence and Violence groups.

In Figure 10.4, the mean is depicted by a square, and the bars represent 1 standard deviation above and below the mean. Thus, although the means differ by 0.6 units, one can see from the standard deviation bars that there is quite a bit of overlap between the two sets of scores. Inferential statistics will be needed to determine whether the difference between the means is statistically significant.

In the preceding description of data, we selected a few ways that the data could be summarized in both tables and figures. However, these methods are certainly not exhaustive. We can display these data and other data in a variety of other ways, in both tabular and graphical form, and we encourage students to experiment with these techniques. Remember that the data are your window into participant behavior and thought. You can only obtain a clear view by careful examination of the scores.

Before we turn to inferential statistics, let’s think about the added clarity that descriptive statistics can provide when observed behavior is described. To do this, we return to the report of a study that was first described in Chapter 6 and is considered again here in the box “Thinking Critically About Everyday Information.”

Thinking Critically About Everyday Information: School Backpacks and Posture Revisited

A news report by MSNBC describes a study in which children were observed carrying school backpacks. The article states:

> Thirteen children ages 8 and 9 walked about 1,310 feet without a backpack, and wearing packs weighing 9 and 13 pounds, while researchers filmed them with a high-speed camera. . . . The kids did not change their strides, the images showed. Instead, the youngsters bent forward more as they tried to counter the loads on their backs, and the heavier loads made them bend more, the study found. As they grew more tired, their heads went down, Orloff said.

In Chapter 6, we focused our critical thinking questions on the method of observation. Now, let’s think about the description of the observations:

- The article states that when children carried the backpacks, they “did not change their strides” but “bent forward more.” Although this description is typical of a brief news report, what specific measure of central tendency could be reported for each dependent variable to clarify the description?
- What measure of variability would clarify the description?
- How would you create a graph that would nicely summarize the pattern of results reported in the article?
- Why is it that most reports of research in the everyday media do not report measures of central tendency and variability, with related tables or graphs?

Retrieved June 11, 2003. online at http://www.msnbc.com/news/922623.asp?0si=-

Inferential Statistics

From Descriptions to Inferences

We have examined several descriptive statistics that we use to make sense out of a mass of raw data. We have briefly reviewed the calculation and interpretation of statistics that are used to describe both the central tendency of a distribution of scores or quantities (mean, median, and mode) and the dispersion of scores around the central tendency (range, standard deviation, and variance). Our goal in descriptive statistics is to describe, with both accuracy and economy of statement, aspects of samples selected from the population.

It should be clear that our primary focus is not on the sample statistics themselves. Their value lies primarily in the light that they may shed on characteristics of the population. Thus, we are not interested, as such, in the fact that the mean of the control group was higher or lower than the mean of an experimental group, nor that a sample of 100 voters revealed a higher proportion favoring Candidate A. Rather, our focus shifts from near to far vision; it shifts from the sample to the population. We wish to know if we may justifiably conclude that the experimental variable has had an effect; or we wish to predict that Candidate A is likely to win the election. Our descriptive statistics provide the factual basis for the inductive leap from samples to populations.

In the remainder of this chapter, we will take a conceptual tour of statistical decision making. The purpose is not to dwell on computational techniques but rather to explore the rationale underlying inferential statistics.

The Role of Probability Theory

Recall the distinction between deductive and inductive reasoning. With deductive reasoning, the truth of the conclusion is implicit in the assumptions. Either we draw a valid conclusion from the premises, or we do not. There is no in-between ground. This is not the case with inductive or scientific proof. Conclusions do not follow logically from a set of premises. Rather, they represent extensions of or generalizations based on empirical observations. Hence, in contrast to logical proof, scientific or inductive conclusions are not considered valid or invalid in any ultimate sense. Rather than being either right or wrong, we regard scientific propositions as having a given probability of being valid. If observation after observation confirms a proposition, we assign a high probability (approaching 1.00) to the validity of the proposition. If we have deep and abiding reservations about its validity, we may assign a probability that approaches 0. Note, however, we never establish scientific truth, nor do we disprove its validity, with absolute certainty.

Most commonly, probabilities are expressed either as a proportion or as a percentage. As the probability of an event approaches 1.00, or 100%, we say that the event is likely to occur. As it approaches 0.00, or 0%, we deem the event unlikely to occur. One way of expressing probability is in terms of the number of events favoring a given outcome relative to the total number of events possible. Thus,

$$P_A = \frac{\text{number of events favoring A}}{\text{number of events favoring A} + \text{number of events not favoring A}}$$

To illustrate, if a population of 100,000 individuals contains 10 individuals with the disorder phenylketonuria (PKU), what is the probability that a person, selected at random, will have PKU?

$$P_{PKU} = \frac{10}{10 + 99{,}990} = 0.0001 \text{ or } 0.01\%$$

Thus, the probability is extremely low: 1 in 10,000.

This definition is perfectly satisfactory for dealing with discrete events (those that are counted). However, how do we define probability when the variables are continuous—for example, weight, IQ score, or reaction time? Here, probabilities can be expressed as a proportion of one area under a curve relative to the total area under a curve. Recall the normal distribution. The total area under the curve is 1.00. Between the mean and 1 standard deviation above the mean, the proportion of the total area is 0.3413. If we selected a sample score from a normally distributed population, what is the probability that it would be between the mean and 1 standard deviation above the mean? Because about 34% of the total area is included between these points, $p = 0.34$. Similarly, $p = 0.34$ that a single randomly selected score would be between the mean and 1 standard deviation below the mean. Figure 10.2 shows areas under the standard normal curve and permits the expression of any value of a normally distributed variable in terms of probability.

Probability looms large on the scene of inferential statistics because it is the basis for accepting some hypotheses and rejecting others.

The Null and Alternative Hypotheses

Before beginning an experiment, the researcher sets up two mutually exclusive hypotheses. One is a statistical hypothesis that the experimenter expects to reject. It is referred to as the **null hypothesis** and is usually represented symbolically as $\boldsymbol{H_0}$**.** The null hypothesis states some expectation regarding the value of one or more population parameters. Most commonly, it is a hypothesis of no difference (no effect). Let us look at a few examples:

- If we were testing the honesty of a coin, the null hypothesis (H_o) would read: The coin is unbiased. Stated more precisely, the probability of a head is equal to the probability of a tail: $P_h = P_t = ½ = 0.5$.
- If we were evaluating the effect of a drug on reaction time, the null hypothesis might read: The drug has no effect on reaction time.

The important point to remember about the null hypothesis is that it always states some expectation regarding a population parameter—such as the population mean, median, proportion, standard deviation, or variance. It is never stated in terms of expectations of a sample. For example, we would never state that the sample mean (or median or proportion) of one group is equal to the sample mean of another. It is a fact of sampling behavior that sample statistics are rarely identical, even if selected from the same population. Thus, ten tosses of a single coin will not always yield five heads and five tails. The discipline of statistics sets down the rules for making an inductive leap from sample statistics to population parameters.

The **alternative hypothesis (H_1)** denies the null hypothesis. If the null hypothesis states that there is no difference in the population means from which two samples were drawn, the alternative hypothesis asserts that there is a difference. The alternative hypothesis usually states the investigator's expectations. Indeed, there really would be little sense embarking upon costly and time-consuming research unless we had some reason for expecting that the experimental variable will have an effect. Let's look at a few examples of alternative hypotheses:

- In the study aimed at testing the honesty of a coin, the alternative hypothesis would read: H_1: $P_h \neq P_t \neq 1/2$*;* the probability of a head is not equal to the probability of a tail, which is not equal to one-half.
- In the effect of a drug on reaction time, the alternative hypothesis might read: The administration of a given dosage level of a drug affects reaction time.

The Sampling Distribution and Statistical Decision Making

Now that we have stated our null and alternative hypotheses, where do we go from here? Recall that these hypotheses are mutually exclusive. They are also *exhaustive.* By exhaustive we mean that no other possibility exists. These two possible outcomes in our statistical decision exhaust all possible outcomes. If the null hypothesis is true, then the alternative hypothesis must be false. Conversely, if the null hypothesis is false, then the alternative hypothesis must be true.

Considering these realities, our strategy would appear to be quite straightforward—simply determine whether the null hypothesis is true or false. Unfortunately, there is one further wrinkle. The null hypothesis can never be proved to be true. How would you go about proving that a drug has no effect, or that males and females are equally intelligent, or that a coin is honest? If you flipped it 1,000,000 times and obtained exactly 500,000 heads, wouldn't that be proof positive? No. It would merely indicate that, if a bias does exist, it must be exceptionally small. But we cannot rule out the possibility that a small bias does exist. Perhaps the next million, 5 million, or 10 billion tosses will reveal this bias. So we have a dilemma. If we have no way of proving one of two mutually exclusive and exhaustive hypotheses, how can we establish which of these alternatives has the higher probability of being true?

Fortunately, there is a way out of this dilemma. If we cannot prove the null hypothesis, we can set up conditions that permit us to reject it. For example, if we had tossed the coin 1,000,000 times and obtained 950,000 heads, would anyone seriously doubt the bias of the coin? Clearly, we would reject the null hypothesis that the coin is honest. The critical factor in this decision is our judgment that an outcome this rare is unlikely to have been the result of chance factors. It happened for a reason, and that reason is to be found in the characteristics of the coin or in the way it was tossed.

In this particular example, we did not engage in any formal statistical exercise in order to reject H_o. Our lifelong experience with coin-tossing experiments provided a frame of reference that permitted us to make the judgment. Because the obtained outcome is monumentally rare, we conclude that it did not occur by chance. However, in science, we often do not have frames of reference, based on experience, that permit us to dispense with formal statistical analyses. Nor are we often afforded the luxury of a sample size equal to 1 million. The frame of reference for statistical decision making is provided by the **sampling distribution** of a statistic. A sampling distribution is a theoretical probability distribution of the possible values of some sample statistic that would occur if we were to draw all possible samples of a fixed size from a given population. There is a sampling distribution for every statistic—mean, standard deviation, variance, proportion, median, and so on.

To illustrate, imagine we had a population of six scores: 1, 2, 3, 3, 4, 5. Suppose we randomly select a single score from this population, return it, and randomly select a second score. We call these two scores a random sample of $n = 2$, and we calculate a mean. Now imagine that we selected all possible samples of $n = 2$ from that population and calculated a mean for each. Table 10.4 shows all possible outcomes of this sampling experiment. Each cell shows the mean of the two scores that make up each sample. Thus, if the first selection is l and the second is 1, the mean of the sample is 1.0.

Table 10.4 Sample Means Resulting From Selecting All Possible Samples of n = 2 From a Population of Six Scores

Second Selection	First Selection					
	1	**2**	**3**	**3**	**4**	**5**
1	1.0	1.5	2.0	2.0	2.5	3.0
2	1.5	2.0	2.5	2.5	3.0	3.5
3	2.0	2.5	3.0	3.0	3.5	4.0
3	2.0	2.5	3.0	3.0	3.5	4.0
4	2.5	3.0	3.5	3.5	4.0	4.5
5	3.0	3.5	4.0	4.0	4.5	5.0

Now we can record the frequency with which each mean would be obtained. When we do so, we have constructed a frequency distribution of means of sample size $n = 2$. Table 10.5 shows this frequency distribution of the sample means.

Table 10.5 Frequency Distribution and Sampling Distribution of Means Based on Samples of $n = 2$ Drawn at Random From a Population of Six Numbers

Note that the mean of the distribution of sample means is the same as the population mean.

	Sample Mean	Frequency	*P*
	1.0	1	.028
	1.5	2	.056
	2.0	5	.139
	2.5	6	.167
Mean →	3.0	8	.222
	3.5	6	.167
	4.0	5	.139
	4.5	2	.056
	5.0	1	.028
		Total = 36	Total = 1.002*

*The sum of the probabilities is 1.00. The slight disparity is due to rounding error.

Now, if we divide the frequency with which a given mean was obtained by the total number of sample means (36), we obtain the probability of selecting that mean (last column in Table 10.5). Thus, eight different samples of $n = 2$ would yield a mean equal to 3.0. The probability of selecting that mean is 8/36 = 0.222. Note that, by chance, we would rarely select a mean of 1.0 (p = 0.028) or a mean of 5.0 (p = 0.028).

In this example, we used a very small population to illustrate a sampling distribution of a statistic. In real life, the populations are often extremely large. Let us imagine that we had a large population of scores and we selected a large number of samples of a given size (say, n = 30). We could construct a distribution of the sample means. We would find that the mean of this distribution equals the mean of the population, and the form of the distribution would tend to be normal—even if the population distribution is not normal. In fact, the larger the sample size, the more closely the distribution of sample means will approximate a normal curve (see Figure 10.5).

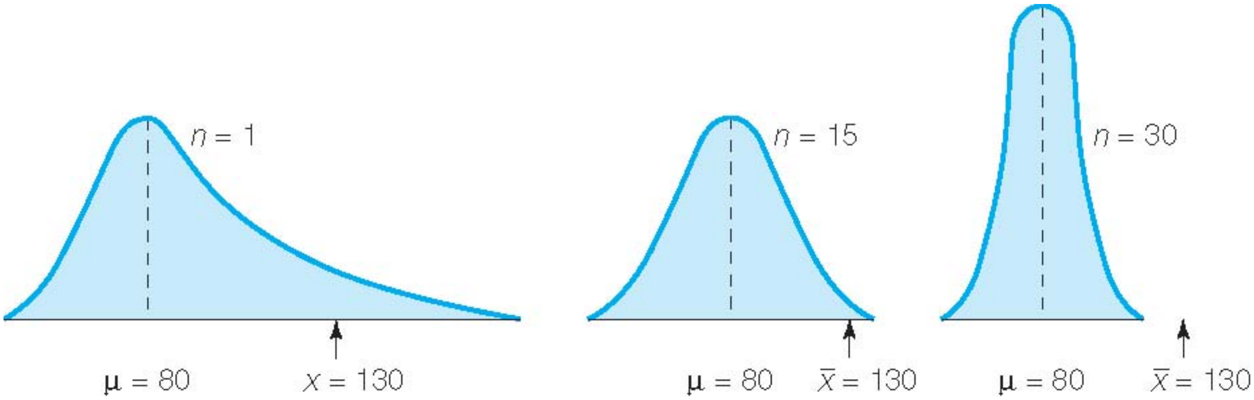

Figure 10.5 Distribution of scores of a population (*n* = 1) and sampling distributions of means when samples are randomly selected from that population and the sample sizes are *n* = 15 and *n* = 30, respectively. Note that the parent distribution is markedly skewed. As we increase *n*, the distribution of sample means tends toward normality and the dispersion of sample means decreases. (The Greek letter mu (μ) represents a population mean.)

It is a fortunate fact of statistical life that distributions of sample statistics often take on the form of other distributions with known mathematical properties. This permits us to use this known distribution as a frame of reference against which to evaluate a given sample statistic. Thus, knowing that the distribution of sample means tends toward normality when the sample size exceeds 30 permits us to evaluate the relative frequency of a sample mean in terms of the normal distribution. We are then able to label certain events or outcomes as common, others as somewhat unusual, and still others as rare. For example, note that in Figure 10.5 a score of 130 is fairly common when n = 1, whereas a mean of 130 is unusual when n = 15, and rare when n = 30. If we find that the occurrence of an event or the outcome of an experiment is rare, we conclude that nonchance factors (such as the experimental variable) are responsible for or have caused this rare or unusual outcome.

Table 10.6 shows the sampling distribution of a coin-tossing experiment when a single coin is tossed 12 times or when 12 coins are tossed once. It is the sampling distribution of a binomial or two-category variable when $n = 12$, and the probability of each elementary event (a head or a tail) is equal to 1/2.

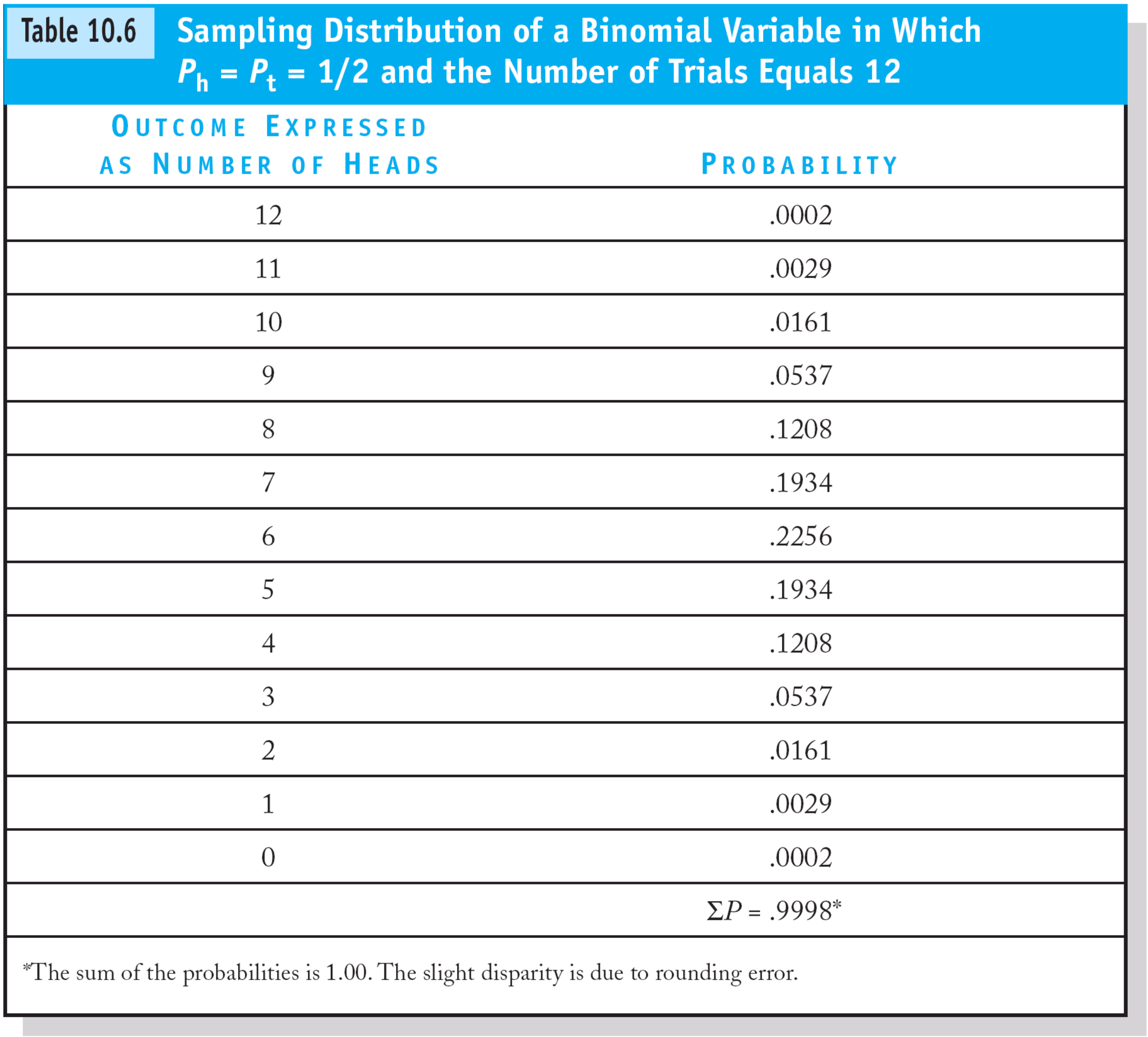

Table 10.6 Sampling Distribution of a Binomial Variable in Which $P_h = P_t = 1/2$ and the Number of Trials Equals 12

Outcome Expressed as Number of Heads	Probability
12	.0002
11	.0029
10	.0161
9	.0537
8	.1208
7	.1934
6	.2256
5	.1934
4	.1208
3	.0537
2	.0161
1	.0029
0	.0002
	$\Sigma P = .9998$*

*The sum of the probabilities is 1.00. The slight disparity is due to rounding error.

This sampling distribution provides a frame of reference for answering questions about possible outcomes of experiments. Is 12 out of 12 heads a rare outcome? Yes, exceedingly rare. It occurs, by chance, about twice out of every 10,000 repetitions of the experiment. Note that 0 out of 12 is equally rare. What about 7 heads out of 12? This would be not at all unusual. This outcome will occur about 19 times out of every 100 repetitions of the coin-tossing experiment. Can we define *rare* more precisely? This will require an agreement among fellow scientists. If the event in question or an event more unusual would have occurred less than 5% of the time, most psychological researchers are willing to make the judgment that the outcome is rare and ascribe it to nonchance factors. In other words, they reject H_o and assert H_1. This cutoff point for inferring the operation of nonchance factors is referred to as the 0.05 significance level. When we reject the null hypothesis at the 0.05 level, it is conventional to refer to the outcome of the experiment as statistically significant at the 0.05 level.

Other scientists, more conservative about rejecting H_o, prefer to use the 0.01 significance level, or the 1% significance level. When the observed event or one that is more deviant would occur by chance less than 1% of the time, we assert that the results are due to nonchance factors. It is conventional to refer to the results of the experiment as statistically significant at the 0.01 level.

The level of significance set by the investigator for rejecting H_o is known as the **alpha (α) level.** When we employ the 0.01 significance level, $\alpha = 0.01$. When we use the 0.05 significance level, $\alpha = 0.05$.

Let us look at a few examples of this statistical decision-making process:

- Jeanette conducted a study in which she used $\alpha = 0.05$. Upon completing her statistical analysis, she found that the probability of obtaining such an outcome simply by chance was 0.02. Her decision? She rejects H_o and asserts that the experimental variable had an effect on the dependent measure.
- Roger obtained a probability of 0.06, using $\alpha = 0.05$. Because his results failed to achieve the 0.05 cutoff point, he does not reject the null hypothesis. Note that he cannot claim to have proved the null hypothesis, nor should he claim "there is a trend toward significance." Once a significance level is set, its boundaries should be considered as quite rigid and fixed.
- A research team set alpha at 0.01 and found the probability of their obtained outcome to be $p = 0.03$. They fail to reject H_o because the probability of this outcome is greater than α.

The statistical decision-making process is summarized in Table 10.7.

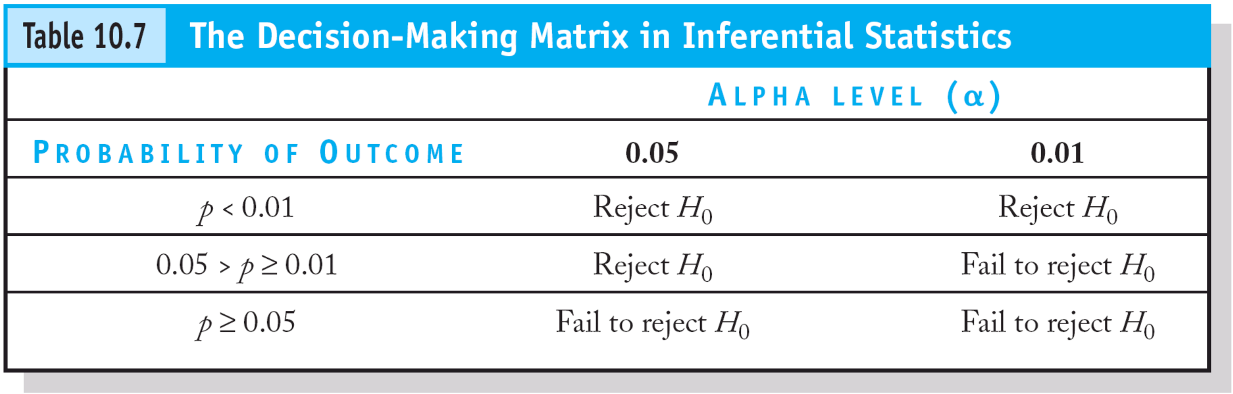

Table 10.7 The Decision-Making Matrix in Inferential Statistics

	Alpha level (α)	
Probability of Outcome	**0.05**	**0.01**
$p < 0.01$	Reject H_0	Reject H_0
$0.05 > p \geq 0.01$	Reject H_0	Fail to reject H_0
$p \geq 0.05$	Fail to reject H_0	Fail to reject H_0

Many of the tables that researchers and statisticians use do not provide probability values for the sampling distributions to which they refer. Rather, they present **critical values** that define the **region of rejection** at various levels of α. The region of rejection is that portion of the area under a curve that includes those values of a test statistic that lead to rejection of the null hypothesis. However, Roger's results are so close to the cutoff point, he would be wise to consider repeating the study and increasing the *n*, if still possible.

To illustrate, the two curves in Figure 10.6 show the regions for rejecting H_o when the standard normal curve is used.

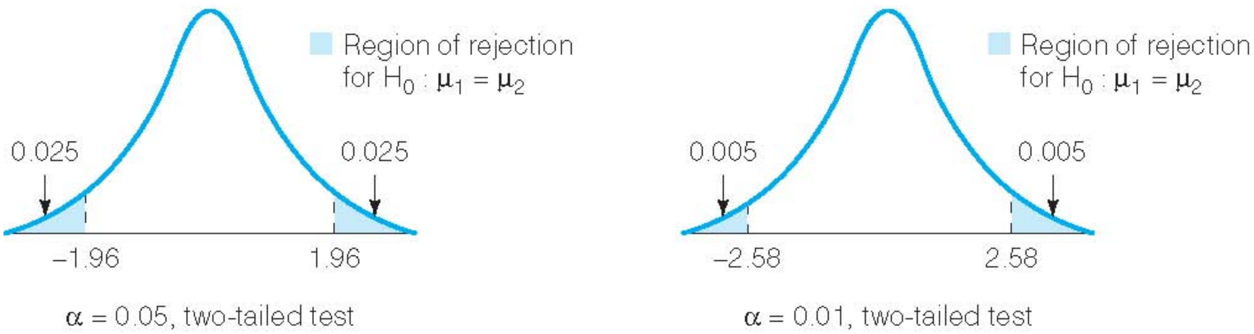

Figure 10.6 Region of rejection at α = 0.05 and α = .01 under the standard normal curve. If the obtained statistic is greater than 1.96 or less than –1.96 and α = 0.05, we reject H_o. If α = 0.01 and the test statistic exceeds the absolute value of 2.58, we reject H_o.

Let's look at a few examples:

- If α = 0.05 and the test statistic equals 1.43, we fail to reject H_o because the test statistic does not achieve the critical value.
- If α = 0.01 and the test statistic equals 2.83, we reject H_o because the test statistic is in the region of rejection.

- If $\alpha = 0.05$ and the test statistic equals 2.19, we reject H_o because the test statistic is in the region of rejection for $\alpha = 0.05$.

You may have noticed that in the above discussion we assumed a two-tailed test of significance. That is, the probability associated with the alpha level was split into the left and right tails of the sampling distribution to determine the critical values for the significance test. Most researchers use two-tailed tests, and most statistical software programs conduct two-tailed tests as the default. The two-tailed test is slightly more conservative, and it lets us reject the null hypothesis if the treatment increases *or* decreases scores on the dependent variable. However, if you are confident that your independent variable will only affect scores on the dependent variable in a particular direction, you may decide to use a one-tailed test of significance. In that case, you would place all of the probability associated with your alpha level into one tail of the sampling distribution. This results in a slightly more powerful analysis but risks the chance that the effect may go in the direction that is opposite of your prediction.

Type I Errors, Type II Errors, and Statistical Power

As we saw earlier in the chapter, we can make two types of statistical decisions: reject H_o when the probability of the event of interest achieves an acceptable α level (usually $p < 0.05$ or $p < 0.01$), or fail to reject H_o when the probability of the event of interest is greater than α. Each of these decisions carries an associated risk of error.

If we reject H_o (conclude that H_o is false) when in fact H_o is true, we have made the error of falsely rejecting the null hypothesis. This type of error is called a **Type I error**. If we fail to reject H_o (we do not assert the alternative hypothesis) when in fact H_o is false, we have made the error of falsely accepting H_o. This type of error is referred to as a **Type II error.**

Let's look at a few examples:

- H_o: $\mu_1 = \mu_{2,}$ $\alpha = 0.05$. Obtained $p = 0.03$.
  Statistical decision: Reject H_o. Actual status of H_o: True.
  Error: Type I (rejecting a true H_o).
- H_o: $\mu_1 = \mu_{2,}$ $\alpha = 0.05$. Obtained $p = 0.04$.
  Statistical decision: Reject H_o. Actual status of H_o: False.
  Error: None (conclusion is correct).
- H_o: $\mu_1 = \mu_{2,}$ $\alpha = 0.01$. Obtained $p = 0.10$.
  Statistical decision: Fail to reject H_o. Actual status of H_o: False.
  Error: Type II (failing to reject a false H_o).

- H_o: $\mu_1 = \mu_{2,}$ $\alpha = 0.01$. Obtained $p = 0.06$.

  Statistical decision: Fail to reject H_o. Actual status of H_o: True.

  Error: None (conclusion is correct).

You should know that a Type I error can be made only when H_o is true because this type of error is defined as the mistaken rejection of a true hypothesis. The probability of a Type I error is given by α. Thus, if α = 0.05, about 5 times out of 100 we will falsely reject a true null hypothesis. In contrast, a Type II error can be made only when H_o is false because this type of error is defined as the mistaken retention of a false hypothesis.

Table 10.8 summarizes the probabilities associated with retention or rejection of H_o depending on the true status of the null hypothesis.

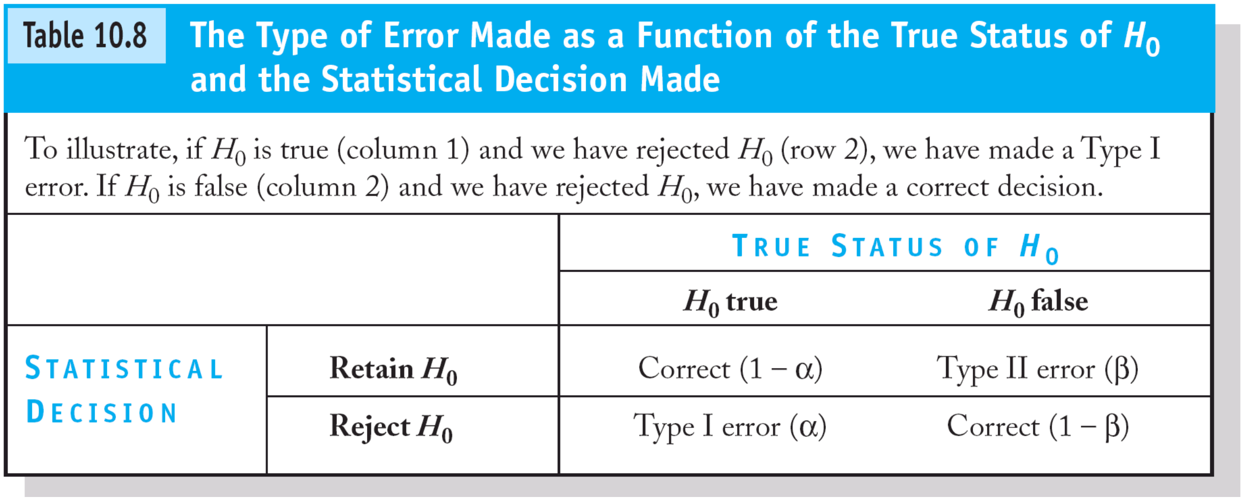

Table 10.8 **The Type of Error Made as a Function of the True Status of H_0 and the Statistical Decision Made**

To illustrate, if H_0 is true (column 1) and we have rejected H_0 (row 2), we have made a Type I error. If H_0 is false (column 2) and we have rejected H_0, we have made a correct decision.

		True Status of H_0	
		H_0 true	**H_0 false**
Statistical Decision	**Retain H_0**	Correct (1 – α)	Type II error (β)
	Reject H_0	Type I error (α)	Correct (1 – β)

The probability of a Type II error must be obtained by calculation. It is beyond the scope of this book to delve into the calculation of β probabilities. Interested students may consult a more advanced statistics book. The concept of this error is important because, among other things, it relates to the economics of research. It would make little sense to expend large amounts of funds, tie up laboratory space and equipment, and devote hours of human effort to the conceptualization, conduct, and statistical analysis of research if, for example, the β probability were as high as 90 percent. This would mean that the probability of making a correct decision—rejecting the false null hypothesis—would be only 10 percent. It would hardly seem worth the effort. This probability—the probability of correctly rejecting the null hypothesis when it is false—is known as the **power** of the test. Power is defined as 1 – β. In other words, the power of a statistical test is the ability of the test to detect an effect of the IV when it is there.

There is a further risk when conducting research in which the power is low. A failure to find a significant difference may cause a researcher to prematurely abandon a promising line of experimentation. As a consequence, potentially important discoveries may never be made because the researcher relegated a seminal idea to the junk heap.

Clearly, one of the goals of the careful researcher must be to reduce the probability of β error and, thereby, increase the power of the test. A number of factors influence statistical power. Among them are sample size, alpha level, and precision in estimating experimental error. Fortunately, all are under the control of the experimenter.

Other things being equal, as you increase the sample size, you increase the power of your statistical test. In research in which the cost per participant is low, increasing the sample size may be an attractive way to boost power. However, the relationship between sample size and power is one of diminishing returns. Beyond a certain point, further increases in sample size lead to negligible increases in power.

As the α level is decreased, we decrease the probability of a Type I error and increase the probability of a Type II error. Conversely, as the α level is increased, we increase the probability of a Type I error and decrease the probability of a Type II error. Because the power of the test is inversely related to the probability of a Type II error (power increases as the probability of a Type II error decreases), it follows that the power can be increased by setting a higher alpha level for rejecting H_o.

Balanced against this is the fact that increasing the α level also increases the probability of falsely rejecting a true null hypothesis. The researcher must decide which of these risks is more acceptable. If the consequences of making a Type I error are serious (claiming that a chemical compound cures a serious disease when it does not), it is desirable to set a low α level. However, the commission of a Type II error can also have serious consequences, as when failure to reject the null hypothesis is treated as if the null hypothesis has been proved. Thus, a chemical compound “proved safe after exhaustive testing” could lead to the introduction of a lethal compound into the marketplace.

The third factor, control over the precision in estimating experimental error, is the one that should receive the most attention from a careful researcher. Any steps that lead to increased precision of our measurement of experimental error will also increase the power of the test. We can increase our precision in measuring experimental error in numerous ways, including improving the reliability of our criterion measure, standardizing the experimental technique, and using correlated measures. In a correlated samples design, for example, the power of the test will increase as the correlation between paired measures increases. We had more to say about this important topic in earlier chapters, where we noted that a feature of various research designs was the degree of precision in estimating both the effects of experimental treatments and the error variance.

In our consideration of Type I and Type II errors, it is important to remember that only one type of error is possible in any given analysis and that researchers never know whether one of these errors has occurred (if they knew, then they would obviously change their conclusion to avoid the error!). Thus, it is critical for the researcher to consider all of the factors discussed in the previous paragraphs to increase confidence that an erroneous conclusion will not be made.

Effect Size

Generally speaking, we want a high degree of power in our experiment. That is, we want to be able to detect a difference if one in fact exists. As noted above, we can increase the power of our experiment by increasing the sample size. How far should we take this? Is it the case that the larger the sample the better? Not exactly. At some point, we can use such large samples and have such high power that we begin to detect statistically significant differences that are, in fact, practically meaningless (the differences between the treatment group and control group are small and trivial). Significance tests (such as the *t* test or ANOVA) provide a way to decide whether an effect exists, but do not provide a clear indication of how large an effect is. Measures of **effect size** provide an indication of the size of an effect (strength of the IV) and, therefore, provide important additional information. Measuring effect size is a helpful technique for separating statistical significance and practical significance.

Several different measures of effect size are available. The APA Publication Manual (p. 25) provides a list and encourages authors to include an appropriate measure of effect size in the results section of a research report. You can use Table 10.9 as a guide to select the most appropriate measure of effect size for some of the most basic statistical analyses.

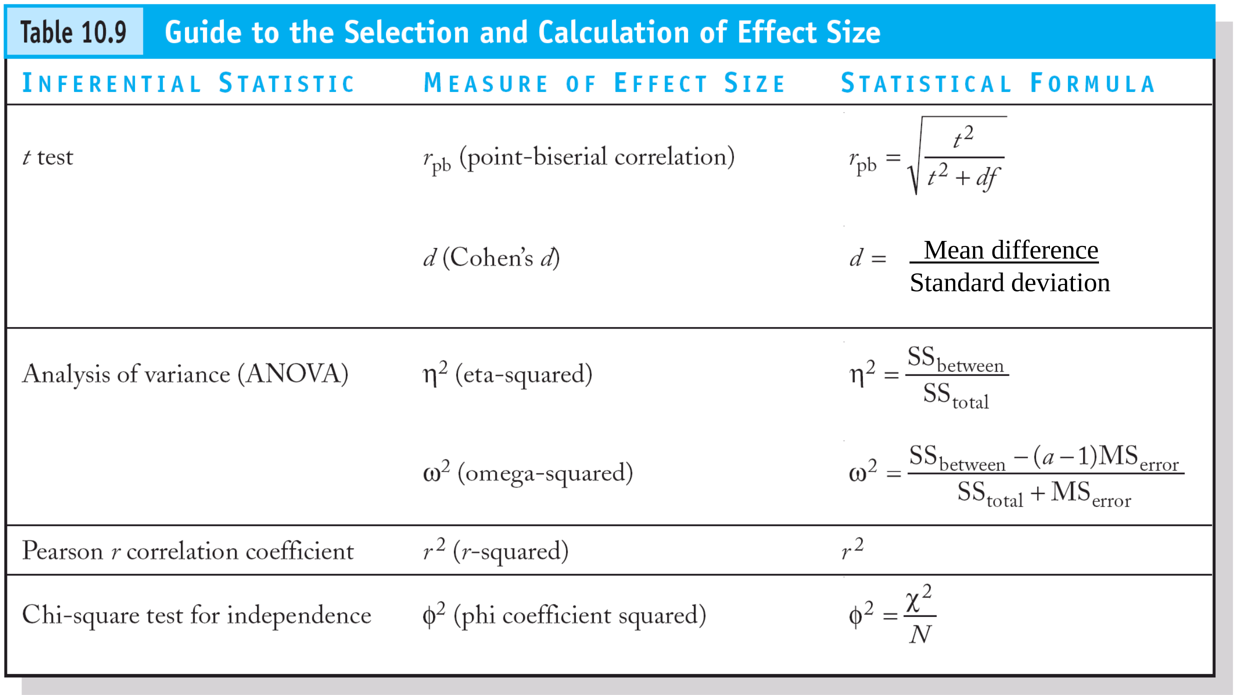

Table 10.9 Guide to the Selection and Calculation of Effect Size

Inferential Statistic	Measure of Effect Size	Statistical Formula
t test	r_{pb} (point-biserial correlation)	$r_{\text{pb}} = \sqrt{\dfrac{t^2}{t^2 + df}}$
	d (Cohen's *d*)	$d = \dfrac{\text{Mean difference}}{\text{Standard deviation}}$
Analysis of variance (ANOVA)	η^2 (eta-squared)	$\eta^2 = \dfrac{\text{SS}_{\text{between}}}{\text{SS}_{\text{total}}}$
	ω^2 (omega-squared)	$\omega^2 = \dfrac{\text{SS}_{\text{between}} - (a-1)\text{MS}_{\text{error}}}{\text{SS}_{\text{total}} + \text{MS}_{\text{error}}}$
Pearson *r* correlation coefficient	r^2 (*r*-squared)	r^2
Chi-square test for independence	ϕ^2 (phi coefficient squared)	$\phi^2 = \dfrac{\chi^2}{N}$

Your obvious next question is, What do these numbers tell you? Without getting into deep statistical explanations (you can find these in more advanced statistics books), let's summarize what are considered small, medium, and large effect sizes. For the point-biserial correlation, values less than .30 are considered small, values of .30–.50 are considered moderate, and values greater than .50 are considered large (Thompson & Buchanan, 1979). For Cohen's *d*, values of .20–.50 are considered small, values of .50–.80 are considered moderate, and values greater than .80 are considered large (Cohen, 1977). For η^2 and ω^2, the value indicates the proportion of variance (0.0–1.0) in the dependent variable that can be explained by the levels of the independent variable. For r^2 and ϕ^2, the value indicates the proportion of variance in the criterion variable that can be explained by the predictor variable. The larger the proportion of variance that can be explained, the larger is the effect size.

Meta-analysis

Whereas measures of effect size provide important information for a particular study, **meta-analysis** is a statistical technique that provides an indication of the size of an effect across the results of many studies. As different researchers continue to explore a particular research question, published studies begin to accumulate. After some period of time, it is common for someone to publish a review article to summarize the different studies that have been done and their findings. These review articles often reveal mixed findings; that is, some studies report effects, and some do not.

Meta-analysis provides a statistical method for combining the effects across studies to reach a decision regarding whether a particular independent variable affects a particular dependent variable. Essentially, a measure of effect size is calculated for each study and then weighted according to the sample size and quality of the study. These measures are then averaged across studies to produce an overall effect size. This overall value provides a measure of effect size in standard deviation units. Thus, a meta-analysis that produced an effect size of .33 would indicate that the size of the effect is one-third of the average standard deviation across studies.

Let's examine an actual meta-analysis. Do you believe that prevention programs targeted to children in school can reduce the incidence of child sexual abuse? A variety of such programs have been developed, implemented, and evaluated. Two researchers conducted a meta-analysis of 27 such studies (Davis & Gidycz, 2000). They reported an overall effect size of 1.07, which indicated that children who participated in prevention programs performed more than 1 standard deviation higher on outcome measures. Based on their analysis, they also concluded that long-term programs that required active involvement from the children were more effective. Such analyses effectively summarize a body of literature and direct further research.

Parametric Versus Nonparametric Analyses

Many data are collected in the behavioral sciences that either do not lend themselves to analysis in terms of the normal probability curve or fail to meet the basic assumptions for its use. For example, researchers explore many populations that consist of two categories—for example, yes/no, male/female, heads/tails, right/wrong. Such populations are referred to as dichotomous, or two-category, populations. Other populations consist of more than two categories—for example, political affiliation or year in college. (We dealt with these in Chapter 5 under the heading Nominal Scale.) Other data are best expressed in terms of ranks—that is, on ordinal scales. When comparing the attributes of objects, events, or people, we are often unable to specify precise quantitative differences. However, we are frequently able to state ordered relationships—for example, Event A ranks the highest with respect to the attribute in question, Event B the second highest, and so on. In addition to equivalence and nonequivalence, then, the mathematical relationships germane to such data are "greater than" (>) and "less than" (<). The relationship $a > b$ may mean that *a* is taller than *b,* of higher rank than *b*, more prestigious than *b,* prettier than *b,* and so on. Similarly, the relationship $a < b$ may mean that *a* is less than *b,* of lower rank than *b,* less prestigious than *b,* and so on.

Finally, many data collected by psychologists are truly quantitative. They may be meaningfully added, subtracted, multiplied, and divided. These data are measured on a scale with equal intervals between adjacent values—that is, an interval or ratio scale. For example, in a timed task, a difference of 1 second is the same throughout the time scale. Most commonly, **parametric statistics** are used with such variables. Parametric tests of significance include the *t* test and analysis of variance (ANOVA).

Parametric tests always involve two assumptions. One is that the populations for the dependent variable are normally distributed. That is, the distribution of scores conforms to a bell-shaped distribution rather some other shape of distribution (such as positively or negatively skewed, or multimodal). The risk of a nonnormal distribution is particularly great with small *n*'s. With large *n*'s, the sampling distributions of most statistics approach the normal curve even when the parent distributions deviate considerably from normality. The second assumption is termed **homogeneity of variance.** Homogeneity of variance is the assumption that the populations for the dependent variable have equal variances. That is, the degree to which the scores are spread out around the mean is the same in the populations represented by the groups in the study. It is worth noting that parametric tests are robust. As long as the assumptions are not seriously violated, the conclusions derived from parametric tests will be accurate.

For data measured on a nominal scale, an ordinal scale, an interval scale with a nonnormal distribution, or a ratio scale with a nonnormal distribution, the investigator should use **nonparametric statistics** for the analysis. For data on a nominal scale, nonparametric analyses include the chi-square test for goodness of fit, the chi-square test for independence, the binomial test, and the median test. For data on an ordinal scale or for data on an interval/ratio scale that do not satisfy the assumption of normality, nonparametric analyses include the Wilcoxon test and the Mann–Whitney test. (It is beyond the scope of this text to review the many nonparametric tests that are available. If you wish to read further, you may consult a statistics textbook.)

Selecting the Appropriate Analysis: Using a Decision Tree

As you can see, deciding on the appropriate descriptive and inferential statistics for a given study is not easy and involves consideration of several factors. To aid in these decisions, we have included several decision trees. Figure 10.7 illustrates how to choose a descriptive statistic. Figure 10.8 illustrates how to choose a parametric statistic to evaluate group differences. Figure 10.9 illustrates how to choose a nonparametric statistic to evaluate group differences. Finally, Figure 10.10 illustrates how to choose a statistic to measure the relationship between two variables.

Choosing Descriptive Statistics: A Decision Map

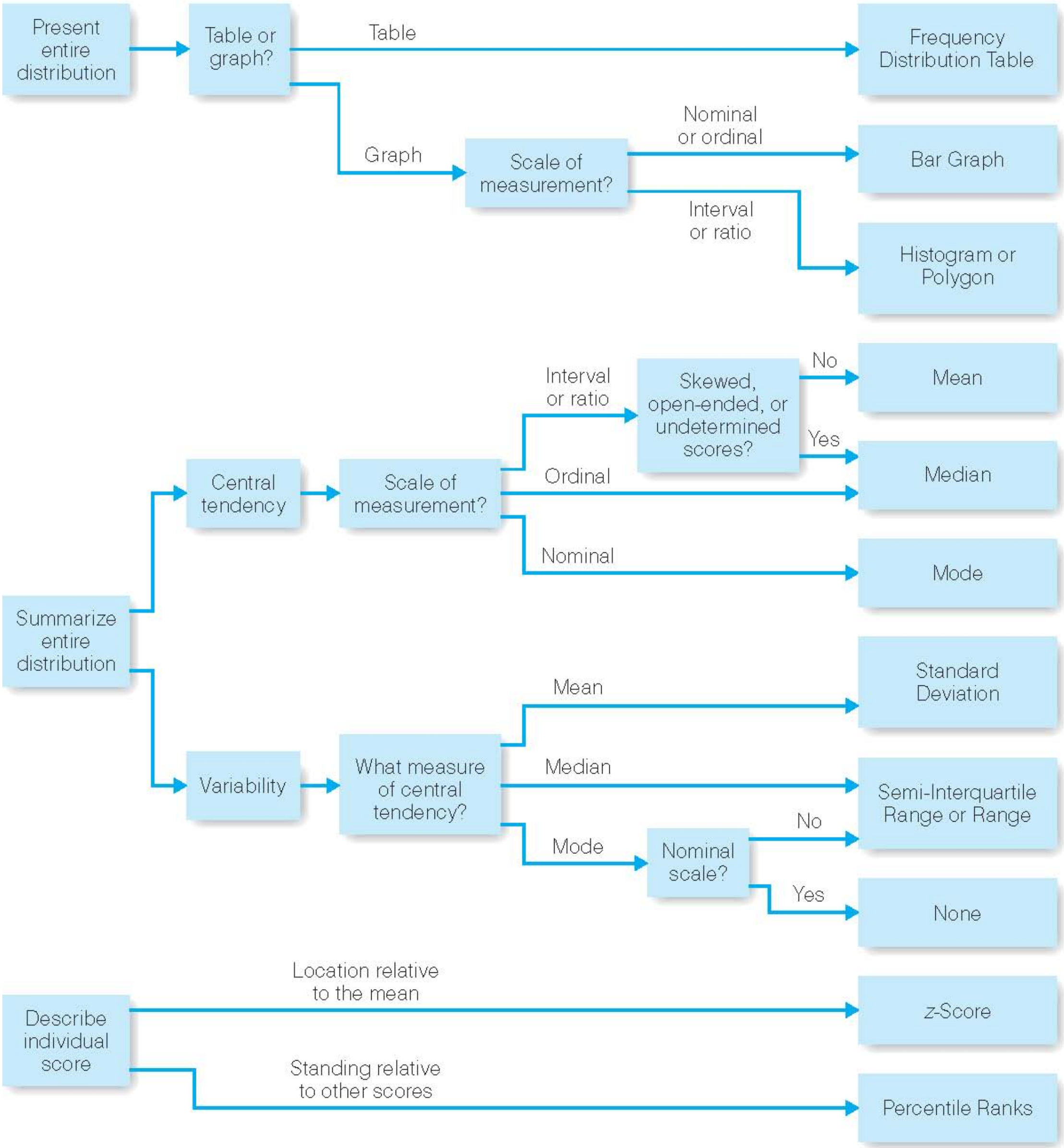

Figure 10.7 Decision tree to select a descriptive statistic.

Choosing a Parametric Test: A Decision Map for Making Inferences About Population Means or Mean Differences

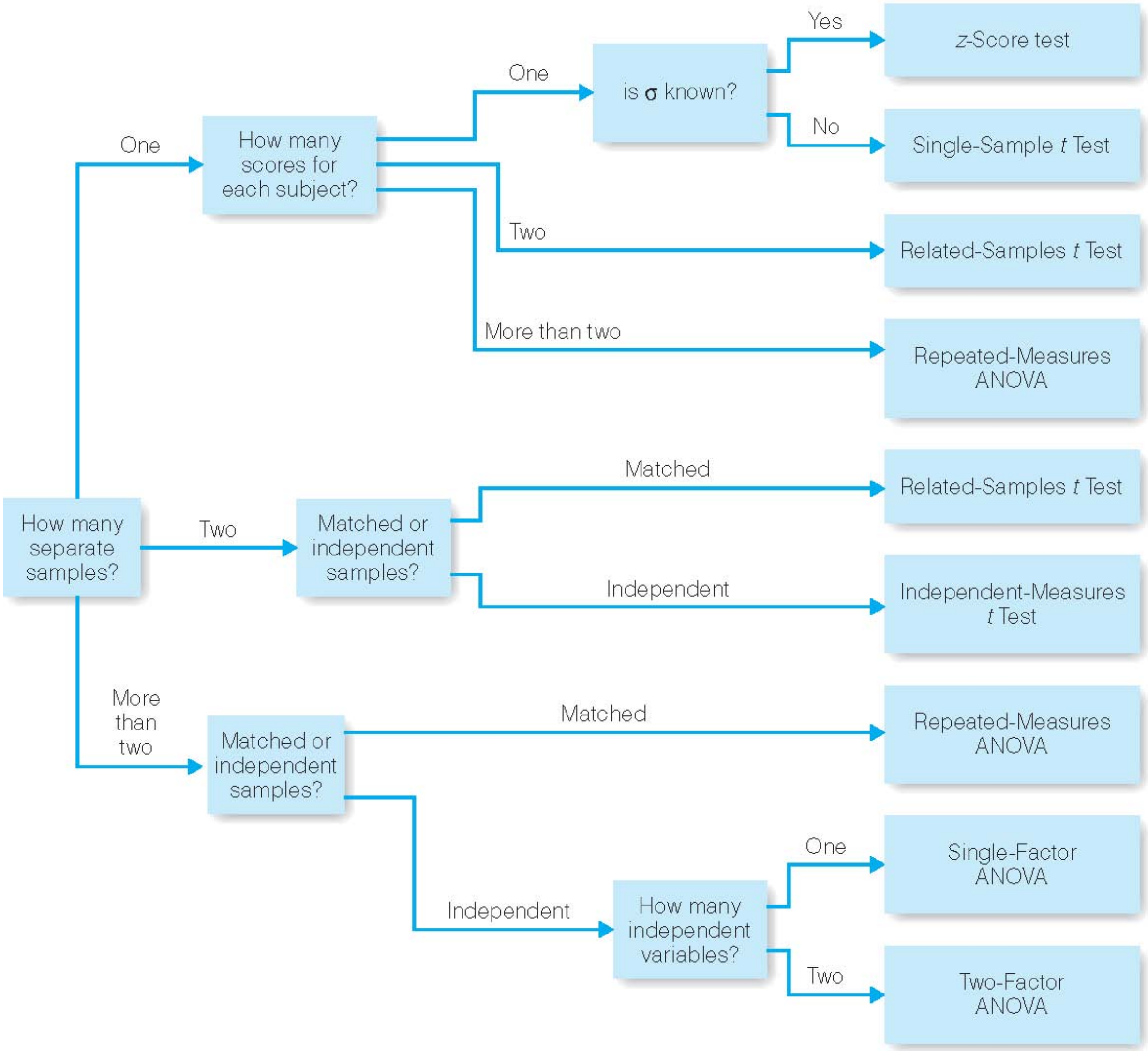

Figure 10.8 Decision tree to select a parametric statistic to evaluate group differences.

Choosing a Nonparametric Test: A Decision Map for Making Inferences About Populations or Population Differences

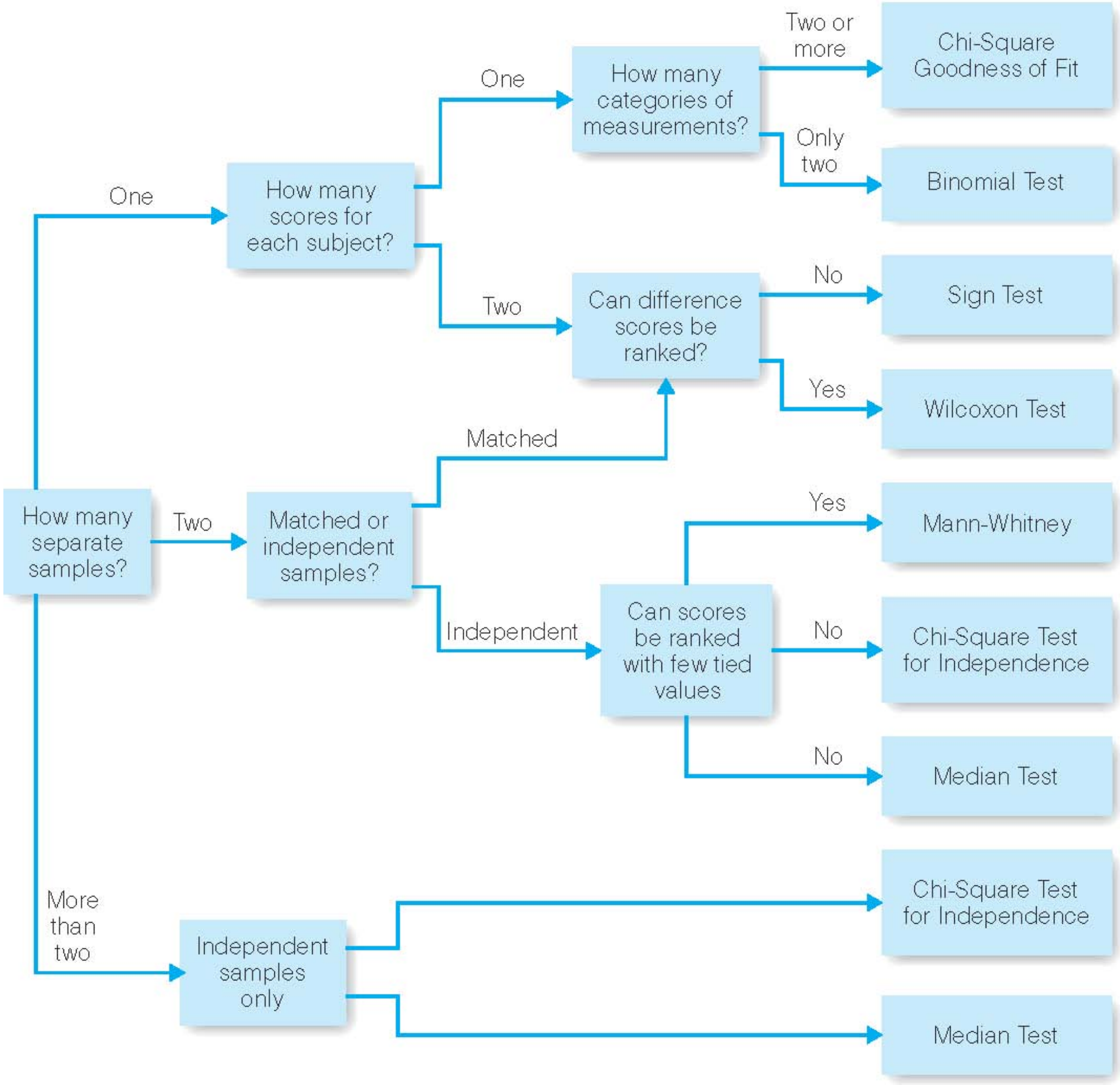
How many separate samples?
One
How many scores for each subject?
One
How many categories of measurements?
Two or more
Chi-Square Goodness of Fit
Only two
Binomial Test
Two
Can difference scores be ranked?
No
Sign Test
Yes
Wilcoxon Test
Two
Matched or independent samples?
Matched
Independent
Can scores be ranked with few tied values
Yes
Mann-Whitney
No
Chi-Square Test for Independence
No
Median Test
More than two
Independent samples only
Chi-Square Test for Independence
Median Test

Figure 10.9 Decision tree to select a nonparametric statistic to evaluate group differences.

Choosing a Measure of Relationship Between Two Variables: A Decision Map

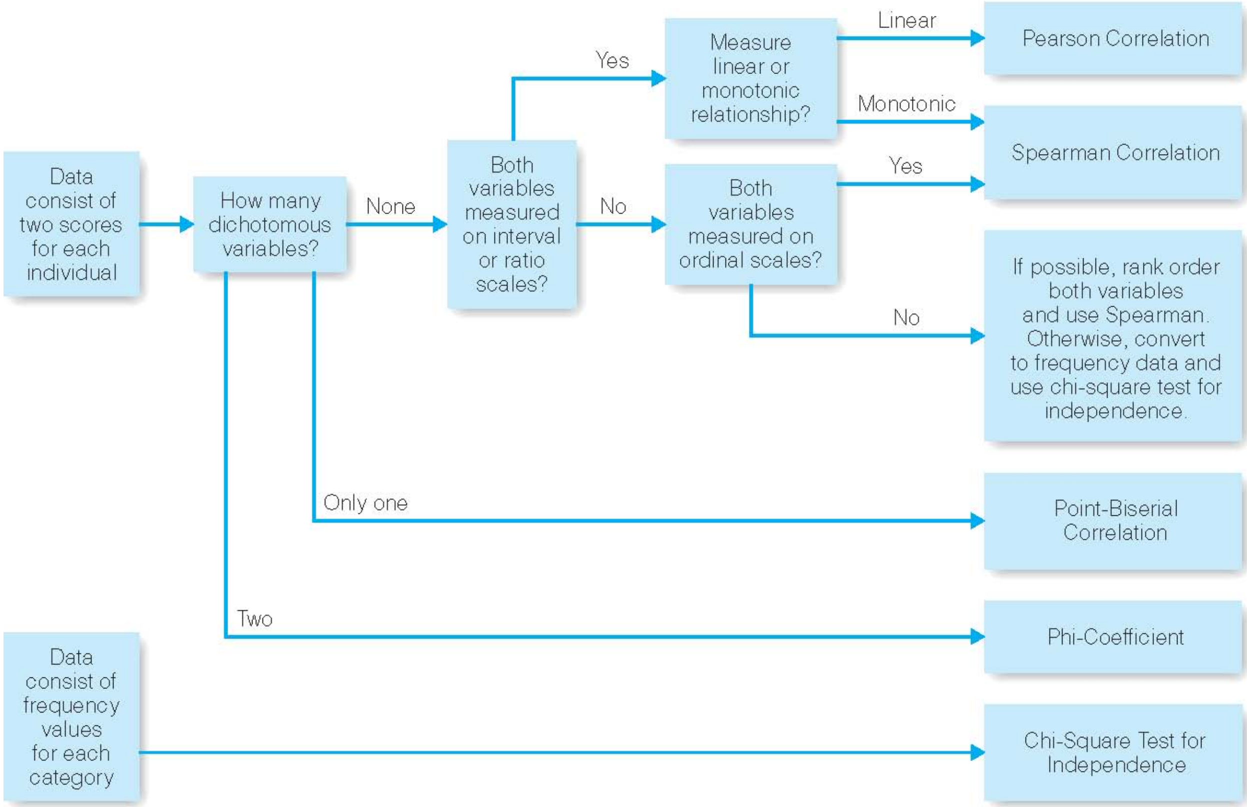
Data consist of two scores for each individual
How many dichotomous variables?
None
Both variables measured on interval or ratio scales?
Yes
Measure linear or monotonic relationship?
Linear
Pearson Correlation
Monotonic
Spearman Correlation
No
Both variables measured on ordinal scales?
Yes
No
If possible, rank order both variables and use Spearman. Otherwise, convert to frequency data and use chi-square test for independence.
Only one
Point-Biserial Correlation
Two
Phi-Coefficient
Data consist of frequency values for each category
Chi-Square Test for Independence

Figure 10.10 Decision tree to select a statistic to measure the relationship between two variables.

Using Statistical Software

Computers have greatly increased the efficiency with which research data can be analyzed. Researchers rarely calculate statistics by hand. Typically, data are entered into some form of spreadsheet. Statistics can then be calculated by database software (such as Excel, Access, Lotus123, or Quattro pro) or by statistical software (such as SPSS, SAS, SYSTAT, or STATVIEW). The particular software that you use as a student researcher will depend on which software is available at your university and which software is familiar to your instructor. Each software program has advantages and disadvantages. Although we do not want to recommend a particular program, we do suggest that you learn at least one of them.

The statistical output that we present in the next several chapters is presented in generic form rather than the format of a particular software package. Whichever package you use, you should be able to locate the same information in the output.

A caution is in order. The ease with which inferential statistics can be calculated by the computer creates a temptation to simply enter the data and click on the button to perform the inferential analysis so that a conclusion statement can be written. Be sure to perform descriptive statistics first. Get a strong feel for your data by calculating measures of central tendency and measures of variability. Let your data talk to you through various graphs that not only depict summary statistics, but also depict the distribution of raw scores. These graphs can show outliers and patterns in the data that would be overlooked by the inferential statistic and may, in fact, create inferential analyses that are misleading.

Case Analysis

One recent area of research in the behavioral sciences involves the positive impact that natural environments have on mental, social, and physical health. This research on “restorative environments” has implications for the design of homes, office space, and hospital recovery rooms. You decide to study this phenomenon by comparing two recovery rooms at your local hospital. The rooms are identical except that the view from one looks out over a park with trees and grass, and the view from the other simply shows a brick wall of the building. Patients recovering from routine surgeries are randomly assigned to one of the recovery rooms, and you record the number of days of recovery prior to discharge from the hospital. Because you understand the concept of confounding variables, you make sure that the patients and nurses are unaware of the experiment. The data are shown in the following table:

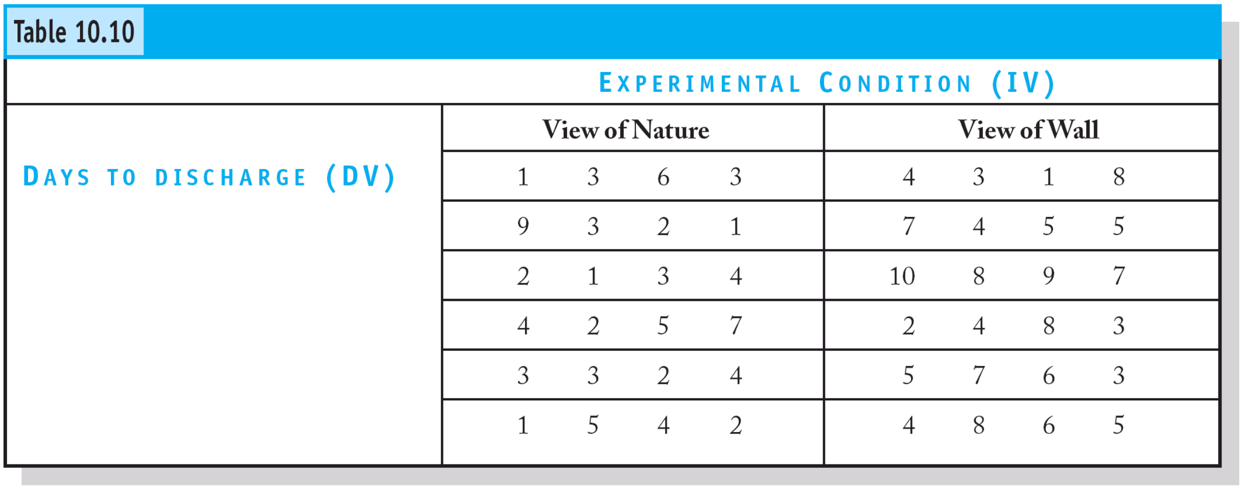

Table 10.10

	Experimental Condition (IV)							
	View of Nature				View of Wall			
Days to discharge (DV)	1	3	6	3	4	3	1	8
	9	3	2	1	7	4	5	5
	2	1	3	4	10	8	9	7
	4	2	5	7	2	4	8	3
	3	3	2	4	5	7	6	3
	1	5	4	2	4	8	6	5

Critical Thinking Questions

1. Which measure of central tendency would you use? Why?
2. Which measure of variability would you use? Why?

3. Which type of graph would you use to illustrate the average days to recovery as a function of the type of view? Why?
4. Which type of inferential analysis would you use to determine whether there was any effect of the type of view on recovery rate? Why?

General Summary

Statistics provide a way to summarize and interpret behavioral observations. Descriptive statistics, such as measures of central tendency, measures of variability, and graphical representations of data, summarize observations. Central tendency can be measured with the mean, median, or mode. The mean, the most common measure, has the advantage of considering the specific value of every score in the distribution. The median is often appropriate when the distribution of scores is skewed or contains a few extreme scores. The mode can be used when the data are measured on a nominal scale of measurement. Variability is most often measured using standard deviation, which provides an indication of how far, on average, scores are from the mean.

Inferential statistics, such the *t* test and analysis of variance, provide a way to interpret the data and arrive at conclusions. The interpretation is based on a calculation of probabilities. Thus, inferential statistics never prove anything. Rather, they allow the researcher to draw a conclusion with some degree of confidence. Because probabilities are involved, there always exists the possibility of a Type I error or a Type II error. A Type I error occurs when the researcher concludes that an effect exists when, in fact, it does not. A Type II error occurs when the researcher concludes that no effect exists when, in fact, it does. Increasing the power of your study can reduce the chance of a Type II error.

Significance tests are often supplemented with a measure of effect size. Such measures provide an indication of how large an effect is and, thus, whether a significant effect is, in fact, a meaningful effect. Similarly, meta-analyses provide a way to summarize the size of effects across multiple studies so that an entire body of literature can be interpreted.

Inferential statistics come in the form of parametric statistics and nonparametric statistics. Parametric statistics are more powerful but require interval or ratio data and also require that certain assumptions about the nature of the data be met. A variety of factors must be considered to determine the most appropriate statistical analysis. Decision trees can be used to make such decisions, and computer programs can be used to perform the statistical analyses.

In the next chapter, we will begin to explore particular types of experimental design.

Detailed Summary

1. Statistics provide an objective approach to understanding and interpreting the behaviors that we observe and measure.
2. Descriptive statistics are used to describe and summarize data. They include measures of central tendency (mean, median, mode) and measures of variability (range, variance, standard deviation). Descriptive statistics are often presented in the form of graphs.
3. Measures of central tendency provide an indication of the "center" of the distribution of scores, whereas measures of variability provide an indication of the spread of scores.
4. The mean is the arithmetic average of a set of scores. It considers the precise value of each score in the distribution. It is the preferred measure of central tendency for interval or ratio data unless the distribution of scores is skewed by outlier (extreme) scores.
5. The median is the middle point in the distribution. That is, half of the scores are above the median, and half are below the median.
6. The mode is the most frequent score in the distribution—that is, the score that occurs most often.
7. The range is the number of units between the highest and lowest scores in the distribution. Because the range only considers the values of the two most extreme scores, it is less stable than other measures of variability and may not adequately reflect the overall spread of scores.
8. The variance is the average squared deviation of scores from the mean. The square root of variance is the standard deviation. Thus, standard deviation reflects, on average, how far scores are from the mean. It is the preferred measure of variability.
9. Many variables result in a distribution of scores that is normal in shape. This observation, along with the calculated mean and standard deviation, provide a wealth of additional information regarding the proportion of scores in particular parts of the distribution or the probability (or percentage chance) of obtaining a particular score in the distribution.
10. Variability is an essential concept in behavioral research because most of the principles of good research design involve methods to reduce variability due to extraneous variables so that variability due to systematic sources (our independent variables) is clear.
11. Researchers should make extensive use of tables and graphs to summarize data. Such techniques provide the researcher with a better "feel" for the data.
12. We usually conduct research on samples of participants and then want to draw conclusions about populations. Inferential statistics are tools used to make such inferences.
13. The conclusions made using inferential statistics are based on probabilities—specifically, the probabilities that certain events would occur simply by chance. Thus, our research hypotheses are never *proven* correct or incorrect. They are either retained or rejected based on probabilities.

14. The null hypothesis typically states that there is no difference in population parameters (usually population means), whereas the alternative hypothesis typically states that there is a difference in population parameters. The null hypothesis is the one that is statistically tested and either retained or rejected, whereas the alternative hypothesis usually reflects the researcher's expectation.
15. The frame of reference for statistical decision making is provided by the sampling distribution of a statistic. A sampling distribution is a theoretical probability distribution of the possible values of some sample statistic that would occur if we were to draw all possible samples of a fixed size from a given population.
16. If the probability of obtaining a sample statistic by chance is very rare, very unlikely, less than our alpha level (often 0.05), then we conclude that the sample did not come from the population and that our independent variable had a significant effect (that is, we reject the null hypothesis).
17. Power is the probability of finding a certain size effect assuming that it, in fact, exists. Power can be increased by increasing sample size and by using control techniques to reduce extraneous variability.
18. Because all conclusions are based on probabilities, our conclusions can, in fact, be wrong. If we conclude that there is an effect and there really is not, then we have made a Type I error. If we conclude that there is no effect and there really is one, then we have made a Type II error. Good research designs and experimental control will reduce the chance of making these errors.
19. The decision to reject a null hypothesis does not reflect the *size* of an effect. Other statistics measure effect size, providing another valuable tool in data analysis.
20. A particular inferential technique called meta-analysis provides a statistical method for combining the effects across studies to reach a decision regarding whether a particular independent variable affects a particular dependent variable.
21. Parametric statistics are used when data are measured on an interval or ratio scale and meet a few additional assumptions regarding sample size and variability. Nonparametric statistics are used when data are measured on a nominal or ordinal scale or do not meet the assumptions of parametric statistics.
22. During data analysis, the researcher must decide on the most appropriate descriptive and inferential statistics. These decisions are not always easy, and flowcharts can be a useful aid.
23. Statistical software makes data analysis much more efficient and less prone to errors in calculation. However, it is the responsibility of the researcher to understand what the software is doing to the data and to not blindly click the mouse on a series of buttons.

Key Terms

alpha (α) level

alternative hypothesis (H_1)

critical values

effect size

frequency distribution

homogeneity of variance

mean

measure of central tendency

measure of variability

median

meta-analysis

mode

nonparametric statistics

null hypothesis (H_0)

parametric statistics

power

range

region of rejection

sampling distribution

standard deviation

Type I error

Type II error

variance

Review Questions / Exercises

Use the hypothetical study described in the Case Analysis above.

1. Either by hand or using a computer program, calculate the mean, median, and mode for each group. Comparing these values, what do you notice?
2. Either by hand or using a computer program, calculate the range and standard deviation for each group. Comparing these values, what do you notice?
3. Construct a graph similar to the one shown in Figure 10.4. Again, what do you notice?
4. In words, write both the null hypothesis and the alternative hypothesis.

5. Either by hand or using a computer program, calculate an appropriate inferential statistic that will test the null hypothesis. What is the probability of such a statistic assuming the null hypothesis is true? Do you reject or retain the null hypothesis?
6. Which type of statistical error could you be making with the statistical decision in question 5?
7. Either by hand or using a computer program, calculate an appropriate measure of effect size. What does this value indicate?
8. Based on all of the above information, write a conclusion statement for the study.