CHAPTER 1

Introduction to Statistics and Frequency Distributions

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain how you can be successful in this course
- Explain why many academic majors require a statistics course
- Use common statistical terms correctly in a statistical context
  - Statistic, parameter, sample, population, descriptive statistics, inferential statistics, sampling error, and hypothesis testing
- Identify the scale of measurement of a variable (nominal, ordinal, or interval/ratio)
- Determine if a variable is discrete or continuous
- Create and interpret frequency distribution tables, bar graphs, histograms, and line graphs
- Explain when to use a bar graph, histogram, and line graph
- Enter data into SPSS and generate frequency distribution tables and graphs

HOW TO BE SUCCESSFUL IN THIS COURSE

Have you ever read a few pages of a textbook and realized *you were not thinking about what you were reading*? Your mind wandered to topics completely unrelated to the text, and you could not identify the point of the paragraph (or sentence) you just read. Most people will admit to having this experience, at least occasionally. It is easy to let your mind wander when you are not *reading with purpose*. Force yourself to read with purpose. As you read each paragraph ask, "What is the purpose of this paragraph?" or "What am I supposed to learn from this paragraph?" The sole purpose of reading a textbook is to extract information. If you don't remember what you've read, you've wasted your time. Most of us are too busy to waste time.

Reading Question

1. Reading with purpose means
   a. thinking about other things while you are reading a textbook.
   b. actively trying to extract information from a text.

Try to read the next four paragraphs with purpose. What are you supposed to learn from each paragraph?

This text is structured to make it easier for you to read with purpose. The short chapters have frequent reading questions embedded in the text that make it easier for you to remember key points from preceding paragraphs. Resist the temptation to go immediately to the reading questions and search for answers in the preceding paragraphs. *Read first, and then answer the questions*. Using this approach will increase your memory for the material in this text.

Reading Question

2. Is it better to read the paragraph and then answer the reading question or to read the reading question and then search for the answer? It's better to
   a. read the paragraph, then answer the reading question.
   b. read the reading question, then search for the question's answer.

This text also features "Activities," which present new material that is NOT covered in the short chapters. When completing these activities, you will demonstrate your understanding of basic material (by answering questions) before you learn more advanced topics. You will be provided with answers to every activity question. Therefore, your emphasis when working the activities should be on understanding your answers. If you generate a wrong answer, figure out your error. Every error is an opportunity to learn. If you find your errors and correct them, you will probably not repeat the error. Resist the temptation to "get the right answer quickly." It is more important that you understand why every answer is correct.

Reading Question

3. Which of the following best describe the activities in this book?
   a. Activities introduce new material that was not included in the chapter reading.
   b. All of the new material is in the reading. The activities are simply meant to give you practice with the material in the reading.

Reading Question

4. When completing activities, your primary goal should be to get the correct answer quickly.
   a. True
   b. False

At the end of each chapter, there are "Practice Problems." After you complete the assigned activities in a chapter (and you understand why every answer is correct), you

should complete all of the practice problems. Most students benefit from a few repetitions of each problem type. The additional practice helps consolidate what you have learned so you don't forget it during tests. Finally, use the activities and the practice problems to study. Then, *after* you understand all of the activities and all of the practice problems, assess your understanding by taking a self-test. Try to duplicate a testing situation as much as possible. Just sit down with a calculator and have a go at it. If you can do the self-test, you should expect to do well on the actual exam. Taking a practice test days before your actual test will give you time to review material if you discover you did not understand something. Testing yourself is also a good way to lessen the anxiety that can occur during testing. Practice test questions are available on the SAGE website (www.sagepub.com/carlson).

Reading Question

5. How should you use the self-tests?
   a. Use them to study; complete them open-book so you can be sure to look up all the answers.
   b. Use them to test what you know days before the exam; try to duplicate the testing situation as much as possible.

MATH SKILLS REQUIRED IN THIS COURSE

Students often approach their first statistics course with some anxiety. The primary source of this anxiety seems to be a general math anxiety. The good news is that the math skills required in this course are fairly basic. You need to be able to add, subtract, multiply, divide, square numbers, and take the square root of numbers using a calculator. You also need to be able to do some basic algebra. For example, you should be able to solve the following equation for X: $22 = \frac{X}{3}$. [The correct answer is $X = 66$.]

Reading Question

6. This course requires basic algebra.
   a. True
   b. False

Reading Question

7. Solve the following equation for X: $30 = \frac{X}{3}$.
   a. 10
   b. 90

You will also need to follow the correct order of mathematical operations. As a review, the correct order of operations is (1) the operations in parentheses, (2) exponents, (3) multiplication or division, and (4) addition or subtraction. Some of you may have learned the mnemonic, Please Excuse My Dear Aunt Sally, to help remember the correct order. For example, when solving the following equation, $(3 + 4)^2$, you would first add $(3 + 4)$ to get 7

and then square the 7 to get 49. Try to solve the next more complicated problem. The answer is 7.125. If you have trouble with this problem, talk with your instructor about how to review the necessary material for this course.

$$X = \frac{(6-1)3^2 + (4-1)2^2}{(6-1)+(4-1)}$$

Reading Question

8. Solve the following equation for X: $X = \dfrac{(3-1)4^2 + (5-1)3^2}{(3-1)+(5-1)}$
   a. 11.33
   b. 15.25

You will be using a calculator to perform computations in this course. You should be aware that order of operations is very important when using your calculator. Unless you are very comfortable with the parentheses buttons on your calculator, we recommend that you do one step at a time rather than trying to enter the entire equation into your calculator.

Reading Question

9. Order of operations is only important when doing computations by hand, not when using your calculator.
   a. True
   b. False

Although the math in this course should not be new, you will see new notation throughout the course. When you encounter this new notation, relax and realize that the notation is simply a shorthand way of giving instructions. While you will be learning how to *interpret* numbers in new ways, the actual mathematical skills in this course are no more complex than the order of operations. The primary goal of this course is teaching you to use numbers to make decisions. Occasionally, we will give you numbers solely to practice computation, but most of the time you will use the numbers you compute to make decisions within a specific, real-world context.

WHY DO YOU HAVE TO TAKE STATISTICS?

You are probably reading this book because you are required to take a statistics course to complete your degree. Students majoring in business, economics, nursing, political science, premedicine, psychology, social work, and sociology are often required to take at least one statistics course. There are a lot of different reasons why statistics is a mandatory course for students in these varied disciplines, but the primary reason is that in every one of these

disciplines people make decisions that have the potential to improve people's lives, and these decisions should be informed by data. For example, a psychologist may conduct a study to determine if a new treatment reduces the symptoms of depression. Based on this study, the researcher will need to decide if the treatment is effective or not. If the wrong decision is made, an opportunity to help people with depression may be missed. Even more troubling, a wrong decision might harm people. While statistical methods will not eliminate wrong decisions, understanding statistical methods will allow you to reduce the number of wrong decisions you make. You are taking this course because the professionals in your discipline recognize that statistical methods improve decision making. Statistics make us better at our professions.

Reading Question

10. Why do many disciplines require students to take a statistics course? Taking a statistics course
    a. is a way to employ statistics instructors, which is good for the economy.
    b. can help people make better decisions in their chosen professions.

STATISTICS AND THE HELPING PROFESSIONS

When suffering from a physical or mental illness, we expect health professionals (e.g., medical doctors, nurses, clinical psychologists, and counselors) to accurately diagnose us and then prescribe effective treatments. We expect them to ask us detailed questions and then to use our answers (i.e., the data) to formulate a diagnosis. Decades of research has consistently found that health professionals who use statistics to make their diagnoses are more accurate than those who rely on their personal experience or intuition.

For example, lawyers frequently ask forensic psychologists to determine if someone is likely to be violent in the future. In this situation, forensic psychologists typically review the person's medical and criminal records as well as interview the person. Based on the records and the information gained during the interview, forensic psychologists make a final judgment about the person's potential for violence in the future. While making their professional judgment, forensic psychologists weigh the relative importance of the information in the records (i.e., the person's behavioral history) and the information obtained via the interview. This is an extremely difficult task. Fortunately, through the use of statistics, clinicians have developed methods that enable them to optimally gather and interpret data. One concrete example is the Violence Risk Appraisal Guide (Harris, Rice, & Quinsey, 1993). The guide is a list of questions that the psychologist answers after reviewing someone's behavioral history and conducting an interview. The answers to the Guide questions are mathematically combined to yield a value that predicts the likelihood of future violence. Research indicates that clinicians who use statistical approaches like the Violence Risk Appraisal Guide make more accurate clinical judgments than those who rely solely on their own judgment. Today, statistical procedures help psychologists predict many things including violent behavior, academic

success, marital satisfaction, and work productivity. Statistical approaches also help professionals determine which interventions are most effective.

Reading Question

11. Decades of research indicates that professionals in the helping professions make better decisions when they rely on
    a. statistics.
    b. their intuition and clinical experience.

HYPOTHESIS TESTING AND SAMPLING ERROR

The statistical decisions you will make in this course revolve around specific hypotheses. The primary purpose of this book is to introduce the statistical process of **null hypothesis testing**, *a formal multiple-step procedure for evaluating the likelihood of a prediction, called a null hypothesis*. Knowledge of null hypothesis testing, also called **significance testing**, is fundamental to those working in the behavioral sciences, medicine, and the counseling professions. In later chapters, you will learn a variety of statistics that test different hypotheses. All of the hypothesis testing procedures that you will learn are needed because of one fundamental problem that plagues all researchers, namely the problem of sampling error. For example, researchers evaluating a new depression treatment want to know if it effectively lowers depression in all people with depression, called the population of people with depression. However, researchers cannot possibly study every depressed person in the world. Instead, researchers have to study a subset of this population, perhaps a sample of 100 people with depression. *The purpose of any sample is to represent the population from which it came.* In other words, if the 100 people with depression are a good sample, they will be similar to the population of people with depression. Thus, if the average score on a clinical assessment of depression in the population is 50, the average score of a good sample will also be 50. Likewise, if the ratio of women with depression to men with depression is 2:1 in the population, it will also be 2:1 in a good sample. Of course, you do not really expect a sample to be exactly like the population. *The differences between a sample and the population create* **sampling error**.

Reading Question

12. All hypothesis testing procedures were created so that researchers could
    a. study entire populations rather than samples.
    b. deal with sampling error.

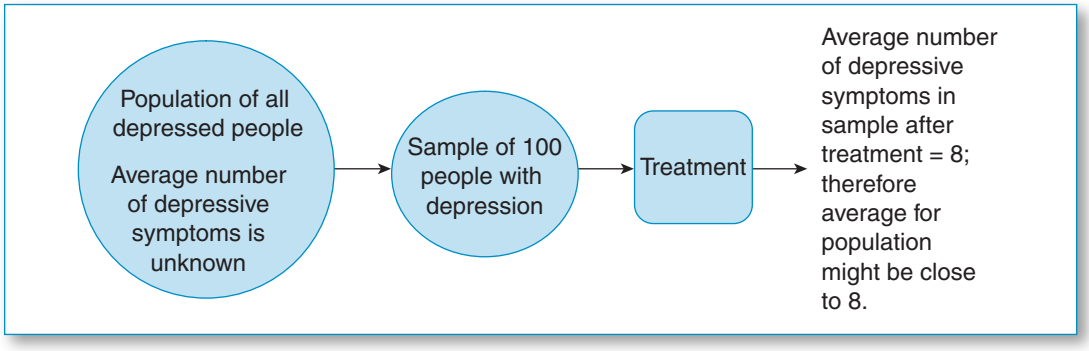
Reading Question

13. If a sample represents a population well, it will
    a. respond in a way that is similar to how the entire population would respond.
    b. generate a large amount of sampling error.

POPULATIONS AND SAMPLES

Suppose that the researcher studying depression gave a new treatment to a sample of 100 people with depression. Figure 1.1 is a pictorial representation of this research scenario. The large circle on the left represents a **population**, *a group of all things that share a set of characteristics*. In this case, the "things" are people, and the characteristic they all share is depression. Researchers want to know what the mean depression score for the population would be if all people with depression were treated with the new depression treatment. In other words, researchers want to know the **population parameter**, *the value that would be obtained if the entire population were actually studied*. Of course, the researchers don't have the resources to study every person with depression in the world, so they must instead study a **sample**, *a subset of the population that is intended to represent the population*. In most cases, the best way to get a sample that accurately represents the population is by taking a **random sample** from the population. When taking a **random sample**, *each individual in the population has the same chance of being selected for the sample*. In other words, while researchers want to know a population parameter, their investigations usually produce a **sample statistic**, *the value obtained from the sample*. The researchers then use the sample statistic value as an estimate of the population parameter value. The researchers are making an *inference* that the sample statistic is a value similar to the population parameter value based on the premise that the characteristics of those in the sample are similar to the characteristics of those in the entire population. *When researchers use a sample statistic to infer the value of a population parameter* it is called **inferential statistics**. For example, in Figure 1.1, the sample of 100 people with depression was given a new treatment. After the treatment, the average number of depressive symptoms from the sample was 8. If the researchers then inferred that the entire population of people with depression would have an average of 8 depressive symptoms after getting the new treatment, they would be basing their conclusion

Figure 1.1 A pictorial representation of using a sample to estimate a population parameter (i.e., inferential statistics).



on inferential statistics. It should be clear to you that if the sample did not represent the population well (i.e., if there was a lot sampling error), the sample statistic would NOT be similar to the population parameter. In fact, **sampling error** is defined as *the difference between a sample statistic value and an actual population parameter value.*

Reading Question

14. The value obtained from a population is called a
    a. statistic.
    b. parameter.

Reading Question

15. Parameters are
    a. always exactly equal to sample statistics.
    b. often estimated or inferred from sample statistics.

Reading Question

16. When a statistic and parameter differ,
    a. it is called an inferential statistic.
    b. there is sampling error.

The researchers studying depression were using inferential statistics because they were using data from a sample to infer the value of a population parameter. The component of the process that makes it inferential is that researchers are using data they actually have to estimate (or infer) the value of data they don't actually have. In contrast, researchers use **descriptive statistics** *when their intent is to describe the data that they actually collected.* For example, if a clinical psychologist conducted a study in which she gave some of her clients a new depression treatment and she wanted to describe the average depression score of only those clients who got the treatment, she would be using descriptive statistics. Her intent is only to describe the results she observed in the clients who actually got the treatment. However, if she then wanted to estimate what the results would be if she were to give the same treatment to additional clients, she would then be performing inferential statistics.

Reading Question

17. Researchers are using descriptive statistics if they are using their results to
    a. estimate a population parameter.
    b. describe the data they actually collected.

Reading Question

18. Researchers are using inferential statistics if they are using their results to
    a. estimate a population parameter.
    b. describe the data they actually collected.

INDEPENDENT AND DEPENDENT VARIABLES

Researchers design experiments to test if one or more variables cause changes to another variable. For example, if a researcher thinks a new treatment reduces depressive symptoms he could design an experiment to test this prediction. He might give a sample of people with depression the new treatment and withhold the treatment from another sample of people with depression. Later, if those who received the new treatment had lower levels of depression, he would have evidence that the new treatment reduces depression. In this experiment, the type of treatment each person received (i.e., new treatment vs. no treatment) is the **independent variable (IV)**. In this study, the experimenter manipulated the IV by giving one sample of people with depression the new treatment and giving another sample of people with depression a placebo treatment that is not expected to reduce depression. In this experiment, the IV has two **IV levels**: (1) the new treatment and (2) the placebo treatment. The main point of the study is to determine if the two different IV levels were differentially effective at reducing depressive symptoms. More generally, *the IV is a variable with two or more levels that are expected to have different effects on another variable.* In this study, after both samples of people with depression were given their respective treatment levels, the amount of depression in each sample was compared by counting the number of depressive symptoms in each person. In this experiment, the number of depressive symptoms observed in each person is the **dependent variable (DV)**. Given that the researcher expects the new treatment to work and the placebo treatment not to work, he expects the new treatment DV scores to be lower than the placebo treatment DV scores. More generally, *the DV is the outcome variable that is used to compare the effects of the different IV levels.*

Reading Question

19. The IV (independent variable) in a study is the
    a. variable expected to change the outcome variable.
    b. outcome variable.

Reading Question

20. The DV (dependent variable) in a study is the
    a. variable expected to change the outcome variable.
    b. outcome variable.

The term *IV* is always used to refer to the variable being tested in a **true experiment** (i.e., studies in which the IV was manipulated by the researcher). However, in this text, we also use IV in a more general way. The IV is any variable predicted to influence another variable even when the IV was not manipulated. If you take a research methods course, you will learn an important distinction between manipulated IVs and *measured* IVs (sometimes called quasi-experimental variables or subject variables). Very briefly, the ultimate goal of science is to discover causal relationships, and manipulated IVs allow researchers to draw causal conclusions while measured IVs do not. You can learn more about this important distinction and its implications for drawing causal conclusions in a research methods course.

SCALES OF MEASUREMENT

All research is based on measurement. For example, if researchers are studying depression, they will need to devise a way to measure depression accurately and reliably. The way a variable is measured has a direct impact on the type of statistical procedures that can be used to analyze that variable. Generally speaking, researchers want to devise measurement procedures that are as precise as possible because more precise measurements enable more sophisticated statistical procedures. Researchers recognize four different **scales of measurement**: (1) nominal, (2) ordinal, (3) interval, and (4) ratio. Each of these scales of measurement is increasingly more precise than its predecessor, and therefore, each succeeding scale of measurement allows more sophisticated statistical analyses than its predecessor.

Reading Question

21. The way a variable is measured
    a. determines the kinds of statistical procedures that can be used on that variable.
    b. has very little impact on how researchers conduct their statistical analyses.

For example, researchers could describe depression using a nominal scale by categorizing people with different kinds of major depressive disorders into groups, including those with melancholic depression, atypical depression, catatonic depression, seasonal affective disorder, or postpartum depression. **Nominal scales** of measurement *categorize things into groups that are qualitatively different from other groups*. Because nominal scales of measurement involve categorizing individuals into qualitatively distinct categories, they yield **qualitative** data. In this case, clinical researchers would interview each person and then decide which type of major depressive disorder each person has. With nominal scales of measurement, it is important to note that the categories are not in any particular order. A diagnosis of melancholic depression is not considered to be "more depressed" than a diagnosis of atypical depression. With all other scales of measurement, the categories are ordered. For example, researchers could also measure depression on an ordinal scale by ranking individual people in terms of the severity of their depression. **Ordinal scales** of measurement *rank order things*. In this case, researchers might interview people and diagnose them with a "mild depressive disorder," "moderate depressive disorder," or "severe depressive disorder." An ordinal scale clearly indicates that people *differ in the amount of something they possess.* Thus, someone who was diagnosed with mild depressive disorder would be less depressed than someone diagnosed with moderate depressive disorder. Although ordinal scales rank diagnoses by severity, they do not quantify how much more depressed a moderately depressed person is relative to a mildly depressed person. To make statements about how much more depressed one person is than another, an interval or ratio measurement scale is required. Researchers could measure depression on an interval scale by having people complete a multiple choice questionnaire that is designed to yield a score reflecting

the amount of depression each person has. **Interval scales** of measurement *involve quantifying how much of something people have*. While the ordinal scale indicates that some people have more or less of something than others, the interval scale is more precise indicating exactly *how much* of something someone has. Another way to think about this is that for interval scales, the intervals are equivalent whereas for ordinal scales, the intervals are not equivalent. For example, on an ordinal scale, the interval (or distance) between a mild depressive disorder and a moderate depressive disorder may not be the same as the interval between a moderate depressive disorder and a severe depressive disorder. However, on an interval scale, the distances between values are equivalent. For example, if people completed a well-designed survey instrument that yielded a score between 1 and 50, the difference in the amount of depression between scores 1 and 2 would be the same as the difference in the amount of depression between scores 11 and 12. Most questionnaires used for research purposes yield scores that are measured on an interval scale of measurement. **Ratio scales** of measurement also *involve quantifying how much of something people have but a score of zero on a ratio scale indicates that the person has none of the thing being measured*. Because they involve quantifying how much of something an individual has, interval and ratio scales yield **quantitative** data. Interval and ratio scales are similar in that they both determine how much of something someone has but some interval scales can yield a negative number while the lowest score possible on a ratio scale is zero. Within the behavioral sciences, the distinction between interval and ratio scales of measurement is not usually very important. Researchers typically use the same statistical procedures to analyze variables measured on interval and ratio scales of measurement.

Although most variables can be easily classified as nominal, ordinal, or interval/ratio, there is one type of data that can be difficult to classify. Researchers often ask people to respond to questions using a Likert scale, where they are asked to indicate the extent to which they agree with a particular statement. For example, a professor may ask students at the end of the semester how much they agree with the statement, "I enjoyed taking this statistics course." Students may respond with 1 = *strongly agree*, 2 = *agree*, 3 = *neither agree nor disagree*, 4 = *disagree*, 5 = *strongly disagree*. Although there is not complete agreement among statisticians, most researchers classify these Likert questions as interval because there is evidence that people perceive the distance between categories as equivalent. Thus, in this course, responses to these types of questions will be considered interval/ratio data.

Reading Question

22. Researchers typically treat Likert scale responses (i.e., 1 = *strongly agree*, 2 = *agree*, etc.) as which scale of measurement?

    a. Nominal scale of measurement
    b. Ordinal scale of measurement
    c. Interval scale of measurement

When trying to identify the scale of measurement of a variable, it can also be helpful to think about what each scale of measurement allows you to do. For example, if you can only count the number of things in a given category, you know that you have a nominal scale.

Table 1.1 summarizes what you can do with each type of scale and provides examples of each scale of measurement:

Table 1.1 The four scales of measurement, what they allow, and examples.

Scale of Measurement	*What the Scale Allows You to Do*	*Examples*
Nominal	COUNT the number of things within different categories	Pets: 5 dogs, 12 cats, 7 fish, 2 hamsters
		Marital status: 12 married, 10 divorced, 2 separated
Ordinal	RANK some things as having more of something than others (but NOT QUANTIFY how much of it they have)	Annual income: above average, average, or below average
		Speed (measured by place of finish in a race): 1st, 2nd, 3rd, etc.
Interval	QUANTIFY how much of something there is but a score of zero does not mean the absence of the thing being measured	Temperature: –2° F, 98° F, 57° F; 0° F is not the absence of heat
Ratio	QUANTIFY how much of something there is and a score of zero means the absence of the thing being measured	Annual income: $25,048, $48,802, $157,435, etc.
		Number of text messages sent in a day: 0, 3351, 15, etc.

Reading Question

23. The scale of measurement that quantifies the thing being measured (i.e., indicates *how much* of it there is) is _____ scale(s) of measurement.
    a. the nominal
    b. the ordinal
    c. both the interval and ratio

Reading Question

24. The scale of measurement that categorizes objects into different kinds of things is _____ scale(s) of measurement.
    a. the nominal
    b. the ordinal
    c. both the interval and ratio

Reading Question

25. The scale of measurement that indicates that some objects have more of something than other objects but not how much more is _____ scale(s) of measurement
    a. the nominal
    b. the ordinal
    c. both the interval and ratio

DISCRETE VERSUS CONTINUOUS VARIABLES

Variables can also be categorized as discrete or continuous. A **discrete variable** *is measured in whole units rather than fractions of units*. For example, the variable "number of siblings" is a discrete variable because someone can only have a whole number of siblings (i.e., no one can have 2.7 siblings). A **continuous variable** *is measured in fractions of units*. For example, the variable "time to complete a test" is a continuous variable because someone can take a fraction of minutes to complete a test (i.e., 27.39 minutes). Nominal and ordinal variables are always discrete variables. Interval and ratio variables can be either discrete or continuous.

Reading Question

26. If a variable can be measured in fractions of units, it is a _____ variable.
    a. discrete
    b. continuous

GRAPHING DATA

The first step in all statistical analyses is to graph your data. Creating graphs gives you a picture of the data that you can inspect to "get a feel" for the data. For example, if you were looking at the number of siblings college students have, you could begin by looking at a graph to determine how many siblings most students have. Inspection of the graph also allows you to find out if there is anything odd in the data file that requires further examination. For example, if you graphed the data and found that most people reported having between 0 and 4 siblings but one person reported having 20 siblings, you should probably investigate to determine if that 20 was an error.

There are three basic types of graphs that we use for most data: (1) **bar graphs**, (2) **histograms**, and (3) **line graphs**. The names of the first two are a bit misleading because both are created using bars. The only difference between a bar graph and a histogram is that in a bar graph the bars do not touch while the bars do touch in a histogram. In general, bar graphs are used when the data are discrete or qualitative. The space between the bars of a bar graph emphasize that there are no possible values between any two categories. For example, when graphing the number of children in a family, a bar graph is appropriate because there is no possible value between any two categories (e.g., 1 and 2 children).

When the data are continuous, we use a histogram. The bars touch in a histogram to indicate that there are possible values between any two categories. For example, if we were graphing time to complete a test, the bars would touch to indicate that there are possible values between any two times (e.g., 27 and 28 minutes).

Reading Question

27. What type of graph is used for discrete data or qualitative data?
    a. bar graph
    b. histogram

Reading Question

28. What type of graph is used for continuous data?
    a. bar graph
    b. histogram

Reading Question

29. In bar graphs, the bars _____.
    a. touch
    b. don't touch

Reading Question

30. In histograms, the bars _____.
    a. touch
    b. don't touch

To create either a bar graph or a histogram, you should put categories on the x-axis and the number of scores in a particular category (i.e., the frequency) on the y-axis. For example, suppose we asked 19 students how many siblings they have and obtained the following responses:

0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 6

To graph these responses, you would list the range of responses to the question "How many siblings do you have?" on the x-axis (i.e., in this case 0 through 6). The y-axis is the frequency within each category. For each response category, you will draw a bar with a height equal to the number of times that response was given. For example, in the bar graph (Figure 1.2), 4 people said they had 0 siblings and so the bar above the 0 has a height of 4.

Reading Question

31. Use the graph to determine how many people said they had 1 sibling.
    a. 4
    b. 5
    c. 6

Figure 1.2 Bar graph of variable, number of siblings, collected from a sample of 19 students.

The procedure for creating a histogram is similar to that for creating a bar graph. The only difference is that the bars should touch. For example, suppose that you recorded the height of players on a volleyball team and obtained the following heights rounded to the nearest inch:

65, 67, 67, 68, 68, 68, 69, 69, 70, 70, 70, 71, 72

Height in inches is continuous because there are an infinite number of possible values between any two categories (e.g., between 68 and 69 inches). The data are continuous so we create a histogram (i.e., we allow the bars to touch) (Figure 1.3).

Whenever a histogram is appropriate, you may also use a **line graph** in its place. To create a line graph, you use dots to indicate frequencies and connect adjacent dots with lines (Figure 1.4).

Whether the data are discrete or continuous should determine how the data are graphed. You should use a bar graph for discrete data and a histogram or a line graph for continuous data. Nominal data should be graphed with a bar graph. Throughout the text we will use these guidelines, but you should be aware of the unfortunate fact that histograms and bar graphs are often used interchangeably outside of statistics classes.

Reading Question

32. Line graphs can be used whenever a _____ is appropriate.
    a. histogram
    b. bar graph

Figure 1.3 Frequency histogram of variable, height in inches, collected from a sample of 13 volleyball players.

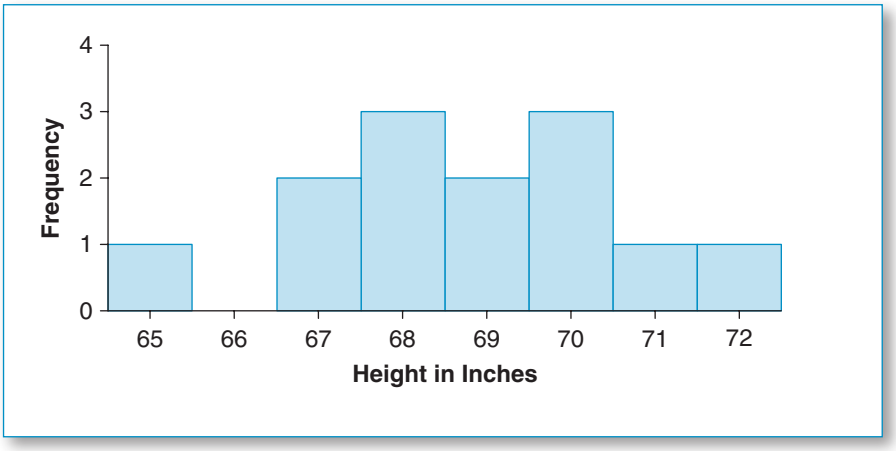
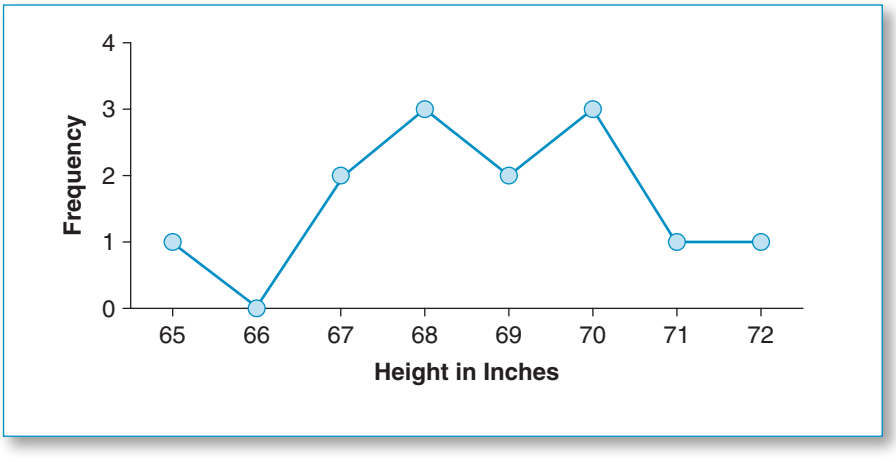


Figure 1.4 Frequency line graph of variable, height in inches, collected from a sample of 13 volleyball players.



Reading Question

33. What type of graph should be used if the data are measured on a nominal scale?

a. histogram

b. bar graph

FREQUENCY DISTRIBUTION TABLES

Graphing data is typically the best way to see patterns in the data and to look for potential problems. However, some precision is often lost with graphs. Therefore, it is sometimes useful to look at the raw data in a **frequency distribution table**. To create a frequency distribution table, you need to know the measurement categories as well as the number of responses within a given measurement category. For example, suppose that a market researcher asked cell phone users to respond to the following statement: "I am very happy with my cell phone service provider." People were asked to respond with 1 = *strongly agree*, 2 = *agree*, 3 = *neither agree nor disagree*, 4 = *disagree*, 5 = *strongly disagree*. The responses are listed below:

1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5

It is probably obvious that a string of numbers is not a particularly useful way to present data. A frequency distribution table organizes the data, so it is easier to interpret; one is shown in Table 1.2.

The first column (*X*) represents the possible response categories. People *could* respond with any number between 1 and 5, therefore the *X* column (i.e., the measurement categories) must include all of the *possible* response values, namely 1 through 5. In this case, we chose to put the categories in ascending order from 1 to 5, but they could also be listed in descending order from 5 to 1.

The next column (*f*) is where you record the frequency of each response. For example, 4 people gave responses of 5 (*strongly disagree*) and so a 4 is written in the "*f*" column across from the response category of 5 (*strongly disagree*).

Table 1.2 Frequency distribution table of the variable "I am very happy with my cell phone service provider."

	X	*f*
Strongly agree	1	2
Agree	2	4
Neither agree nor disagree	3	7
Disagree	4	6
Strongly disagree	5	4

Reading Question

34. The value for "*f*" represents the
    a. number of measurement categories.
    b. number of responses within a given measurement category.

Reading Question

35. In the above frequency table, how many people responded with an answer of 3?
    a. 2
    b. 4
    c. 7

SPSS

We will be using a statistical package called **IBM SPSS** to conduct many of the statistical analyses in this course. Our instructions and screenshots were developed with version 18. There are some minor differences between version 18 and other versions but you should have no difficulty using our instructions with other SPSS versions.

It is likely that your school has a site license for SPSS allowing you to access it on campus. Depending on your school's site license, you may also be able to access the program off campus. You may also purchase or "lease" a student or graduate version of SPSS for this course. Your instructor will tell you about the options available to you.

Data File

After you open SPSS, click on the Data View tab near the bottom left of the screen. Enter the data you want to analyze in a single column.

We have used the cell phone data from the previous page to help illustrate how to use SPSS. In the screen shot in Figure 1.5, a variable name "happycellphone" is shown at the top of the column of data. To add this variable name, double click on the blue box at the top of a column in the Data View screen. Doing so will take you to the Variable View screen. You can also access the Variable View screen by pressing the Variable View tab at the bottom left of the screen. In the first column and first row of the variable view screen, type the name of the variable you want to appear in the data spreadsheet (e.g., happycellphone—the variable name cannot have spaces or start with a number). To go back to the Data View, click on the blue Data View tab at the bottom left of the screen.

The data file you created should look like the screen shot in Figure 1.5. The exact order of the data values is not important, but all 23 scores should be in a single column. As a general rule, all of the data for a variable must be entered in a single column.

Reading Question

36. The Variable View screen is where you
    a. enter the variable names.
    b. enter the data.

Reading Question

37. The Data View screen is where you
    a. enter the variable names.
    b. enter the data.

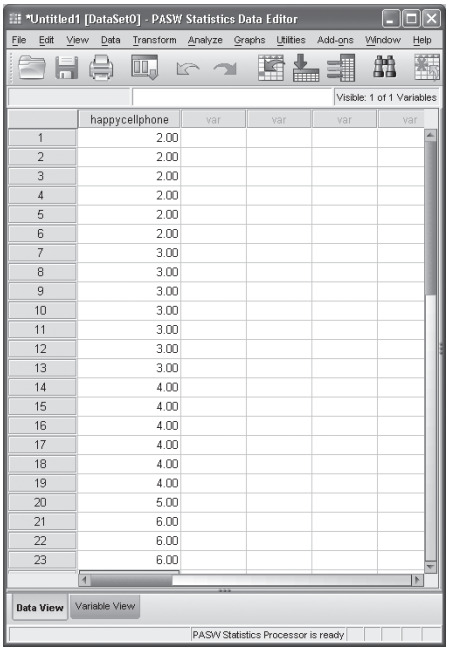
Obtaining Frequency Distribution Tables and Graphs

SPSS can create frequency tables and graphs. To create a frequency graph of the data you just entered, do the following:

- From the Data View screen, click on Analyze, Descriptive Statistics, and then Frequencies.

- To create a graph, click on the Charts button and then choose the type of graph you want to create (Bar chart, Pie chart, or Histogram). Click on the Continue button.
- Be sure that the Display Frequency Tables box is checked if you want to create a frequency distribution table.
- Click on the OK button to create the frequency distribution table and graph.

Figure 1.5 Screenshot of SPSS data entry screen.



Annotated Output

After performing the steps outlined above, a frequency distribution graph and table will appear in the SPSS output screen. Use the SPSS output provided in Figure 1.6 to answer the following three questions.

Reading Question

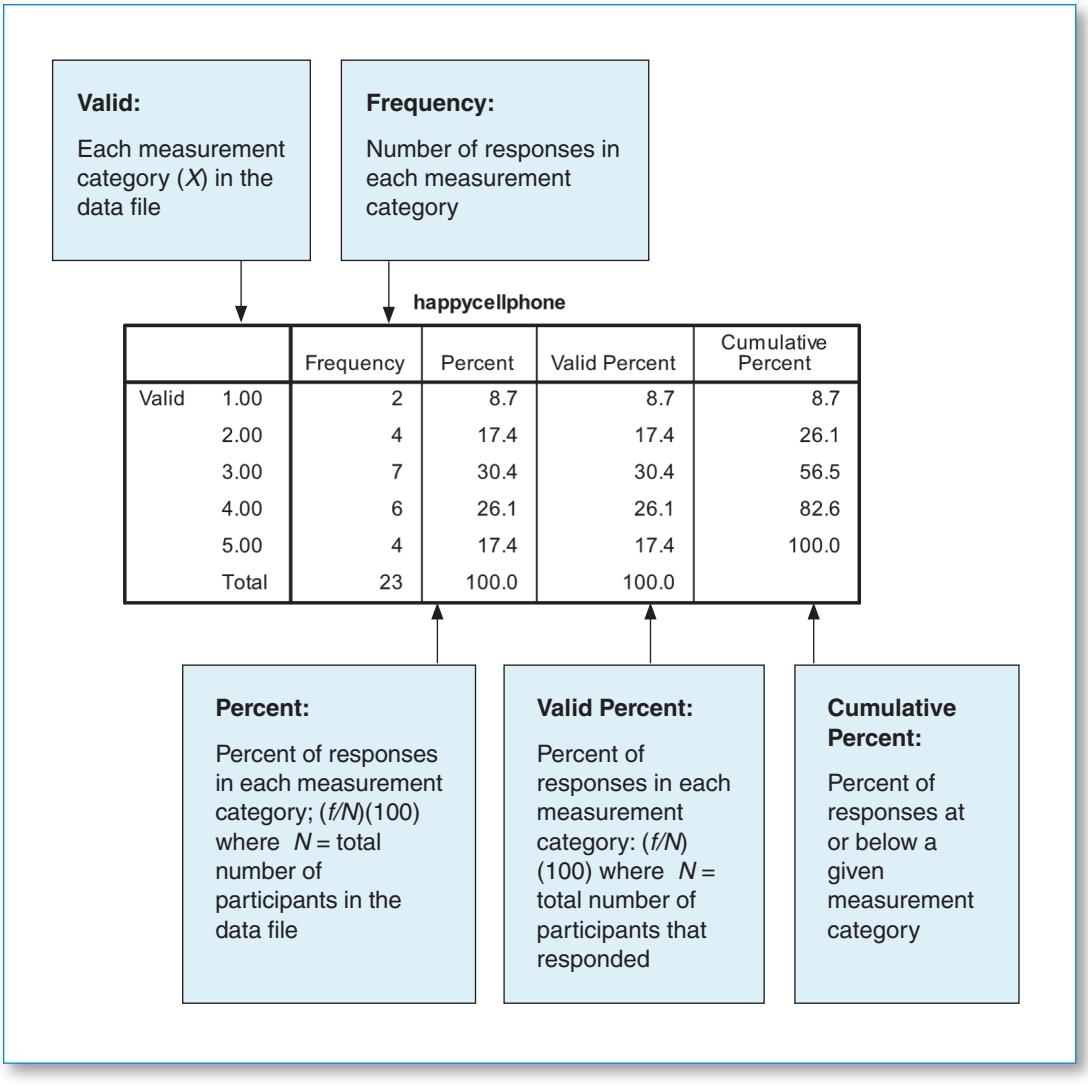
38. How many people responded with a 3 to the question "I am very happy with my cell phone provider?"

a. 2

b. 4

c. 7

Figure 1.6 Annotated SPSS frequency table output.



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	2	8.7	8.7	8.7
	2.00	4	17.4	17.4	26.1
	3.00	7	30.4	30.4	56.5
	4.00	6	26.1	26.1	82.6
	5.00	4	17.4	17.4	100.0
	Total	23	100.0	100.0	

Reading Question

39. What percentage of the respondents answered the question with a response of 4?

a. 30.4

b. 26.1

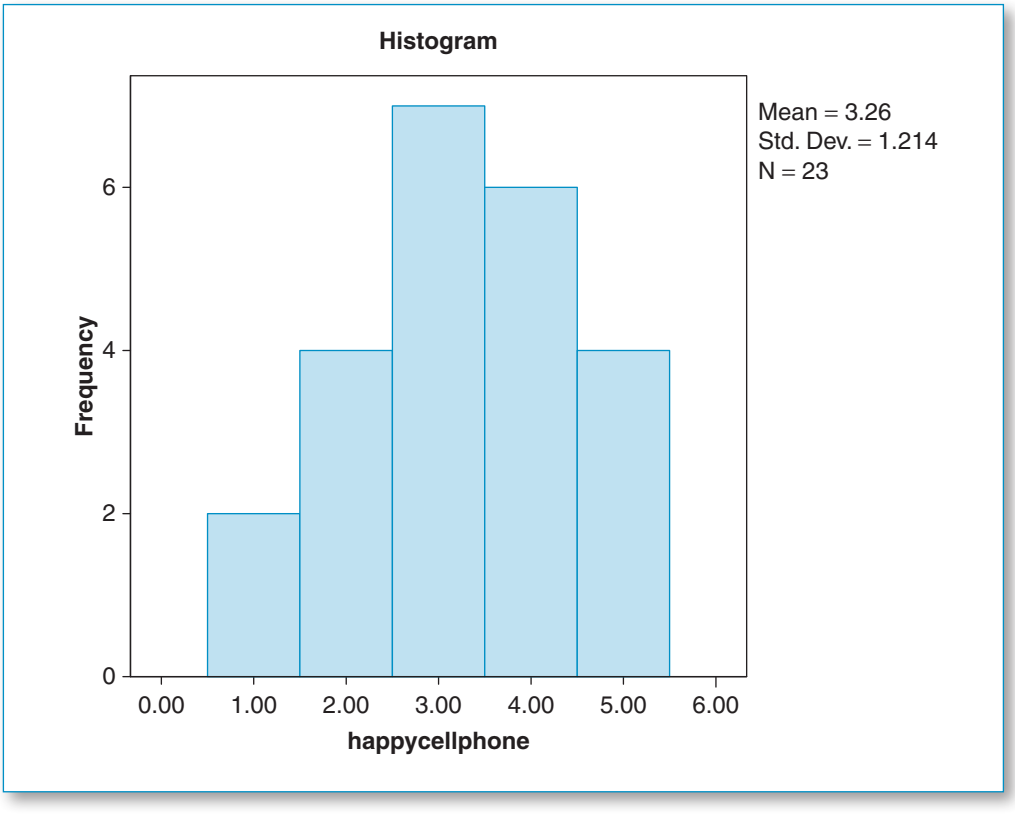
c. 17.4

Reading Question

40. What percentage of the respondents answered the question with a response of 4 or a lower value?
    a. 56.5
    b. 82.6
    c. 100

Use the histogram in Figure 1.7 to answer the following two questions.

Figure 1.7 Frequency histogram of "I am very happy with my cell phone service provider" data.



Reading Question

41. What is the most common response in the data?
    a. 2
    b. 3
    c. 4
    d. 5

Reading Question

42. How many people responded with the most common response?
    a. 7
    b. 6
    c. 5
    d. 4

SPSS is a great tool for creating graphs to help you gain a better understanding of your data. However, it is not necessarily intended for creating presentation-quality graphs. You can customize graphs in SPSS by double clicking on the graph once you create it and then, once the image is open, double click on any aspect of the graph to change it. This is trickier than it sounds because there are a lot of options. We are not going to work on editing graphs in this course, but if you would like to edit graphs you can use the Help menu to obtain further information. There are several other ways to create more advanced graphs in SPSS. You can explore these options by clicking on Graphs menu.

Reading Question

43. It is possible to change the appearance of graphs created by SPSS.
    a. True
    b. False

REFERENCE

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior*, *20*, 315–335.

Activity 1-1: Frequency Distributions

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Use common statistical terms correctly in a statistical context
- Construct a frequency distribution table from a bar graph
- Write a meaningful paragraph based on data in a frequency distribution
- Use SPSS to create a frequency table
- Interpret an SPSS frequency table
- Sketch a frequency distribution
- Identify distributions that are bell shaped, positively skewed, and negatively skewed

THERAPEUTIC TOUCH ACTIVITY

There is quite a bit of evidence that human touch is beneficial to our psychological and physical health. Hugs are associated with lower blood pressure, skin-to-skin contact can help preterm infants gain weight, and touch can improve immune system function. Although there is little doubt of the benefits of touch, a treatment known as "Therapeutic Touch" (TT) is far more controversial. Therapeutic touch involves no actual physical contact. Instead, practitioners use their hands to move "human energy fields" (HEFs) in an attempt to promote healing. Proponents of this approach claim that it can help with relaxation, reduce pain, and improve the immune system.

Emily Rosa (who was just 9 years old at the time) and her colleagues (including her parents) investigated the basis of these claims by putting a sample of actual TT practitioners to the test (Rosa, Rosa, Sarner, & Barrett, 1998). In their study, Rosa and colleagues designed a method to determine if TT practitioners could actually detect HEFs. As the figure to the right illustrates, individual practitioners sat at a table facing a large divider that prevented them from seeing their own hands or Emily. The practitioners placed both of their hands through the divider on the table, palms up. Practitioners were told to indicate whether Emily was holding her hand above their right or left hand. Emily began each trial by flipping a coin to determine where to place her hand. She then placed her hand 8 to 10 cm above one of the practitioner's hands. The practitioners had to "sense" the HEF allegedly emanating from Emily's hand to determine if Emily's hand was over their right hand or left hand. Each practitioner went through a total of 10 of these trials.

If the TT practitioners can actually sense HEFs they should be able to choose the correct hand 10 out of 10 times. However, if they really can't detect HEFs and the practitioners were guessing you would expect them to choose the correct hand 5 out of 10 times. Some may get more than 5 correct others may get less than 5 correct but the most common number of correct answers would be about 5 of 10, if the practitioners were guessing.

1. As mentioned previously, the researchers had a sample of TT practitioners participate in the experiment described above. They used the results from this sample to infer what the results would be if they had collected data from the entire population of TT practitioners. The purpose of their study was
   a. descriptive.
   b. inferential.

2. Use three of the following terms to fill in the blanks: parameters, statistics, inferential, descriptive, sampling error.

   If the sample of TT practitioners represented the population of TT practitioners well, the sample _____ would be similar to the population _____ and the study would have a relatively small amount of _____.

3. After the experiment was complete, the researchers counted the number of correct responses generated by each participant out of 10. The number of correct responses out of 10 ranged between a low of 1 correct to a high of 8 correct. The variable "number of correct responses out of 10 trials" is measured on which scale of measurement?

   a. Nominal

   b. Ordinal

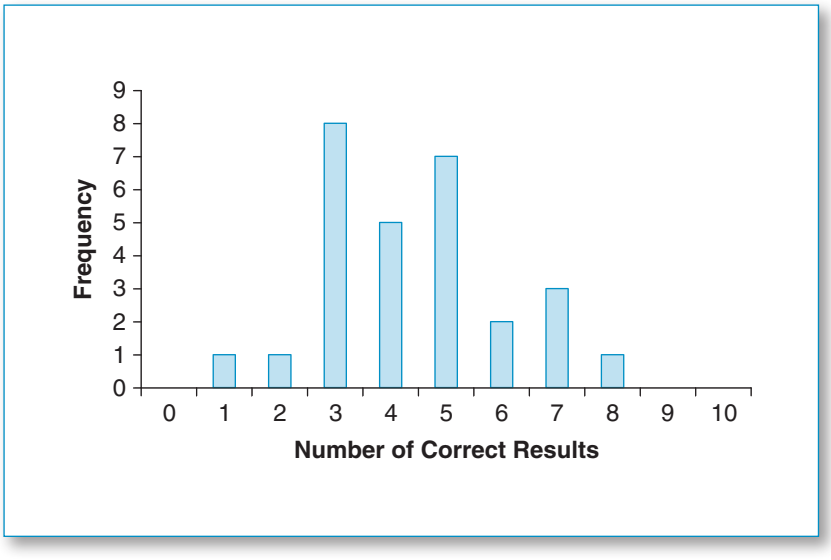
   c. Interval/Ratio

4. Is the number of correct responses out of 10 a continuous or discrete variable?

   a. Continuous

   b. Discrete

The following bar graph is an accurate re-creation of the actual data from the experiment. The graph is a frequency distribution of the number of correct responses generated by each practitioner out of 10 trials. Use these data to answer the following questions:



5. Use the data from the graph to complete the frequency table on the right.

X	*f*

6. How many practitioners were in the sample?
   a. 8
   b. 10
   c. 28
7. How many practitioners did *better* than chance (i.e., did better than 5 correct out of 10)?
   a. 3
   b. 6
   c. 13
8. What *percentage* of the practitioners performed *at or below* chance?
   a. 100
   b. 78.6
   c. 53.6
9. Compose several sentences that summarize the data. In other words, does the data support the conclusion that TT practitioners can detect HEFs or does the data support the conclusion that they cannot and instead are guessing? What evidence from the data supports your conclusion?

Every 2 years the National Opinion Research Center asks a random sample of adults in the United States to complete the **General Social Survey** (GSS). All of the GSS data are available at www.norc.org. You will be using a small portion of the GSS that we placed in a file titled "gss2010.sav." Your instructor will tell you how to access this file. Load this file into SPSS.

Part of the GSS assesses respondents' science knowledge. In 2010, respondents answered questions from a variety of different sciences, such as "True or False. Antibiotics kill viruses as well as bacteria" and "True or False. Lasers work by focusing sound waves." For this assignment, we created the variable "ScientificKnowledge" by summing the total number of correct answers each participant gave to 10 science questions. The resulting "ScientificKnowlege" variable was measured on a ratio scale and had a possible range of 0 to 10 correct answers.

Use SPSS to create a frequency distribution table and graph of "ScientificKnowledge" scores. To create a frequency distribution table and graph, do the following:

- From the Data View screen, click on Analyze, Descriptive Statistics, and then Frequencies.
- Move the variable(s) of interest into the Variable(s) box. In this case, you will move "ScientificKnowledge" into the Variable(s) box.

- Make sure the Display Frequency Tables box is checked.
- To create a graph, click on the Charts button and then choose Bar chart. Click on the Continue button.
- Click on the OK button to create the frequency distribution table and graph.

10. Use the frequency distribution table you created in SPSS to determine how many people responded to the "ScientificKnowledge" questions.

11. How many people got all 10 questions right?

12. What *percentage* of people got all 10 questions right?

13. How many people got all 10 questions wrong?

14. What *percentage* of people got all 10 questions wrong?

15. All of the questions had just two response options. Thus, if people answered every question and they were *guessing* on every question, we would expect them to get 50% of the questions correct. What percentage of people got exactly 5 of the 10 questions correct?

16. What percentage of people scored at or below chance (i.e., 5 correct responses out of 10) on this test?

17. After taking standardized tests you typically get a raw score as well as a **percentile rank**, *the percentage of scores a given score is higher than*. For example, if you scored at the 95th percentile you would know that you scored as well or better than 95% of the people who took the test. The same thing can be done for this data file by using the cumulative percent column of the frequency distribution table? What Science Knowledge test score is at the 95th percentile?

18. What Science Knowledge test score is at the 9th percentile?

19. What is the percentile rank for a Science Knowledge test score of 7?

20. What is the percentile rank for a Science Knowledge test score of 3?

21. On the GSS, people were asked how many years of school they completed. Before you graph the data in SPSS, what two responses to this question do you think will

be the most common in the United States? Explain why you think these answers would be given frequently.

22. Create a bar graph of the YearsofEducation variable. What was the most frequently occurring response?

23. What percentage of the respondents completed exactly 12 years of school?

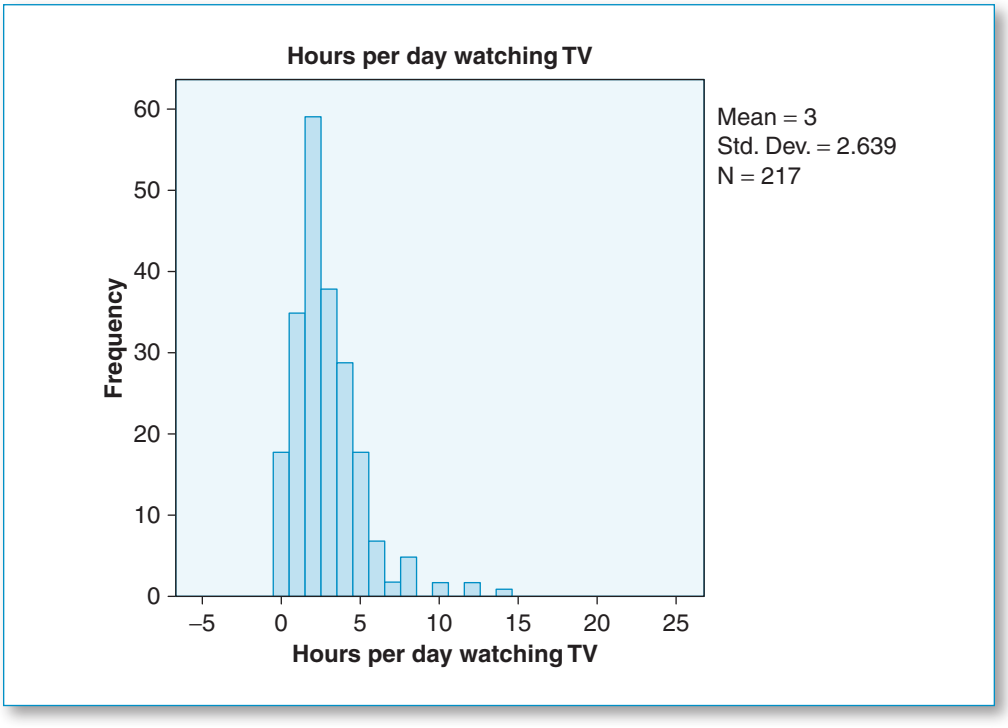
24. What percentage of the respondents completed 16 or more years of school?

25. What percentage of the respondents completed fewer than 12 years of school?

26. As you work with data throughout this course you will find that frequency distribution graphs (i.e., bar graphs and histograms) can look quite different for different variables. By far, the most commonly seen shape in statistics courses is a bell-shaped distribution. For example, the Science Knowledge scores that you graphed above are approximately bell shaped. Sketch a bar graph of the Science Knowledge scores below:

This **bell shape** occurs when most of the scores are concentrated around the mean of the scores, and there are fewer and fewer scores the further you get from the mean. For example, the average score on the Scientific Knowledge test was 5.95, and most of the scores are very close to 5.95. The further a score is from 5.95, the less frequently it occurred. If you look at the Scientific Knowledge graph closely, you will see that it is not perfectly bell shaped because it is not perfectly symmetrical. The right and left sides of the distribution do not look exactly alike. Statisticians often refer to the declining slopes to the right and left of a distribution's peak as "tails." When the right tail is longer than the left tail, the distribution is **positively skewed**. Conversely, when the left tail is longer than the right tail the distribution is **negatively skewed**. Therefore, if the right tail (which is over the larger numbers) is longer, the distribution is positively skewed. Conversely, if the left tail (which is over the smaller numbers) is longer, the distribution is negatively skewed.

27. Is the graph of the number of hours people report watching TV each day positively or negatively skewed?



28. Make a rough sketch of a frequency distribution graph that is negatively skewed.

29. On the GSS, respondents were asked how old they were when their first child was born (variable is named AgeFirstChildBorn). Use SPSS to create a histogram for this variable. Does the data look to be positively skewed, negatively skewed, or bell shaped?

30. As part of the GSS, respondents were given a vocabulary test that consisted of 10 words. Create a bar chart of the number of correct words on the vocabulary test (VocabTest). Do the data look to be positively skewed, negatively skewed, or bell shaped?

31. If a test is relatively easy and most people get between 90% and 100%, but a few people get low scores (10–20%), would that data be positively skewed or negatively skewed?

    a. Positively skewed

    b. Negatively skewed

32. A recent study revealed that the brains of new mothers grow bigger after giving birth. The researchers performed MRIs (magnetic resonance imagings) on the brains of 19 women and found that the volume of the hypothalamus was greater after giving birth than prior to giving birth. Circle the scale of measurement used for each of the variables listed below:

    a. Volume of the hypothalamus: Nominal Ordinal Interval/ratio

    b. Before and after birth: Nominal Ordinal Interval/ratio

33. A researcher designs a study in which participants with low levels of HDL (high-density lipoprotein) cholesterol are randomly assigned to exercise for 0 minutes, 30 minutes, 60 minutes, or 90 minutes a day. After 3 months, HDL levels are measured. Women and men often react differently to treatments, and so the researcher also records the gender of each participant. Circle the scale of measurement used for each of the variables listed below:

    a. Amount of exercise (0, 30, 60, or 90 minutes a day): Nominal Ordinal Interval/ratio

    b. HDL cholesterol levels: Nominal Ordinal Interval/ratio

    c. Gender of the participants: Nominal Ordinal Interval/ratio

34. A researcher used government records to classify each family into one of seven different income categories ranging from "below the poverty line" to "more than $1 million dollars a year." The researcher used police records to determine the number of times each family was burglarized. Circle the scale of measurement used for each of the variables listed below:

    a. Income category: Nominal Ordinal Interval/ratio

    b. Number of times burglarized: Nominal Ordinal Interval/ratio

REFERENCE

Rosa, L., Rosa, E., Sarner, L., & Barrett, S. (1998). A close look at therapeutic touch. *Journal of the American Medical Association, 279*(13), 1005–1010.

Activity 1-2: Practice Problems

1. Practice determining if variables are continuous or discrete.

    a. Weight of a rat in ounces Continuous Discrete

    b. Number of heart attacks Continuous Discrete

    c. Grade point average Continuous Discrete

    d. The number of students in a class Continuous Discrete

    e. Percent body fat Continuous Discrete

2. Practice identifying the scale of measurement as nominal, ordinal, or interval/ratio.

a.	Weight of a rat in ounces	Nominal	Ordinal	Interval/ratio
b.	Number of heart attacks	Nominal	Ordinal	Interval/ratio
c.	Football jersey number	Nominal	Ordinal	Interval/ratio
d.	Grade point average	Nominal	Ordinal	Interval/ratio
e.	Class standing (freshmen, sophomore, etc.)	Nominal	Ordinal	Interval/ratio
f.	Percent body fat	Nominal	Ordinal	Interval/ratio
g.	Military rank (private, corporal, etc.)	Nominal	Ordinal	Interval/ratio
h.	Student identification number	Nominal	Ordinal	Interval/ratio
i.	The number of students in a class	Nominal	Ordinal	Interval/ratio

3. True or False. Interval/Ratio data are always continuous.

4. A student distributes a questionnaire to 21 students in her statistics course. On this questionnaire, people are asked to indicate the extent to which they agree with this statement: "I am happy with my life." Responses were made using a 7-point Likert scale, where 1 = *strongly disagree* and 7 = *strongly agree*. The data she obtained are as follows: 7, 7, 7, 7, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 4, 3, 3, 1

   a. What scale of measurement is this variable measured on (i.e., nominal, ordinal, interval/ratio)?

   b. Create a frequency table for these data.

   c. Use your frequency distribution table to determine how many people had scores of 4. Don't just count from the raw data. Make sure you understand the table.

   d. Enter the data into SPSS and create a frequency distribution table. What percentage of people had scores of 4?

   e. What percentage of people had scores of 4 or below?

   f. Create a histogram of these data by hand and using SPSS.

   g. Is this distribution best described as bell shaped, positively skewed, or negatively skewed?

5. A political pollster asked voters to indicate the degree of their agreement with the following question: "Global warming is real and is at least partially caused by human activity." The voters responded using the following scale: 1 = *strongly disagree* to 5 = *strongly agree*. The SPSS output for the pollster's data are presented below.

globalwarming

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	5	21.7	21.7	21.7
	2.00	7	30.4	30.4	52.2
	3.00	1	4.3	4.3	56.5
	4.00	4	17.4	17.4	73.9
	5.00	6	26.1	26.1	100.0
	Total	23	100.0	100.0	

   a. How many people responded to this survey question?
   b. How many people *strongly agreed* with the statement?
   c. What percentage of people *strongly disagreed* with the statement?
   d. What percentage of the sample disagreed with the statement about global warming (i.e., gave an answer of 2 or lower)?
   e. What score is at the 74th (73.9) percentile?
   f. What is the percentile rank for a score of 3?
   g. Are these data normally distributed (i.e., bell shaped)?
   h. Suppose that these scores came from all of the students in a political science class. If these data were being used to describe the class, would you use descriptive or inferential statistics?
   i. If these data were intended to represent all voters, would you use descriptive or inferential statistics?

6. A chair of a foreign languages department is interested in determining the language background of incoming freshmen. She asks a random sample of freshmen to report the number of languages they speak fluently. The following data were obtained:

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 5

   a. What scale of measurement is this variable measured on (i.e., nominal, ordinal, interval/ratio)?
   b. Are these data discrete or continuous?

c. Create a frequency table for these data.

d. Create an appropriate graph for these data.

e. Enter the data into SPSS and create a frequency distribution table. What percentage of students report that they speak two languages fluently?

f. What percentage of students report that they speak two or fewer languages fluently?

g. Is this distribution best described as bell shaped, positively skewed, or negatively skewed?