of heads to 0.5 after two tosses. The next three tosses gave a tail followed by two heads, so the proportion of heads after five tosses is 3/5, or 0.6.

The proportion of tosses that produce heads is quite variable at first, but it settles down as we make more and more tosses. Eventually this proportion gets close to 0.5 and stays there. We say that 0.5 is the *probability* of a head. The probability 0.5 appears as a horizontal line on the graph.

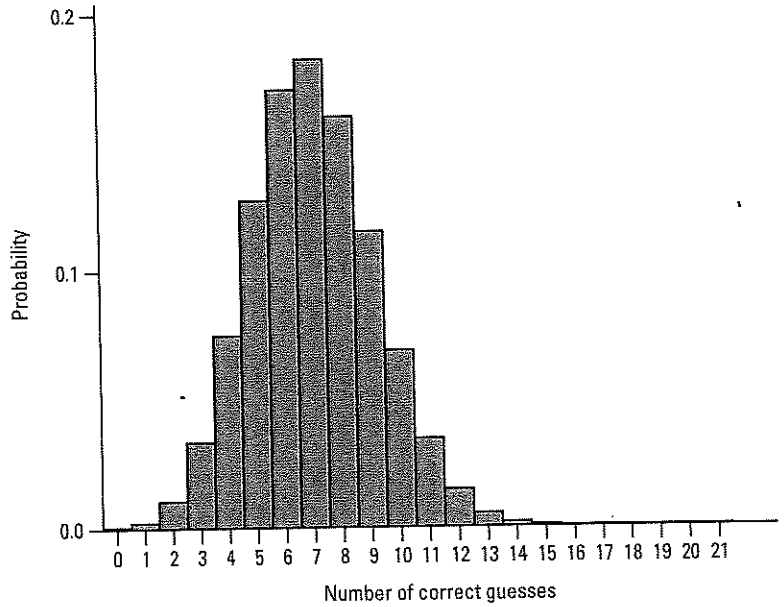
Example P.9 illustrates the big idea of probability: **chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run.** Casinos rely on this fact to make money every day of the year. We can use probability rules to analyze games of chance, like roulette, blackjack, and Texas hold 'em.

Probability plays an even more important role in the study of *variation*. If we toss a coin 30 times, will we get exactly 15 heads? Perhaps. Could we get as few as 11 heads? More than 24 heads? Probability tells us that there's about a 10% chance of getting 11 or fewer heads and less than a 1-in-1000 chance of getting more than 24 heads. If we toss our coin 30 times over and over and over again, the number of heads we obtain will vary. Probability quantifies the pattern of chance variation.

Example P.10 *Water, water everywhere*
Using probability to measure "how likely"

How can probability help us determine whether students can distinguish bottled water from tap water? Let's return to the Activity (page 4). Suppose that in Mr. Bullard's class, 13 out of 21 students made correct identifications. If we assume that the students in his

Figure P.6 *Graph showing the probability for each possible number of correct guesses in Mr. Bullard's class.*



class *cannot* tell bottled water from tap water, then each one is basically guessing, with a 1-in-3 chance of being correct. So we'd expect about one-third of his 21 students, that is, about 7 students, to guess correctly. How likely is it that as many as 13 of his 21 students would guess correctly?

Figure P.6 is a graph of the probability values for the number of correct guesses in Mr. Bullard's class. As you can see from the graph, the chance of guessing 13 or more correctly is quite small. In fact, the actual probability of doing so is 0.0068.

So what do we conclude? Either Mr. Bullard's students are guessing, and they have incredibly good luck, or the students are not guessing. Since the students have less than a 1% chance of getting so many right "just by chance," we feel pretty sure that they are not guessing. It seems that they can detect the difference in taste between tap and bottled water.

As the previous example shows, probability allows us to decide whether an observed outcome is too unlikely to be due to chance variation. Too many students were able to identify which of their three cups contained a different type of water for us to believe that they were guessing. In effect, we tested the claim that the students were guessing. This is our first encounter with ***statistical inference.*** Notice the important role that probability played in leading us to a conclusion.

statistical inference

Statistical Inference: Drawing Conclusions from Data

How prevalent is cheating on tests? Representatives from the Gallup Organization were determined to find out. They conducted an Internet survey of 1200 students, aged 13 to 17, between January 23 and February 10, 2003. The question they posed was "Have you, yourself, ever cheated on a test or exam?" Forty-eight percent of those surveyed said "Yes." If Gallup had asked the same question of *all* 13- to 17-year-old students, would exactly 48% have answered "Yes"?

Gallup is trying to estimate the unknown percent of students in this age group who would say they have cheated on a test. (Notice that we didn't say the percent of students who actually *had* cheated on a test!) Their best estimate, given the survey results, would be 48%. But the folks at Gallup know that samples vary. If they had selected a different sample of 1200 students to respond to the survey, then they would probably have gotten a different estimate. *Variation is everywhere!*

Fortunately, probability provides a description of how the sample results will vary in relation to the true population percent. Based on the sampling method that Gallup used, we can say that their estimate of 48% is very likely to be within 3% of the true population percent. That is, we can be quite confident that between 45% and 51% of *all* teenage students would say that they have cheated on a test.

Statistical inference allows us to use the results of properly designed experiments, sample surveys, and other observational studies to draw conclusions that go beyond the data themselves. Whether we are testing a claim, as in the bottled versus tap water Activity, or computing an estimate, as in the Gallup survey, we rely on probability to help us answer research questions with a known degree of confidence. Unfortunately, we cannot be *certain* that our conclusions are correct. The following example shows you why.

Example P.11 *Do mammograms help?*
Experiments and inference

Most women who reach middle age have regular mammograms to detect breast cancer. Do mammograms really reduce the risk of dying of breast cancer? To seek answers, doctors rely on "randomized clinical trials" that compare different ways of screening for breast cancer. We will see later that data from randomized comparative experiments are the gold standard. The conclusion from 13 such trials is that mammograms reduce the risk of death in women aged 50 to 64 years by 26%.[9]

On average, then, women who have regular mammograms are less likely to die of breast cancer. Of course, the results are different for different women. Some women who have mammograms every year die of breast cancer, and some who never have mammograms live to 100 and die when they crash their motorcycles. In spite of this individual variation, the results of the 13 clinical trials provide convincing evidence that women who have mammograms are less likely to die from breast cancer. That's because probability tells us that the large difference in death rates between women who had regular mammograms and those who didn't was unlikely to have occurred by chance. Can we be *sure* that mammograms reduce risk on the average? No, we can't be sure. **Because variation is everywhere, we cannot be certain about our conclusions.** However, statistics helps us better understand variation so that we can make reasonable conclusions.

Statistics gives us a language for talking about uncertainty that is used and understood by statistically literate people everywhere. In the case of mammograms, the doctors use that language to tell us that "mammography reduces the risk of dying of breast cancer by 26% (95% confidence interval, 17% to 34%)." According to Arthur Nielsen, head of the country's largest market research firm, that 26% is "a shorthand for a range that describes our actual knowledge of the underlying condition."[10] The range is 17% to 34%, and we are 95% confident that the true percent lies in that range. You will soon learn how to understand this language.

We can't escape variation and uncertainty. Learning statistics enables us to deal more effectively with these realities.

Statistical Thinking and You

The purpose of this book is to give you a working knowledge of the ideas and tools of practical statistics. Because data always come from a real-world context, doing statistics means more than just manipulating data. *The Practice of Statistics* is full of data, and each set of data has some brief background to help you understand what the data say. Examples and exercises usually express some brief understanding gained from the data. In practice, you would know much more about the background of the data you work with and about the questions you hope the data will answer. No textbook can be fully realistic. But it is important to form the habit of asking, "What do the data tell me?" rather than just

concentrating on making graphs and doing calculations. This book tries to encourage good habits.

Still, statistics involves lots of calculating and graphing. The text presents the techniques you need, but you should use a calculator or computer software to automate calculations and graphs as much as possible.

Ideas and judgment can't (at least yet!) be automated. They guide you in telling the computer what to do and in interpreting its output. This book tries to explain the most important ideas of statistics, not just teach methods.

You learn statistics by doing statistical problems. This book offers four types of exercises, arranged to help you learn. Short problem sets appear after each major idea. These are straightforward exercises that help you solidify the main points before going on. The Section Exercises at the end of each numbered section help you combine all the ideas of the section. Chapter Review Exercises look back over the entire chapter. Finally, the Part Review Exercises provide challenging, cumulative problems like you might find on a final exam. At each step you are given less advance knowledge of exactly what statistical ideas and skills the problems will require, so each step requires more understanding.

Each chapter ends with a Chapter Review that includes a detailed list of specific things you should now be able to do. Go through that list, and be sure you can say "I can do that" to each item. Then try some chapter exercises.

The basic principle of learning is persistence. The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. Once you put it all together—data analysis, data production, probability, and inference—statistics will help you answer important questions for yourself and for those around you.

Exercises

P.13 TV viewing habits You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student. Give the units of measurement for the quantitative variables.

P.14 Roll the dice What is the probability of getting a "6" if you roll a fair six-sided die? Explain carefully what your answer means.

P.15 Tap water or bottled water, I Refer to Example P.10 (page 22). Which of the following results would provide more convincing evidence that Mr. Bullard's class could tell bottled water from tap water: 12 out of 21 correct identifications or 14 out of 21 correct identifications? Explain your answer.

P.16 Tap water or bottled water, II Refer to Example P.10 (page 22). Estimate the probability of getting 11 or more correct answers if the students were simply guessing. What would you conclude about whether Mr. Bullard's students could distinguish bottled water from tap water?

P.17 Spinning pennies Hold a penny upright on its edge under your forefinger on a hard surface, then snap it with your other forefinger so that it spins for some time before falling.

Is the coin equally likely to land heads or tails? Spin the coin a total of 20 times, recording whether it lands heads or tails each time.

(a) Make a graph like the one in Figure P.5 (page 21) that shows the proportion of heads after each toss.

(b) Based on your results, estimate the proportion of all spins of the coin that would be heads.

(c) What would you conclude about whether the coin lands heads half the time? Justify your answer.

(d) IN CLASS: Pool your results with those of your classmates. Would you change the conclusion you made in (c)? Why or why not?

P.18 Abstinence or not? An August 2004 Gallup Poll asked 439 teens aged 13 to 17 whether they thought young people should abstain from sex until marriage. 56% said "Yes."

(a) If Gallup had asked *all* teens aged 13 to 17 this question, would exactly 56% have said "Yes"? Explain.

(b) In this sample, 48% of the boys and 64% of the girls said "Yes." Are you convinced that a higher percent of girls than boys aged 13 to 17 feel this way? Why or why not?

CASE CLOSED!

Can magnets help reduce pain?

At the end of each chapter, you will be asked to use what you have learned to resolve the Case Study that opened the chapter. Just like in a court proceeding, you can exclaim "Case Closed!" when you have finished.

Start by reviewing the information in the magnets and pain relief Case Study (page 3). Then answer each of the following questions in complete sentences. Be sure to communicate clearly enough for any of your classmates to understand what you are saying.

1. Data analysis
    a. Answer the key questions: who, what, why, when, where, how, and by whom?
    b. Construct separate dotplots of the pain ratings for the individuals in the active- and inactive-magnet groups. Draw your plots one above the other using the same scale.
    c. Describe what you see in your graphs.
    d. Calculate the *mean* (average) pain rating for each group. Now calculate the difference between the two means.

and inactive-magnet groups 10,000 times, keeping each patient's final pain score the same as in the actual experiment. Each time, it computed the difference between the mean pain scores reported by the two groups. The graph displays the values of these 10,000 differences.

3. Probability
   a. Use the graph to estimate what percent of the time the difference in the groups' mean pain ratings is greater than 0. Explain your method.
   b. Based on the graph, how likely is it that the difference in mean pain ratings is greater than the one observed in this study (4.05) if the active magnets don't relieve pain?
4. Inference
   a. What would you estimate is the difference in mean pain relief when using active versus inactive magnets? Why?
   b. If you were testing the claim that the active magnets did not help reduce pain any better than the inactive magnets, what would you conclude? Explain.

Chapter Review

Summary

Statistics is the art and science of collecting, organizing, describing, analyzing, and drawing conclusions from data. When used properly, the tools of statistics can help us answer important questions about the world around us. This chapter gave you an overview of what statistics is all about: *data production, data analysis, probability,* and *statistical inference.*

Some people make decisions based on personal experiences. Statisticians make decisions based on data. **Data production** helps us answer specific questions with an **experiment** or an **observational study.** Experiments are distinguished from observational studies such as **surveys** by doing something intentionally to the individuals involved. A survey selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data about the sample.

A data set contains information on a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary. Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, like male or female. A quantitative variable has numerical values that measure some characteristic of each individual, like height in centimeters or annual salary in

dollars. Remember to ask the key questions—who, what, why, when, where, how, and by whom?—about any data set.

The **distribution** of a variable describes what values the variable takes and how often it takes these values. To describe a distribution, begin with a graph. You can use **bar graphs** to display categorical variables. A **dotplot** is a simple graph you can use to show the distributions of quantitative variables. When examining any graph, ask yourself "What do I see?"

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them. The conclusions of an exploratory analysis may not generalize beyond the specific data studied.

Probability is the language of chance. Chance behavior is unpredictable in the short run but follows a predictable pattern over many repetitions. When we're dealing with chance behavior, the rules of probability help us determine the likelihood of particular outcomes.

Statistical inference produces answers to specific questions, along with a statement of how confident we can be that the answer is correct. The conclusions of statistical inference are usually intended to apply beyond the individuals actually studied. Successful statistical inference requires production of data intended to answer the specific questions posed.

What You Should Have Learned

Here is a review list of the most important skills you should have acquired from your study of this chapter.

A. Where Do Data Come From?

1. Explain why we should not draw conclusions based on personal experiences.
2. Recognize whether a study is an experiment, a survey, or an observational study that is not a survey.
3. Determine the best method for producing data to answer a specific question: experiment, survey, or other observational study.
4. Locate available data on the Internet to help you answer a question of interest.

B. Dealing with Data

1. Identify the individuals and variables in a set of data.
2. Classify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.
3. Answer the key questions—who, what, why, when, where, how, and by whom?—about a given set of data.

C. Describing Distributions

1. Make a bar graph of the distribution of a categorical variable. Interpret bar graphs.

2. Make a dotplot of the distribution of a quantitative variable. Describe what you see.
3. Given a relationship between two variables, identify variables lurking in the background that might affect the relationship.

D. Probability

1. Interpret probability as what happens in the long run.
2. Use simulations to determine how likely an outcome is to occur.

E. Statistical Inference

1. Use the results of simulations and probability calculations to draw conclusions that go beyond the data.
2. Give reasons why conclusions cannot be certain in a given setting.

Web Links

These sites are excellent sources for available data:

U.S. Census Bureau Home Page www.census.gov

Data and Story Library lib.stat.cmu.edu/DASL/

Chapter Review Exercises

P.19 TV violence A typical hour of prime-time television shows three to five violent acts. Linking family interviews and police records shows a clear association between time spent watching TV as a child and later aggressive behavior.[11]

(a) Explain why this is an observational study rather than an experiment.

(b) Suggest several other variables describing a child's home life that may be related to how much TV he or she watches. Explain why these variables make it difficult to conclude that more TV *causes* aggressive behavior.

P.20 How safe are teen drivers? Find some information to help answer this question. Start with the National Highway and Traffic Safety Administration Web site, www.nhtsa.gov. Keep a detailed written record of your search.

P.21 Give it some gas! Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2004 model motor vehicles:

Make and Model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
Acura NSX	Two-seater	Automatic	6	17	24
BMW 330I	Compact	Manual	6	20	30
Cadillac Seville	Midsize	Automatic	8	18	26
Ford F150 2WD	Standard pickup truck	Automatic	6	16	19

Answer the key questions (who, what, why, when, where, how, and by whom?) for these data. Visit the government's fuel economy Web site www.fueleconomy.gov for more information about how these data were produced. For each variable, tell whether it is categorical or quantitative. Be sure to identify the units of measurement for any quantitative variables.

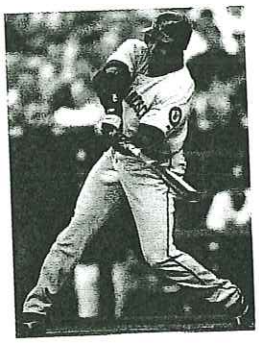
P.22 Wearing bicycle helmets According to the 2003 Youth Risk Behavior Survey, 85.9% of high school students reported rarely or never wearing bicycle helmets. The table below shows additional results from this survey, broken down by gender and grade in school.

	Rarely or never wore bicycle helmets		
Grade	Female (%)	Male (%)	Total (%)
9	80.3	86.4	83.9
10	85.9	88.1	87.1
11	86.8	87.6	87.3
12	86.1	87.5	86.9

(a) Make a bar graph to show the percent of students in each grade who said they rarely or never wore bicycle helmets. Write a few sentences describing what you see.

(b) Now make a side-by-side bar graph to compare the percents of males and females at each grade level who said they rarely or never wore bicycle helmets. Describe what you see in a few sentences.

P.23 Three of a kind You read in a book on poker that the probability of being dealt three of a kind in a five-card poker hand is 1/50. Explain in simple language what this means.



P.24 Baseball and steroids Late in 2004, baseball superstar Barry Bonds admitted using creams and ointments that contained steroids. Bonds said he didn't know that these substances contained steroids. A Gallup Poll asked a random sample of U.S. adults whether they thought Bonds was telling the truth: 42% said "probably not" and 33% said "definitely not."

(a) Why did Gallup survey a random sample of U.S. adults rather than a sample of people attending a Major League Baseball game?

(b) If Gallup had surveyed all U.S. adults instead of a sample, about what percent of the responses would be "probably not"? "Definitely not"? Explain.

(c) Can we conclude based on these results that Barry Bonds is lying? Why or why not?

P.25 Magnets and pain, I Refer to Case Closed! (page 26). Suppose the difference in the mean pain scores of the active and inactive groups had been 2.5 instead of 4.05. What conclusion would you draw about whether magnets help relieve pain in postpolio patients? Explain.

P.26 Magnets and pain, II Refer to the chapter-opening Case Study (page 3). The researchers decided to analyze the patients' final pain ratings. It also makes sense to

examine the *difference* between patients' initial pain ratings and their final pain ratings in the active and inactive groups. Here are the data:

Active: 10, 6, 1, 10, 6, 8, 5, 5, 6, 8, 7, 8, 7, 6, 4, 4, 7, 10, 6, 10, 6, 5, 5, 1, 0, 0, 0, 0, 1
Inactive: 4, 3, 5, 2, 1, 4, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1

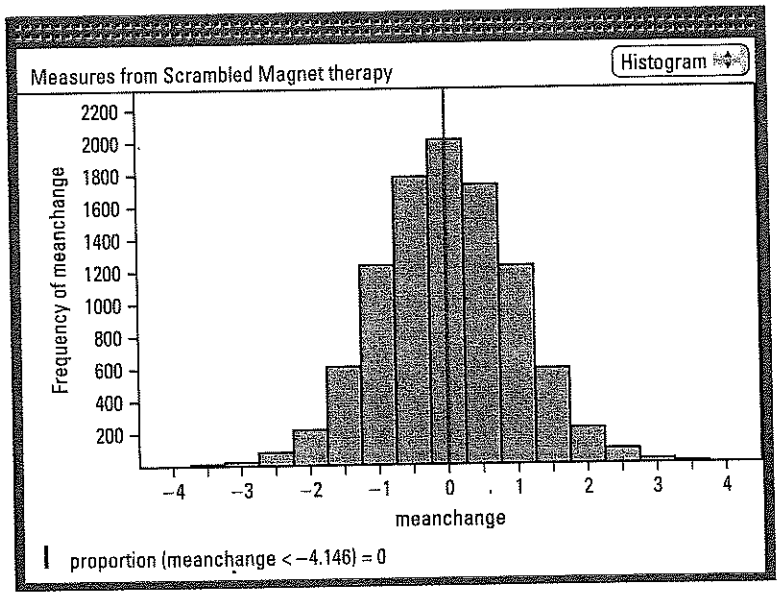
(a) Construct a dotplot for the active group's data. Describe what you see.

(b) Now make a dotplot for the inactive group's data immediately beneath using the same scale as the graph you made in (a). Write a few sentences comparing the changes in pain ratings for patients in the active and inactive groups.

(c) Calculate the mean (average) change in pain rating for each group.

(d) Figure P.8 shows the results of 10,000 repetitions of a computer simulation. As in Case Closed! (page 26), the computer redistributed the patients into the active- and inactive-magnet groups 10,000 times. Each time, it computed the difference between the mean "decrease in pain" scores reported by the two groups. The graph displays the values of these 10,000 differences. If you were testing the claim that the active magnets did not help reduce pain any better than the inactive magnets, what would you conclude? Explain.

Figure P.8 *Graph from Fathom statistical software displaying the difference in average decrease in pain for the two groups in the magnets and pain study for 10,000 trials of a computer simulation.*



P.27 Are you driving a gas guzzler? Table P.2 displays the highway gas mileage for 30 model year 2004 midsize cars.

Table P.2 *Highway gas mileage for 2004 model midsize cars*

Model	MPG	Model	MPG
Acura 3.5RL	24	Jaguar XJR	24
Audi A6 Quattro	25	Lexus GS300	25
BMW 745I	26	Lexus LS430	25
Buick Regal	30	Lincoln-Mercury LS	24
Cadillac Deville	26	Lincoln-Mercury Sable	26
Cadillac Seville	26	Mercedes-Benz E320	27
Chevrolet Malibu	34	Mercedes-Benz E500	20
Chrysler Sebring	30	Mitsubishi Diamante	25
Dodge Stratus	28	Mitsubishi Galant	26
Honda Accord	34	Nissan Maxima	28
Hyundai Sonata	27	Saab 9-3	28
Infiniti G35	26	Saturn L300	28
Infiniti Q45	23	Toyota Camry	33
Jaguar S-Type 3.0	26	Volkswagen Passat	31
Jaguar Vanden Plas	28	Volvo S80	28

Source: U.S. Environmental Protection Agency, *Model Year 2004 Fuel Economy Guide*, found online at www.fueleconomy.gov.

Make a dotplot of these data. Describe what you see in a few sentences.

P.28 Mozart and test scores The Kalamazoo (Michigan) Symphony once advertised a "Mozart for Minors" program with this statement: "Question: Which students scored 51 points higher in verbal skills and 39 points higher in math? Answer: Students who had experience in music."[12]

(a) How do you think these data were obtained—from an experiment, a survey, or an observational study that wasn't a survey? Justify your answer.

(b) Can we conclude that the "Mozart for Minors" program *caused* an increase in students' test scores? Explain. (*Hint:* Think of a variable lurking in the background.)

(c) Describe an experiment to test whether "Mozart for Minors" really leads to higher test scores.

INTRODUCTION Statistics: The Science and Art of Data

LEARNING TARGETS *By the end of the section, you should be able to:*

- Identify the individuals and variables in a set of data.
- Classify variables as categorical or quantitative.

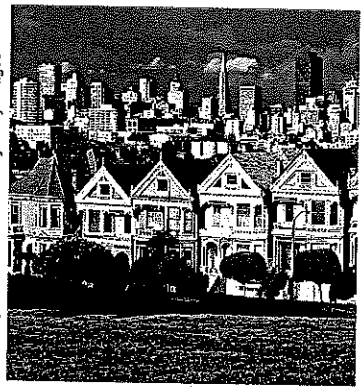
We live in a world of *data*. Every day, the media report poll results, outcomes of medical studies, and analyses of data on everything from stock prices to standardized test scores to global warming. The data are trying to tell us a story. To understand what the data are saying, you need to learn more about **statistics.**

DEFINITION Statistics

Statistics is the science and art of collecting, analyzing, and drawing conclusions from data.

A solid understanding of statistics will help you make good decisions based on data in your daily life.

Organizing Data



Rudy Sulgan/Corbis Documentary/Getty Images

Every year, the U.S. Census Bureau collects data from over 3 million households as part of the American Community Survey (ACS). The table displays some data from the ACS in a recent year.

Household	Region	Number of people	Time in dwelling (years)	Response mode	Household income	Internet access?
425	Midwest	5	2–4	Internet	52,000	Yes
936459	West	4	2–4	Mail	40,500	Yes
50055	Northeast	2	10–19	Internet	481,000	Yes
592934	West	4	2–4	Phone	230,800	No
545854	South	9	2–4	Phone	33,800	Yes
809928	South	2	30+	Internet	59,500	Yes
110157	Midwest	1	5–9	Internet	80,000	Yes
999347	South	1	<1	Mail	8,400	No

Most data tables follow this format—each row describes an **individual** and each column holds the values of a **variable.**

Sometimes the individuals in a data set are called *cases* or *observational units.*

DEFINITION Individual, Variable

An **individual** is an object described in a set of data. Individuals can be people, animals, or things.

A **variable** is an attribute that can take different values for different individuals.

For the American Community Survey data set, the *individuals* are households. The *variables* recorded for each household are region, number of people, time in current dwelling, survey response mode, household income, and whether the dwelling has Internet access. Region, time in dwelling, response mode, and Internet access status are **categorical variables.** Number of people and household income are **quantitative variables.**

Note that household is *not* a variable. The numbers in the household column of the data table are just labels for the individuals in this data set. Be sure to look for a column of labels—names, numbers, or other identifiers—in any data table you encounter.

DEFINITION Categorical variable, Quantitative variable

A **categorical variable** assigns labels that place each individual into a particular group, called a category.

A **quantitative variable** takes number values that are quantities—counts or measurements.

Not every variable that takes number values is quantitative. Zip code is one example. Although zip codes are numbers, they are neither counts of anything, nor measurements of anything. They are simply labels for a regional location, making zip code a categorical variable. Some variables—such as gender, race, and occupation—are categorical by nature. Time in dwelling from the ACS data set is also a categorical variable because the values are recorded as intervals of time, such as 2–4 years. If time in dwelling had been recorded to the nearest year for each household, this variable would be quantitative.

To make life simpler, we sometimes refer to *categorical data* or *quantitative data* instead of identifying the variable as categorical or quantitative.

EXAMPLE Census At School Individuals and Variables

PROBLEM: Census At School is an international project that collects data about primary and secondary school students using surveys. Hundreds of thousands of students from Australia, Canada, Ireland, Japan, New Zealand, South Africa, South Korea, the United Kingdom, and the United States have taken part in the project. Data from the surveys are available online. We used the site's "Random Data Selector" to choose 10 Canadian students who completed the survey in a recent year. The table displays the data.

Garry Black/Alamy

Province	Gender	Number of languages spoken	Handedness	Height (cm)	Wrist circum-ference (mm)	Preferred communication
Saskatchewan	Male	1	Right	175.0	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178.0	174	Facebook
Ontario	Male	2	Right	169.0	160	Cell phone
Ontario	Female	2	Right	166.0	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166.0	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text messaging
Ontario	Female	2	Right	150.5	187	Text messaging
Ontario	Female	1	Right	171.0	180	Text messaging

(a) Identify the individuals in this data set.

(b) What are the variables? Classify each as categorical or quantitative.

SOLUTION:

(a) 10 randomly selected Canadian students who participated in the Census At School survey.

(b) Categorical: Province, gender, handedness, preferred communication method

Quantitative: Number of languages spoken, height (cm), wrist circumference (mm)

We'll see in Chapter 4 why choosing at random, as we did in this example, is a good idea.

There is at least one suspicious value in the data table. We doubt that the girl who is 166 cm tall really has a wrist circumference of 65 mm (about 2.6 inches). Always look to be sure the values make sense!

FOR PRACTICE, TRY EXERCISE 1

AP® EXAM TIP

If you learn to distinguish categorical from quantitative variables now, it will pay big rewards later. You will be expected to analyze categorical and quantitative variables correctly on the AP® exam.

The proper method of data analysis depends on whether a variable is categorical or quantitative. For that reason, it is important to distinguish these two types of variables. The type of data determines what kinds of graphs and which numerical summaries are appropriate.

ANALYZING DATA A variable generally takes values that vary (hence the name *variable*!). Categorical variables sometimes have similar counts in each category and sometimes don't. For instance, we might have expected similar numbers of males and females in the Census At School data set. But we aren't surprised to see that most students are right-handed. Quantitative variables may take values that are very close together or values that are quite spread out. We call the pattern of variation of a variable its **distribution.**

DEFINITION Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes those values.

Let's return to the data for the sample of 10 Canadian students from the preceding example. Figure 1.1(a) shows the distribution of preferred communication

method for these students in a *bar graph*. We can see how many students chose each method from the heights of the bars: cell phone (2), Facebook (1), in person (3), text messaging (4). Figure 1.1(b) shows the distribution of number of languages spoken in a *dotplot*. We can see that 6 students speak one language, 3 students speak two languages, and 1 student speaks four languages.

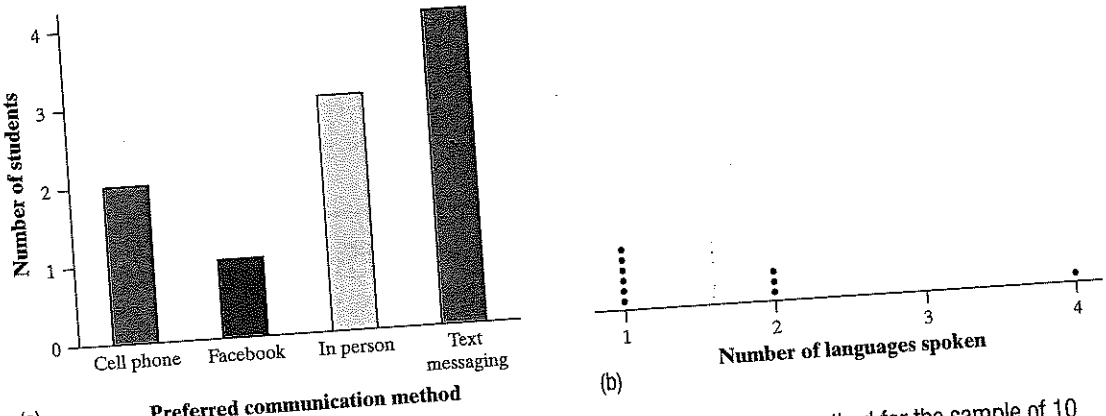


FIGURE 1.1 (a) Bar graph showing the distribution of preferred communication method for the sample of 10 Canadian students. (b) Dotplot showing the distribution of number of languages spoken by these students.

Section 1.1 begins by looking at how to describe the distribution of a single categorical variable and then examines relationships between categorical variables. Sections 1.2 and 1.3 and all of Chapter 2 focus on describing the distribution of a quantitative variable. Chapter 3 investigates relationships between two quantitative variables. In each case, we begin with graphical displays, then add numerical summaries for a more complete description.

HOW TO ANALYZE DATA

- Begin by examining each variable by itself. Then move on to study relationships among the variables.
- Start with a graph or graphs. Then add numerical summaries.

CHECK YOUR UNDERSTANDING

Jake is a car buff who wants to find out more about the vehicles that his classmates drive. He gets permission to go to the student parking lot and record some data. Later, he does some Internet research on each model of car he found. Finally, Jake makes a spreadsheet that includes each car's license plate, model, year, color, highway gas mileage, weight, and whether it has a navigation system.

1. Identify the individuals in Jake's study.
2. What are the variables? Classify each as categorical or quantitative.

From Data Analysis to Inference

Sometimes we're interested in drawing conclusions that go beyond the data at hand. That's the idea of *inference*. In the "Census At School" example, 9 of the 10 randomly selected Canadian students are right-handed. That's 90% of the *sample*. Can we conclude that exactly 90% of the *population* of Canadian students who participated in Census At School are right-handed? No.

If another random sample of 10 students were selected, the percent who are right-handed might not be exactly 90%. Can we at least say that the actual population value is "close" to 90%? That depends on what we mean by "close." The following activity gives you an idea of how statistical inference works.

ACTIVITY Hiring discrimination—it just won't fly!

An airline has just finished training 25 pilots—15 male and 10 female—to become captains. Unfortunately, only eight captain positions are available right now. Airline managers announce that they will use a lottery to determine which pilots will fill the available positions. The names of all 25 pilots will be written on identical slips of paper. The slips will be placed in a hat, mixed thoroughly, and drawn out one at a time until all eight captains have been identified.

Choja/Getty Images

A day later, managers announce the results of the lottery. Of the 8 captains chosen, 5 are female and 3 are male. Some of the male pilots who weren't selected suspect that the lottery was not carried out fairly. One of these pilots asks your statistics class for advice about whether to file a grievance with the pilots' union.

The key question in this possible discrimination case seems to be: *Is it plausible (believable) that these results happened just by chance?* To find out, you and your classmates will *simulate* the lottery process that airline managers said they used.

1. Your teacher will give you a bag with 25 beads (15 of one color and 10 of another) or 25 slips of paper (15 labeled "M" and 10 labeled "F") to represent the 25 pilots. Mix the beads/slips thoroughly. Without looking, remove 8 beads/slips from the bag. Count the number of female pilots selected. Then return the beads/slips to the bag.
2. Your teacher will draw and label a number line for a class *dotplot*. On the graph, plot the number of females you got in Step 1.
3. Repeat Steps 1 and 2 if needed to get a total of at least 40 simulated lottery results for your class.
4. Discuss the results with your classmates. Does it seem plausible that airline managers conducted a fair lottery? What advice would you give the male pilot who contacted you?

Our ability to do inference is determined by how the data are produced. Chapter 4 discusses the two main methods of data production—sampling

and experiments—and the types of conclusions that can be drawn from each. As the activity illustrates, the logic of inference rests on asking, "What are the chances?" *Probability*, the study of chance behavior, is the topic of Chapters 5–7. We'll introduce the most common inference techniques in Chapters 8–12.

Introduction Summary

- **Statistics** is the science and art of collecting, analyzing, and drawing conclusions from data.
- A data set contains information about a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- A **categorical variable** assigns a label that places each individual in one of several groups, such as male or female. A **quantitative variable** has numerical values that count or measure some characteristic of each individual, such as number of siblings or height in meters.
- The **distribution** of a variable describes what values the variable takes and how often it takes them.

Introduction Exercises

The solutions to all exercises numbered in red may be found in the Solutions Appendix, starting on page S-1.

1. pg 3 **A class survey** Here is a small part of the data set that describes the students in an AP® Statistics class. The data come from anonymous responses to a questionnaire filled out on the first day of class.

Gender	Grade level	GPA	Children in family	Homework last night (min)	Android or iPhone?
F	9	2.3	3	0–14	iPhone
M	11	3.8	6	15–29	Android
M	10	3.1	2	15–29	Android
F	10	4.0	1	45–59	iPhone
F	10	3.4	4	0–14	iPhone
F	10	3.0	3	30–44	Android
M	9	3.9	2	15–29	iPhone
M	12	3.5	2	0–14	iPhone

(a) Identify the individuals in this data set.

(b) What are the variables? Classify each as categorical or quantitative.

2. **Coaster craze** Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The following table displays data on several roller coasters that were opened in a recent year.[1]

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (sec)
Wildfire	Wood	187.0	Sit down	70.2	120
Skyline	Steel	131.3	Inverted	50.0	90
Goliath	Wood	165.0	Sit down	72.0	105
Helix	Steel	134.5	Sit down	62.1	130
Banshee	Steel	167.0	Inverted	68.0	160
Black Hole	Steel	22.7	Sit down	25.5	75

(a) Identify the individuals in this data set.

(b) What are the variables? Classify each as categorical or quantitative.

3. **Hit movies** According to the Internet Movie Database, *Avatar* is tops based on box-office receipts worldwide as of January 2017. The following table displays data on several popular movies. Identify the individuals and variables in this data set. Classify each variable as categorical or quantitative.

43. The following partially completed two-way table shows the marginal distributions of gender and handedness for a sample of 100 high school students.

		Gender		
		Male	Female	Total
Dominant hand	Right	x		90
	Left			10
	Total	40	60	100

If there is no association between gender and handedness for the members of the sample, which of the following is the correct value of x?

(a) 20
(b) 30
(c) 36
(d) 45
(e) Impossible to determine without more information.

Recycle and Review

44. **Hotels (Introduction)** A high school lacrosse team is planning to go to Buffalo for a three-day tournament. The tournament's sponsor provides a list of available hotels, along with some information about each hotel. The following table displays data about hotel options. Identify the individuals and variables in this data set. Classify each variable as categorical or quantitative.

Hotel	Pool	Exercise room?	Internet ($/day)	Restaurants	Distance to site (mi)	Room service?	Room rate ($/day)
Comfort Inn	Out	Y	0.00	1	8.2	Y	149
Fairfield Inn & Suites	In	Y	0.00	1	8.3	N	119
Baymont Inn & Suites	Out	Y	0.00	1	3.7	Y	60
Chase Suite Hotel	Out	N	15.00	0	1.5	N	139
Courtyard	In	Y	0.00	1	0.2	Dinner	114
Hilton	In	Y	10.00	2	0.1	Y	156
Marriott	In	Y	9.95	2	0.0	Y	145

SECTION 1.2 Displaying Quantitative Data with Graphs

LEARNING TARGETS *By the end of the section, you should be able to:*

- Make and interpret dotplots, stemplots, and histograms of quantitative data.
- Identify the shape of a distribution from a graph.
- Describe the overall pattern (shape, center, and variability) of a distribution and identify any major departures from the pattern (outliers).
- Compare distributions of quantitative data using dotplots, stemplots, and histograms.

To display the distribution of a categorical variable, use a bar graph or a pie chart. How can we picture the distribution of a quantitative variable? In this section, we present several types of graphs that can be used to display quantitative data.

Dotplots

One of the simplest graphs to construct and interpret is a **dotplot.**

DEFINITION Dotplot

A **dotplot** shows each data value as a dot above its location on a number line.

Here are data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

Figure 1.5 shows a dotplot of these data.

FIGURE 1.5 Dotplot of goals scored in 20 games by the 2016 U.S. women's soccer team.

It is fairly easy to make a dotplot by hand for small sets of quantitative data.

HOW TO MAKE A DOTPLOT

- **Draw and label the axis.** Draw a horizontal axis and put the name of the quantitative variable underneath. Be sure to include units of measurement.
- **Scale the axis.** Look at the smallest and largest values in the data set. Start the horizontal axis at a convenient number equal to or less than the smallest value and place tick marks at equal intervals until you equal or exceed the largest value.
- **Plot the values.** Mark a dot above the location on the horizontal axis corresponding to each data value. Try to make all the dots the same size and space them out equally as you stack them.

Remember what we said in Section 1.1: Making a graph is not an end in itself. When you look at a graph, always ask, "What do I see?" From Figure 1.5, we see that the 2016 U.S. women's soccer team scored 4 or more goals in 6/20 = 0.30 or 30% of its games. That's quite an offense! Unfortunately, the team lost to Sweden on penalty kicks in the 2016 Summer Olympics.

EXAMPLE Give it some gas! Making and interpreting dotplots


Ann Heath

PROBLEM: The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy ratings for cars. To estimate fuel economy, the EPA performs tests on several vehicles of the same make, model, and year. Here are data on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA:

22.4 22.4 22.3 23.3 22.3 22.3 22.5 22.4 22.1 21.5 22.0 22.2 22.7
22.8 22.4 22.6 22.9 22.5 22.1 22.4 22.2 22.9 22.6 21.9 22.4

(a) Make a dotplot of these data.

(b) Toyota reports the highway gas mileage of its 2018 model year 4Runners as 22 mpg. Do these data give the EPA sufficient reason to investigate that claim?

SOLUTION:

(a)

(b) No. 23 of the 25 cars tested had an estimated highway fuel economy of 22 mpg or greater.

To make the dotplot:
- **Draw and label the axis.** Note variable name and units in the label.
- **Scale the axis.** The smallest value is 21.5 and the largest value is 23.3. So we choose a scale from 21.5 to 23.5 with tick marks 0.1 units apart.
- **Plot the values.**

FOR PRACTICE, TRY EXERCISE 45

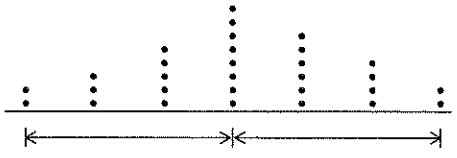
Describing Shape

When you describe the shape of a dotplot or another graph of quantitative data, focus on the main features. Look for major *peaks*, not for minor ups and downs in the graph. Look for *clusters* of values and obvious *gaps*. Decide if the distribution is roughly **symmetric** or clearly **skewed**.

We could also describe a distribution with a long tail to the left as "skewed toward negative values" or "negatively skewed" and a distribution with a long right tail as "positively skewed."

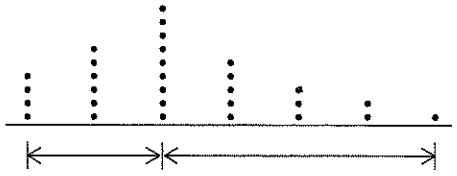
DEFINITION Symmetric and skewed distributions

A distribution is roughly **symmetric** if the right side of the graph (containing the half of observations with the largest values) is approximately a mirror image of the left side.



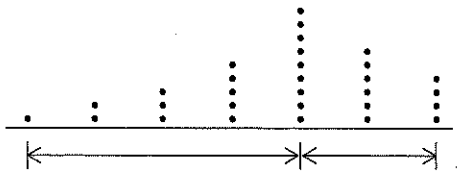
Roughly symmetric

A distribution is **skewed to the right** if the right side of the graph is much longer than the left side.

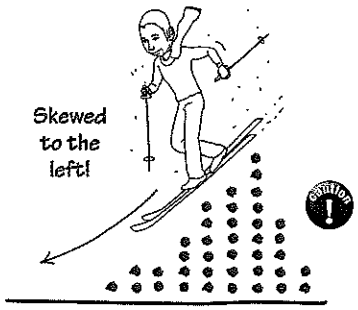


Skewed to the right

A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.

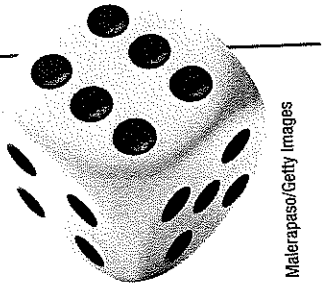


Skewed to the left



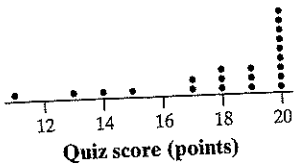
For ease, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed to the right." **The direction of skewness is toward the long tail, not the direction where most observations are clustered.** The drawing is a cute but corny way to help you keep this straight. To avoid danger, Mr. Starnes skis on the gentler slope—in the direction of the skewness.

EXAMPLE Quiz scores and die rolls Describing shape

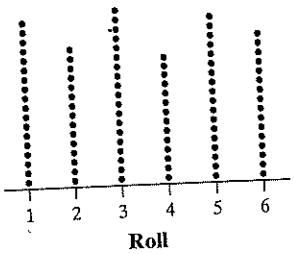

Malerapaso/Getty Images

PROBLEM: The dotplots display two different sets of quantitative data. Graph (a) shows the scores of 21 statistics students on a 20-point quiz. Graph (b) shows the results of 100 rolls of a 6-sided die. Describe the shape of each distribution.

(a)



(b)



SOLUTION:

(a) The distribution of statistics quiz scores is skewed to the left, with a single peak at 20 (a perfect score). There are two small gaps at 12 and 16.

(b) The distribution of die rolls is roughly symmetric. It has no clear peak.

We can describe the shape of the distribution in part (b) as "approximately uniform" because the frequencies are about the same for all possible rolls.

FOR PRACTICE, TRY EXERCISE 49

Some people refer to graphs with a single peak as *unimodal*, to graphs with two peaks as *bimodal*, and to graphs with more than two clear peaks as *multimodal*.

Some quantitative variables have distributions with easily described shapes. But many distributions have irregular shapes that are neither symmetric nor skewed. Some distributions show other patterns, like the dotplot in Figure 1.6. This graph shows the durations (in minutes) of 220 eruptions of the Old Faithful geyser. The dotplot has two distinct clusters and two peaks: one at about 2 minutes and one at about 4.5 minutes. When you examine a graph of quantitative data, describe any pattern you see as clearly as you can.

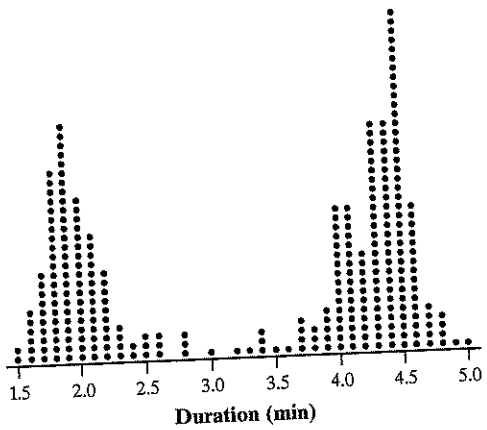


FIGURE 1.6 Dotplot displaying duration (in minutes) of 220 Old Faithful eruptions. This graph has two distinct clusters and two clear peaks.

Some quantitative variables have distributions with predictable shapes. Many biological measurements on individuals from the same species and gender—lengths of bird bills, heights of young women—have roughly symmetric distributions. Salaries and home prices, on the other hand, usually have right-skewed distributions. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a strong right skew.

CHECK YOUR UNDERSTANDING

Knoebels Amusement Park in Elysburg, Pennsylvania, has earned acclaim for being an affordable, family-friendly entertainment venue. Knoebels does not charge for general admission or parking, but it does charge customers for each ride they take. How much do the rides cost at Knoebels? The table shows the cost for each ride in a sample of 22 rides in a recent year.

Name	Cost	Name	Cost
Merry Mixer	$1.50	Looper	$1.75
Italian Trapeze	$1.50	Flying Turns	$3.00
Satellite	$1.50	Flyer	$1.50
Galleon	$1.50	The Haunted Mansion	$1.75
Whipper	$1.25	StratosFear	$2.00
Skooters	$1.75	Twister	$2.50
Ribbit	$1.25	Cosmotron	$1.75
Roundup	$1.50	Paratrooper	$1.50
Paradrop	$1.25	Downdraft	$1.50
The Phoenix	$2.50	Rockin' Tug	$1.25
Gasoline Alley	$1.75	Sklooosh!	$1.75

1. Make a dotplot of the data.
2. Describe the shape of the distribution.

Describing Distributions

Here is a general strategy for describing a distribution of quantitative data.

HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

In any graph, look for the *overall pattern* and for clear *departures* from that pattern.

- You can describe the overall pattern of a distribution by its **shape, center,** and **variability.**
- An important kind of departure is an **outlier,** an observation that falls outside the overall pattern.

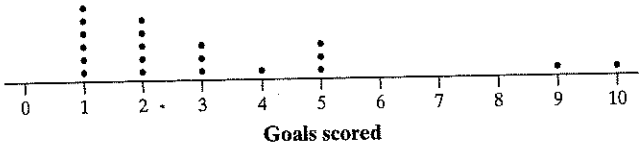
Variability is sometimes referred to as *spread.* We prefer variability because students sometimes think that spread refers only to the distance between the maximum and minimum value of a quantitative data set (the *range*). There are several ways to measure the variability (spread) of a distribution, including the range.

AP® EXAM TIP

Always be sure to include context when you are asked to describe a distribution. This means using the variable name, not just the units the variable is measured in.

We will discuss more formal ways to measure center and variability and to identify outliers in Section 1.3. For now, just use the *median* (middle value in the ordered data set) when describing center and the *minimum* and *maximum* when describing variability.

Let's practice with the dotplot of goals scored in 20 games played by the 2016 U.S. women's soccer team.



When describing a distribution of quantitative data, don't forget: **S**tatistical **O**pinions **C**an **V**ary (**S**hape, **O**utliers, **C**enter, **V**ariability).

Shape: The distribution of goals scored is skewed to the right, with a single peak at 1 goal. There is a gap between 5 and 9 goals.

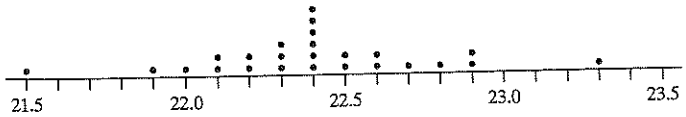
Outliers: The games when the team scored 9 and 10 goals appear to be outliers.

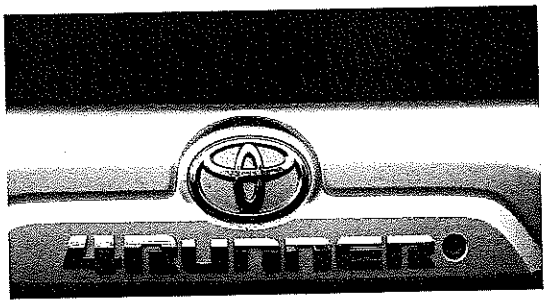
Center: The median is 2 goals scored.

Variability: The data vary from 1 to 10 goals scored.

EXAMPLE Give it some gas! Describing a distribution

PROBLEM: Here is a dotplot of the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA. Describe the distribution.





Daren Starnes

SOLUTION:

Shape: The distribution of highway fuel economy ratings is roughly symmetric, with a single peak at 22.4 mpg. There are two clear gaps: between 21.5 and 21.9 mpg and between 22.9 and 23.3 mpg.

Outliers: The cars with 21.5 mpg and 23.3 mpg ratings are possible outliers.

Center: The median rating is 22.4 mpg.

Variability: The ratings vary from 21.5 to 23.3 mpg.

Be sure to include context by discussing the variable of interest, highway fuel economy ratings. And give the units of measurement: miles per gallon (mpg).

FOR PRACTICE, TRY EXERCISE 53

Section 1.2 Summary

- You can use a **dotplot, stemplot,** or **histogram** to show the distribution of a quantitative variable. A dotplot displays individual values on a number line. Stemplots separate each observation into a stem and a one-digit leaf. Histograms plot the frequencies (counts) or relative frequencies (proportions or percents) of values in equal-width intervals.
- Some distributions have simple shapes, such as **symmetric, skewed to the left,** or **skewed to the right.** The number of peaks is another aspect of overall shape. So are distinct clusters and gaps.
- When examining any graph of quantitative data, look for an *overall pattern* and for clear *departures* from that pattern. **Shape, center,** and **variability** describe the overall pattern of the distribution of a quantitative variable. **Outliers** are observations that lie outside the overall pattern of a distribution.
- When comparing distributions of quantitative data, be sure to compare shape, center, variability, and possible outliers.
- Remember: histograms are for quantitative data; bar graphs are for categorical data. Be sure to use relative frequencies when comparing data sets of different sizes.

1.2 Technology Corner

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

2. Making histograms Page 43

Section 1.2 Exercises

45. **Feeling sleepy?** Students in a high school statistics class pg 31 responded to a survey designed by their teacher. One of the survey questions was "How much sleep did you get last night?" Here are the data (in hours):

9	6	8	7	8	8	6	6.5	7	7	9.0	4	3	4
5	6	11	6	3	7	6	10.0	7	8	4.5	9	7	7

(a) Make a dotplot to display the data.

(b) Experts recommend that high school students sleep at least 9 hours per night. What proportion of students in this class got the recommended amount of sleep?

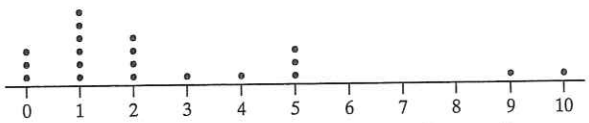
46. **Easy reading?** Here are data on the lengths of the first 25 words on a randomly selected page from Toni Morrison's *Song of Solomon*:

2	3	4	10	2	11	2	8	4	3	7	2	7
5	3	6	4	4	2	5	8	2	3	4	4	

(a) Make a dotplot of these data.

(b) Long words can make a book hard to read. What percentage of words in the sample have 8 or more letters?

47. **U.S. women's soccer—2016** Earlier, we examined data on the number of goals scored by the 2016 U.S. women's soccer team in 20 games played. The following dotplot displays the goal differential for those same games, computed as U.S. goals scored minus opponent goals scored.

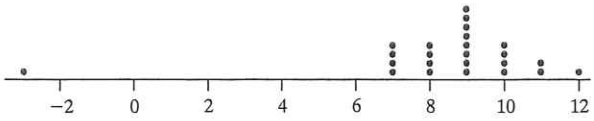


Difference (U.S. − Opponent) in goals scored

(a) Explain what the dot above 3 represents.

(b) What does the graph tell us about how well the team did in 2016? Be specific.

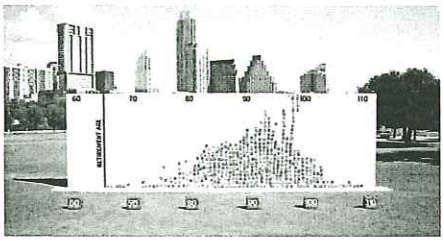
48. **Fuel efficiency** The dotplot shows the difference (Highway − City) in EPA mileage ratings, in miles per gallon (mpg) for each of 24 model year 2018 cars.



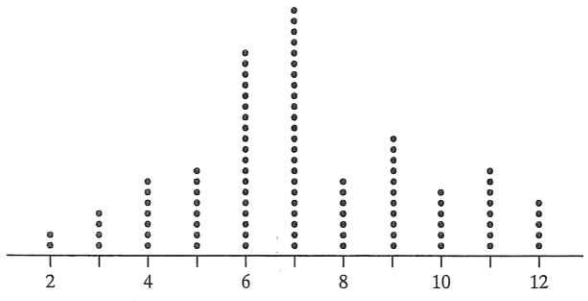
(a) Explain what the dot above −3 represents.

(b) What does the graph tell us about fuel economy in the city versus on the highway for these car models? Be specific.

49. **Getting older** How old is the oldest person you know? Prudential Insurance Company asked 400 people to place a blue sticker on a huge wall next to the age of the oldest person they have ever known. An image of the graph is shown here. Describe the shape of the distribution.



50. **Pair-a-dice** The dotplot shows the results of rolling a pair of fair, six-sided dice and finding the sum of the up-faces 100 times. Describe the shape of the distribution.



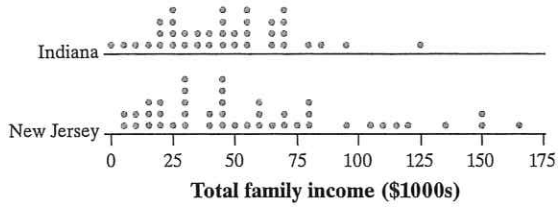
51. **Feeling sleepy?** Refer to Exercise 45. Describe the shape of the distribution.

52. **Easy reading?** Refer to Exercise 46. Describe the shape of the distribution.

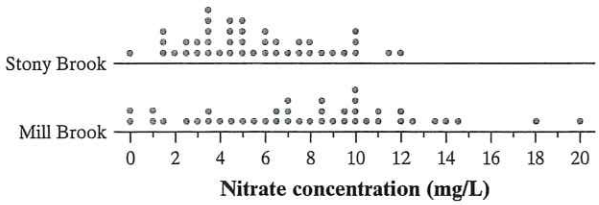
53. **U.S. women's soccer—2016** Refer to Exercise 47. Describe the distribution.

54. **Fuel efficiency** Refer to Exercise 48. Describe the distribution.

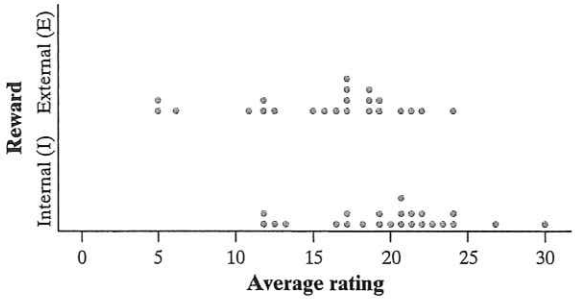
55. **Making money** The parallel dotplots show the total family income of randomly chosen individuals from Indiana (38 individuals) and New Jersey (44 individuals). Compare the distributions of total family incomes in these two samples.



56. **Healthy streams** Nitrates are organic compounds that are a main ingredient in fertilizers. When those fertilizers run off into streams, the nitrates can have a toxic effect on fish. An ecologist studying nitrate pollution in two streams measures nitrate concentrations at 42 places on Stony Brook and 42 places on Mill Brook. The parallel dotplots display the data. Compare the distributions of nitrate concentration in these two streams.



57. **Enhancing creativity** Do external rewards—things like money, praise, fame, and grades—promote creativity? Researcher Teresa Amabile recruited 47 experienced creative writers who were college students and divided them at random into two groups. The students in one group were given a list of statements about external reasons (E) for writing, such as public recognition, making money, or pleasing their parents. Students in the other group were given a list of statements about internal reasons (I) for writing, such as expressing yourself and enjoying wordplay. Both groups were then instructed to write a poem about laughter. Each student's poem was rated separately by 12 different poets using a creativity scale.[26] These ratings were averaged to obtain an overall creativity score for each poem. Parallel dotplots of the two groups' creativity scores are shown here.



(a) Is the variability in creativity scores similar or different for the two groups? Justify your answer.

(b) Do the data suggest that external rewards promote creativity? Justify your answer.