Displaying Distributions with Graphs

11

AP Photo/Daniel Ochoa de Olza

CASE STUDY Nutritionists tell us that a healthy diet should include 20 to 35 grams of fiber daily. Cereal manufacturers, hoping to attract health-conscious consumers, advertise their products as "high-fiber" and other healthy options. The food label on the side of a box of cereal (mandated by the Food and Drug Administration) provides information that allows the consumer to choose a healthy breakfast cereal.

If you go to the breakfast cereal aisle of your local grocery store, you will find lots of different cereals displayed. You could examine all the boxes to see how much fiber each contains, but how do you make sense of all the numbers? Is your favorite cereal, Wheaties, with 3 grams of dietary fiber, among those with the highest fiber content? You could make a list of the fiber content of all the cereals, but it can be difficult to see patterns in large lists. As we saw in Chapter 10, graphs are a powerful way to make sense of large collections of numbers.

In this chapter, we will study two types of graphics—histograms and stemplots (also called stem-and-leaf plots)—that help us make sense of large lists. By the end of this chapter, you will know how to make them and what to look for when you study one of these graphs.

Histograms

Categorical variables record group membership, such as the marital status of a man or the race of a college student. We can use a pie chart or bar graph to display the distribution of categorical variables because they have relatively few values. What about quantitative variables such as the SAT scores of students admitted to a college or the income of families? These variables take so many values that a graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of a quantitative variable is a **histogram**.

Please change this line of text for Step 1 to black. Same for Step 2 and 3 on next page

EXAMPLE 1 How to make a histogram

Table 11.1 presents the percentage of residents aged 65 years and over in each of the 50 states. To make a histogram of this distribution, proceed as follows.

Step 1. Divide the range of the data into classes of equal width. The data in Table 11.1 range from 7.3 to 17.4, so we choose as our classes

$$7.0 \leq \text{percentage over } 65 < 8.0$$

$$8.0 \leq \text{percentage over } 65 < 9.0$$

$$\vdots$$

$$17.0 \leq \text{percentage over } 65 < 18.0$$

Be sure to specify the classes precisely so that each individual falls into exactly one class. In other words, be sure that the classes are *exclusive* (no individual is in more than one class) and *exhaustive* (every individual appears in some class). A state with 7.9% of its residents aged 65 or older would fall into the first class, but 8.0% fall into the second.

TABLE 11.1 Percentage of residents aged 65 and over in the states, July 2008

State	Percent	State	Percent	State	Percent
Alabama	13.8	Louisiana	12.3	Ohio	13.7
Alaska	7.3	Maine	15.1	Oklahoma	13.5
Arizona	13.3	Maryland	12.1	Oregon	13.3
Arkansas	14.3	Massachusetts	13.4	Pennsylvania	15.4
California	11.2	Michigan	13.0	Rhode Island	14.1
Colorado	10.4	Minnesota	12.5	South Carolina	13.3
Connecticut	13.7	Mississippi	12.7	South Dakota	14.4
Delaware	13.9	Missouri	13.6	Tennessee	13.2
Florida	17.4	Montana	14.2	Texas	10.2
Georgia	10.1	Nebraska	13.5	Utah	9.0
Hawaii	14.8	Nevada	11.4	Vermont	14.0
Idaho	12.0	New Hampshire	12.9	Virginia	12.1
Illinois	12.2	New Jersey	13.3	Washington	12.0
Indiana	12.8	New Mexico	13.1	West Virginia	15.7
Iowa	14.8	New York	13.4	Wisconsin	13.3
Kansas	13.1	North Carolina	12.4	Wyoming	12.3
Kentucky	13.3	North Dakota	14.7		

Source: *2010 Statistical Abstract of the United States*; available online at www.census.gov/library/publications/2009/compendia/statab/129ed.html.

Step 2. Count the number of individuals in each class. Here are the counts:

Class	Count	Class	Count	Class	Count
7.0 to 7.9	1	11.0 to 11.9	2	15.0 to 15.9	3
8.0 to 8.9	0	12.0 to 12.9	12	16.0 to 16.9	0
9.0 to 9.9	1	13.0 to 13.9	19	17.0 to 17.9	1
10.0 to 10.9	3	14.0 to 14.9	8		

Step 3. Draw the histogram. Mark on the horizontal axis the scale for the variable whose distribution you are displaying. That's "percentage of residents aged 65 and over" in this example. The scale runs from 5 to 20 because that range spans the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 11.1 is our histogram.

Just as with bar graphs, our eyes respond to the area of the bars in a histogram. Be sure that the classes for a histogram have equal widths. There is no one right choice for the number of classes. Some people recommend between 10 and 20 classes but suggest using fewer when

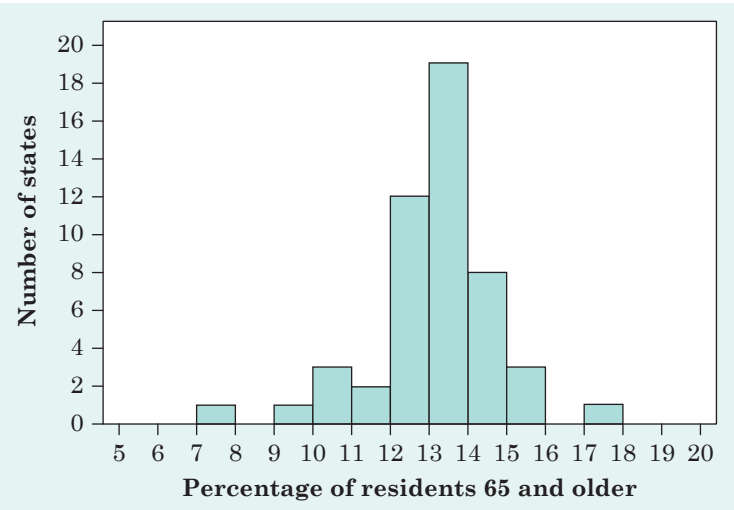

Figure 11.1 Histogram of the percentages of residents aged 65 and older in the 50 states, Example 1. Note the two outliers.

Statistics in Your World

What the eye really sees We make the bars in bar graphs and histograms equal in width because the eye responds to their area. That's roughly true. Careful study by statistician William Cleveland shows that our eyes "see" the size of a bar in proportion to the 0.7 power of its area. Suppose, for example, that one figure in a pictogram is both twice as high and twice as wide as another. The area of the bigger figure is 4 times that of the smaller. But we perceive the bigger figure as only 2.6 times the size of the smaller because 2.6 is 4 to the 0.7 power.

the size of the data set is small. Too few classes will give a "skyscraper" histogram, with all values in a few classes with tall bars. Too many classes will produce a "pancake" graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistics software will choose the classes for you and may use slightly different rules than those we have discussed. The computer's choice is usually a good one, but you can change it if you want. When using statistical software, it is good practice to check what rules are used to determine the classes.

NOW IT'S YOUR TURN

11.1 Personal record for weightlifting. Bodyshop Athletics keeps a dry erase board for members to keep track of their personal record for various events. The "dead lift" is a weightlifting maneuver where a barbell is lifted from the floor to the hips. The following data are the personal records for members at Bodyshop Athletics in pounds lifted during the dead lift.

Member	Weight	Member	Weight	Member	Weight
Baker, B.	175	G.T.C.	250	Pender	205
Baker, T.	100	Harper	155	Porth	215
Birnie	325	Horel	215	Ross	115
Bonner	155	Hureau	285	Stapp	190
Brown	235	Ingram	165	Stokes	305
Burton	155	Johnson	175	Taylor, A.	165
Coffey, L.	135	Jones, J.	195	Taylor, Z.	305
Coffey, S.	275	Jones, L.	205	Thompson	285
Collins, C.	215	LaMonica	235	Trent	135
Collins, E.	95	Lee	165	Tucker	245
Dalick, B.	225	Lord	405	Watson	155
Dalick, K.	335	McCurry	165	Wind, J.	350
Edens	255	Moore	145	Wind, K.	185
Flowers	205	Morrison	145		

Make a histogram of this distribution following the three steps described in Example 1. Create your classes using $75 \leq \text{weight} < 125$, then $125 \leq 175$, and so on.

Interpreting histograms

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you (or your computer) make a graph, always ask, "What do I see?" Here is a general strategy for looking at graphs.

Pattern and deviations

In any graph of data, look for an **overall pattern** and also for striking **deviations** from that pattern.

We have already applied this strategy to line graphs. Trend and seasonal variation are common overall patterns in a line graph. The more drastic decrease in unemployment in mid-2010 for those with less than a high school degree, seen in Figure 10.7 (page 224), is deviating from the overall pattern of small dips and increases apparent throughout the rest of the line. This is an example of a striking deviation from the general pattern that one observes for the rest of the time period. In the case of the histogram of Figure 11.1, it is easiest to begin with deviations from the overall pattern of the histogram. Two states stand out as separated from the main body of the histogram. You can find them in the table once the histogram has called attention to them. Alaska has 7.3% and Florida 17.4% of its residents over age 65. These states are clear *outliers*.

Outliers

An **outlier** in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Is Utah, with 9.0% of its population over 65, an outlier? Whether an observation is an outlier is, to some extent, a matter of judgment, although statisticians have developed some objective criteria for identifying possible outliers. Utah is the smallest of the main body of observations and, unlike Alaska and Florida, is not separated from the general pattern. We would not call Utah an outlier. Once you have spotted outliers, look for an explanation. Many outliers are due to mistakes, such as typing 4.0 as 40. Other outliers point to the special nature of some observations. Explaining outliers usually requires some background information. It is not surprising that Alaska, the northern frontier, has few residents 65 and over and that Florida, a popular state for retirees, has many residents 65 and over.

To see the *overall pattern* of a histogram, ignore any outliers. Here is a simple way to organize your thinking.

Overall pattern of a distribution

To describe the overall pattern of a distribution:

- Describe the **center** and the **variability**.
- Describe the **shape** of the histogram in a few words.

We will learn how to describe center and variability numerically in Chapter 12. For now, we can describe the center of a distribution by its *midpoint*, the value at roughly the middle of all the values in the distribution. We can describe the variability of a distribution by giving the *smallest and largest values*, ignoring any outliers.

EXAMPLE 2 Describing distributions

Look again at the histogram in Figure 11.1. **Shape:** The distribution has a *single peak*. It is roughly *symmetric*—that is, the pattern is similar on both sides of the peak. **Center:** The midpoint of the distribution is close to the single peak, at about 13%. **Variability:** The variability is about 9% to 18% if we ignore the outliers.

Caiaimage/Robert Daly/ Getty Images

EXAMPLE 3 Tuition and fees in Illinois

There are 116 colleges and universities in Illinois. Their tuition and fees for the 2009–2010 school year run from $1974 at Moraine Valley Community College to $38,550 at the University of Chicago. Figure 11.2 is a histogram of the tuition and fees charged by all 116 Illinois colleges and universities. We see that many (mostly community colleges) charge less than $4000. The distribution extends out to the right. At the upper extreme, two colleges charge between $36,000 and $40,000.

The distribution of tuition and fees at Illinois colleges, shown in Figure 11.2, has a quite different **shape** from the distribution in Figure 11.1. There is a strong *peak* in the lowest cost class. Most colleges charge less than $8000, but there is a long right tail extending up to almost $40,000. We call a distribution with a long tail on one side *skewed*. The **center** is roughly $8000 (half the colleges charge less than this). The **variability** is large, from $1974 to more than $38,000. There are no outliers—the colleges with the highest tuition just continue the long right tail that is part of the overall pattern.

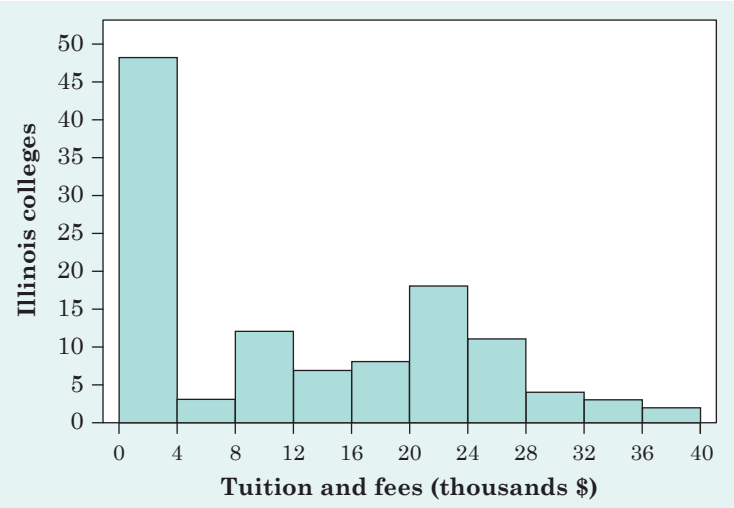

Figure 11.2 Histogram of the tuition and fees charged by 116 Illinois colleges and universities in the 2009–2010 academic year, Example 3. (Data from www.isac.org.)

When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram like those in Figure 11.2. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

Symmetric and skewed distributions

A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

A distribution is symmetric if the two sides of a figure like a histogram are exact mirror images of each other. Data are almost never exactly symmetric, so we are willing to call histograms like that in Figure 11.1 roughly symmetric as an overall description. The tuition distribution in Figure 11.2, on the other hand, is clearly skewed to the right. Here are more examples.

EXAMPLE 4 Lake elevation levels

Lake Murray is a manmade reservoir located in South Carolina. It is used mainly for recreation, such as boating, fishing, and water sports. It is also used to provide backup hydroelectric power for South Carolina Electric and Gas. The lake levels fluctuate with the highest levels in summer (for safe boating and good fishing) and the lowest levels in winter (for water quality). Water can be released at the dam in the case of heavy rains or to let water out to maintain winter levels. The U.S. Geological Survey (USGS) monitors water levels in Lake Murray. The histograms in Figure 11.3 were created using 67,810 hourly elevation levels from November 1, 2007, through August 11, 2015.

The two histograms of lake levels were made from the same data set, and the histograms look identical in shape. The shape of the distribution of lake levels is skewed left because the left side of the histogram is longer. The minimum lake level is 350 feet, and the maximum is 359 feet. Using the histogram on the right, by adding the height of the bar for 358 and 359 feet elevations, we see the lake level is at 358 or 359 roughly 40% of the time. Using this information, it appears that a lake level of 357 feet is the midpoint of the distribution.

Let's examine the difference in the two histograms. The histogram on the left puts the count of observations on the vertical axis (this is called a frequency histogram), while the histogram on the right uses the percentage of times the lake reaches a certain level (this is called a relative frequency histogram). The frequency histogram tells us the lake reached an elevation of 358 feet approximately 24,000 times (24,041, to be exact!). If a fisherman considering a move to Lake Murray cares about how often the lake reaches a certain level, it is more illustrative to use the relative frequency histogram on the right, which reports the percentage of times the lake reached 358 feet. The height of the bar for 358 feet is 35, so the fisherman would know the lake is at the 358 foot elevation roughly 35% of the time.

It is not uncommon in the current world to have very large data sets. Google uses big data to rank web pages and provide the best search results. Banks use big data to analyze spending patterns and learn when to flag your debit or credit card for fraudulent use. Large firms use big data to analyze market patterns and adapt marketing strategies accordingly. Our data set of size 67,810 is actually small in the realm of "big data" but is still big enough to see that it is almost always better to use a relative frequency histogram when sample sizes grow large. A relative

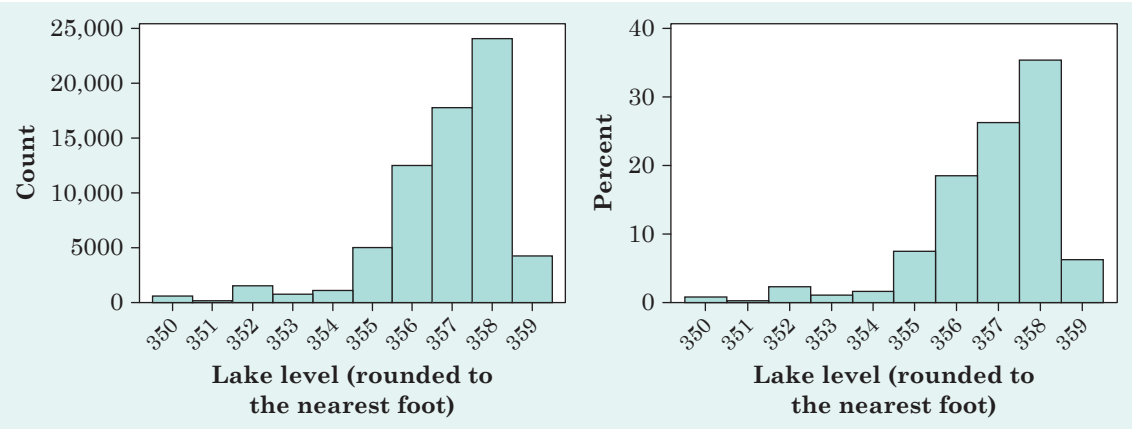

Figure 11.3 U.S. Geological Survey (USGS) reported hourly lake levels from November 1, 2007, through August 11, 2015, for Lake Murray, SC. There are 67,810 lake-level observations. The histogram on the left is showing the number of times the lake was at each level. The histogram on the right shows these same data as the percentage of times Lake Murray reached each level.

frequency histogram is also a better choice if one wants to make comparisons between two distributions.

EXAMPLE 5 Shakespeare's words

Figure 11.4 shows the distribution of lengths of words used in Shakespeare's plays. This distribution has a single peak and is somewhat skewed to the right. There are many short words (three and four letters) and few very long words (10, 11, or 12 letters), so that the right tail of the histogram extends out farther than the left tail. The center of the distribution is about 4. That is, about half of Shakespeare's words have four or fewer letters. The variability is from 1 letter to 12 letters.

Notice that the vertical scale in Figure 11.4 is not the *count* of words but the *percentage* of all of Shakespeare's words that have each length. A histogram of percentages rather than counts is convenient because this was a large data set. Different kinds of writing have different distributions of word lengths, but all are right-skewed because short words are common and very long words are rare.

The overall shape of a distribution is important information about a variable. Some types of data regularly produce distributions that are symmetric or skewed. For example, the sizes of living things of the same species

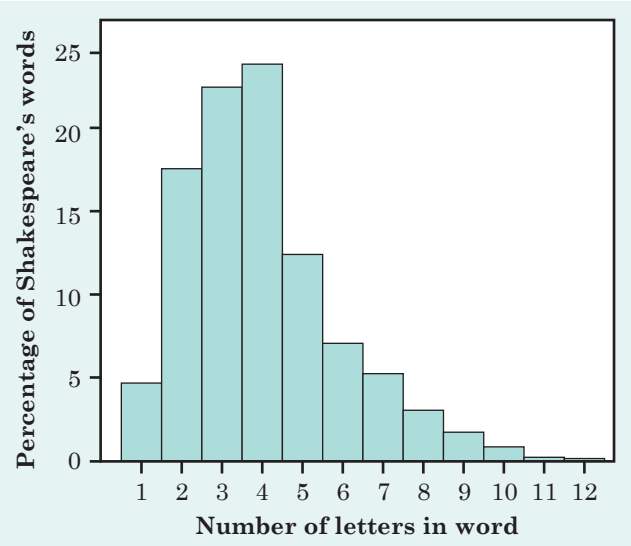

Figure 11.4 The distribution of word lengths used by Shakespeare in his plays, Example 5. This distribution is skewed to the right. (Data from C. B. Williams, *Style and Vocabulary: Numerical Studies*, Griffin, 1970.)

(like lengths of crickets) tend to be symmetric. Data on incomes (whether of individuals, companies, or nations) are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few very large incomes. It is very common for data to be skewed to the right when they have a strict minimum value (often 0). Income and the lengths of Shakespeare's words are examples. Likewise, data that have a strict maximum value (such as 100, as in student test scores) are often skewed to the left. Do remember that many distributions have shapes that are neither symmetric nor skewed. Some data show other patterns. Scores on an exam, for example, may have a cluster near the top of the scale if many students did well. Or they may show two distinct peaks if a tough problem divided the class into those who did and didn't solve it. Use your eyes and describe what you see.

NOW IT'S YOUR TURN

11.2 Height distribution. Height distributions generally have a predictable pattern. In a large introductory statistics class, students were asked to report their height. The histogram in Figure 11.5 displays the distribution of heights, in inches, for 266 females from this class. Describe the shape, center, and variability of this distribution. Are there any outliers?

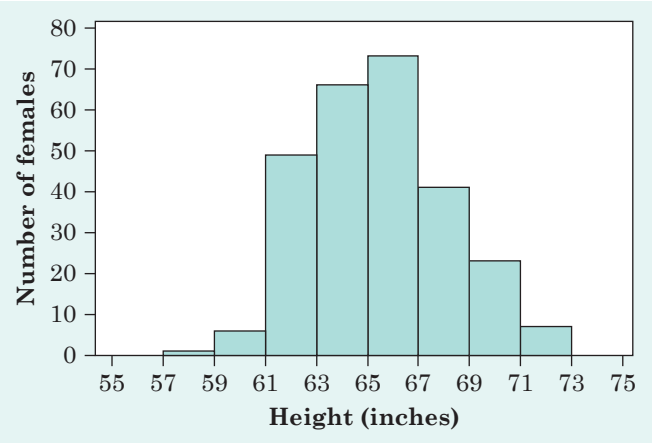

Figure 11.5 Heights in inches for 266 females in a large introductory statistics class.

Stemplots

Histograms are not the only graphical display of distributions. For small data sets, a *stemplot* (sometimes called a *stem-and-leaf plot*) is quicker to make and presents more detailed information.

Stemplot

To make a **stemplot:**

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit. Do not include commas or decimal points with your leaves.

2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

EXAMPLE 6 Stemplot of the "65 and over" data

For the "65 and over" percentages in Table 11.1, the whole-number part of the observation is the stem, and the final digit (tenths) is the leaf. The Alabama entry, 13.8, has stem 13 and leaf 8. Stems can have as many digits as needed, but each leaf must consist of only a single digit. Figure 11.6 shows the steps in making a stemplot for the data in Table 11.1. First, write the stems. Then go through the table adding a leaf for each observation. Finally, arrange the leaves in increasing order out from each stem.

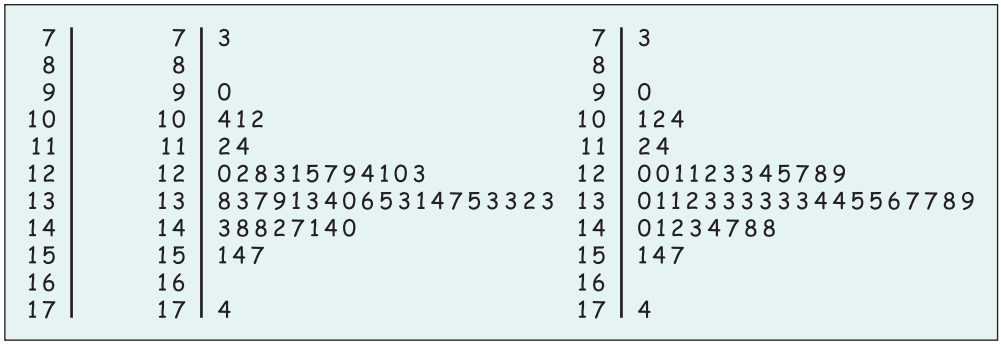

Figure 11.6 Making a stemplot of the data in Table 11.1. Whole percents form the stems, and tenths of a percent form the leaves.

A stemplot looks like a histogram turned on end. The stemplot in Figure 11.6 is just like the histogram in Figure 11.1 because the classes chosen for the histogram are the same as the stems in the stemplot. Figure 11.7 is a stemplot of the Illinois tuition data discussed in Example 3. This stemplot has almost four times as many classes as the histogram of the same data in Figure 11.2. We interpret stemplots as we do histograms, looking for the overall pattern and for any outliers.

You can choose the classes in a histogram. The classes (the stems) of a stemplot are given to you. You can get more flexibility by **rounding** the data so that the final digit after rounding is suitable as a leaf. Do this when the data have too many digits. For example, the tuition charges of Illinois colleges look like

$$\$9{,}500 \quad \$9{,}430 \quad \$7{,}092 \quad \$10{,}672 \quad \ldots$$

A stemplot would have too many stems if we took all but the final digit as the stem and the final digit as the leaf. To make the stemplot in Figure 11.7, we round these data to the nearest hundred dollars:

$$95 \quad 94 \quad 71 \quad 107 \quad \ldots$$

These values appear in Figure 11.7 on the stems 9, 9, 7, and 10.

The chief advantage of a stemplot is that it displays the actual values of the observations. We can see from the stemplot in Figure 11.7, but not from the histogram in Figure 11.2, that Illinois's most expensive college charges $38,600 (rounded to the nearest hundred dollars). Stemplots are also faster to draw than histograms. A stemplot requires that we use the first digit or digits as stems. This amounts to an automatic choice of classes and can give a poor picture of the distribution. Stemplots do not work well with large data sets because the stems then have too many leaves.

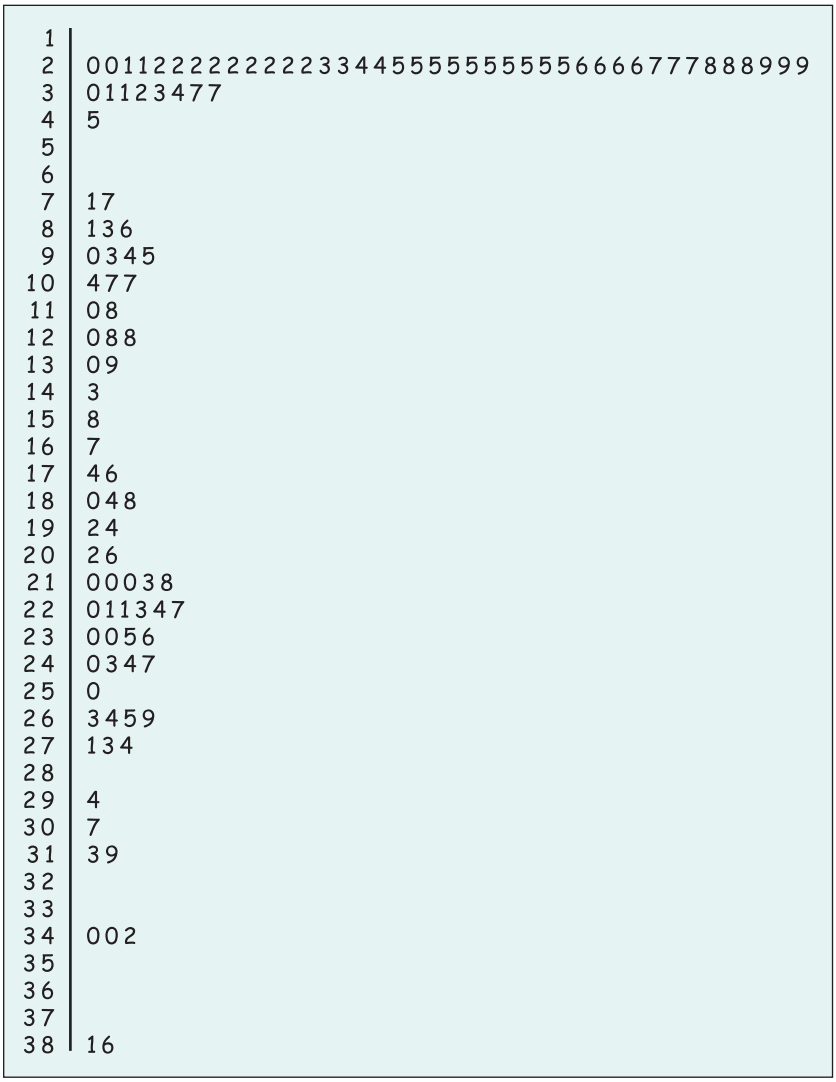

Figure 11.7 Stemplot of the Illinois tuition and fee data. (Data from www.isac.org.)

STATISTICS IN SUMMARY

Chapter Specifics

- The **distribution** of a variable tells us what values the variable takes and how often it takes each value.
- To display the distribution of a quantitative variable, use a **histogram** or a **stemplot**. We usually favor stemplots when we have a small

number of observations and histograms for larger data sets. Make sure to choose the appropriate number of classes so that the distribution shape is displayed accurately. For really large data sets, use a histogram of percents (relative frequency histogram).

- When you look at a graph, look for an **overall pattern** and for **deviations** from that pattern, such as **outliers**.
- We can characterize the overall pattern of a histogram or stemplot by describing its **shape**, **center**, and **variability**. Some distributions have simple shapes, such as **symmetric**, **skewed left**, or **skewed right**, but others are too irregular to describe by a simple shape.

In Chapter 10, we learned how to use tables and graphs to see what data tell us. In this chapter, we looked at two additional graphs—histograms and stemplots—that help us make sense of large collections of numbers. These graphics are pictures of the distribution of a single quantitative variable. Although the bar graphs in Chapter 10 look much like histograms, the difference is that bar graphs are used to display the distribution of a categorical variable, while histograms display the distribution of a quantitative variable. The overall pattern (shape, center, and variability) and deviations from this pattern (outliers) are important features of the distribution of a variable. We will look more carefully at these features in future chapters. In addition, these features will figure prominently in some of the conclusions that we will draw about a variable from data.

CASE STUDY EVALUATED Figure 11.8 is a histogram of the distribution of the amount of dietary fiber in 77 brands of cereal found on the shelves of a grocery store. Use what you have learned in this chapter to answer the following questions.

1. Describe the overall shape, the center, and the variability of the distribution.
2. Are there any outliers?
3. Wheaties has 3 grams of dietary fiber. Is this a low, high, or typical amount?

LaunchPad macmillan learning **Online Resources**

- The StatBoards ~~Video~~ video, *Creating and Interpreting a Histogram*, describes how to constr~~~~ video istogram.
- The Snapshots ~~Video~~, *Visualizing Quantitative Data*, discusses stemplots and histograms in the setting of two real examples.

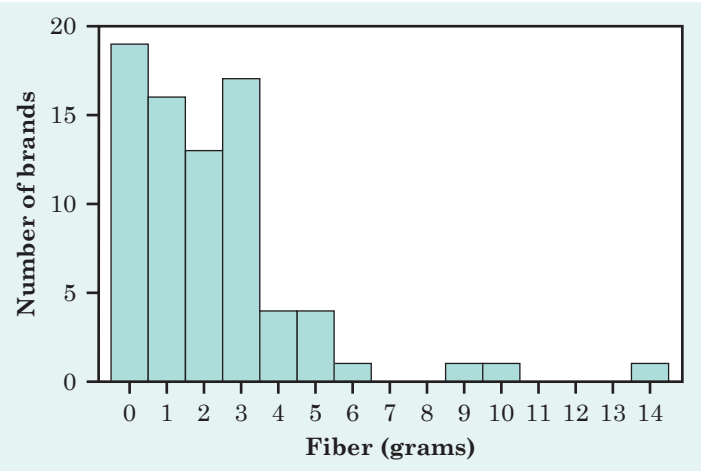

Figure 11.8 Histogram of the amount of dietary fiber (in grams) in 77 different brands of cereal.

CHECK THE BASICS

For Exercise 11.1, see page 246; for Exercise 11.2, see page 252.

11.3 Histograms. Use a histogram when

(a) the number of observations is small.

(b) you want to look at the distribution of a quantitative variable.

(c) you want to look at the distribution of a categorical variable.

(d) you want to show the actual observations.

11.4 Histograms. The heights of the bars on a relative frequency histogram displaying the lengths of rivers will add to

(a) 100%.

(b) the sample size.

(c) the sum of all the river lengths.

(d) the midpoint of the distribution.

11.5 Stemplots. An advantage of a stemplot over a histogram is

(a) they are good for really large data sets.

(b) they are horizontal.

(c) one can recover the actual observations from the display.

(d) the classes are chosen for you.

11.6 Shape of distributions. Figure 11.9 contains exam scores for 500 students. What is the shape of the exam score distribution?

(a) symmetric

(b) skewed right

(c) skewed left

(d) none of the above

11.7 Shape of distributions. In a certain town, most haircuts are between \$10 and \$20, but a few salons cater to high-end clients and charge \$30 to \$60. The distribution of haircut prices is

(a) symmetric.

(b) skewed right.

(c) skewed left.

(d) none of the above.

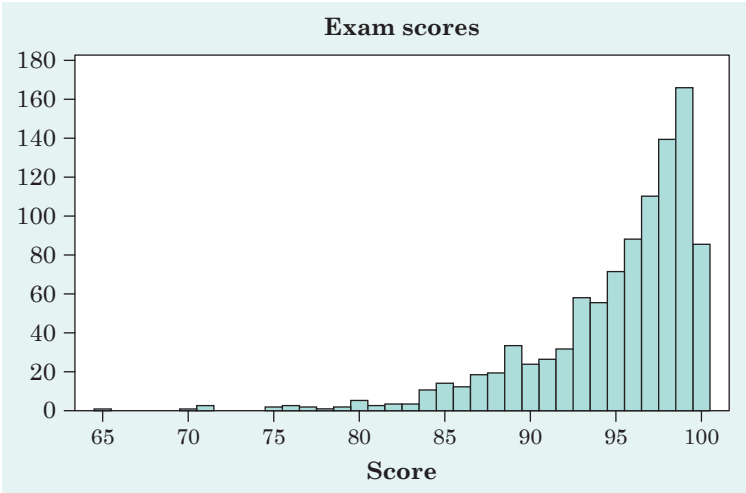

Figure 11.9 Histogram of the exam scores for 500 students.

CHAPTER 11 EXERCISES

11.8 Lightning storms. Figure 11.10 comes from a study of lightning storms in Colorado. It shows the distribution of the hour of the day during which the first lightning flash for that day occurred. Describe the shape, center, and variability of this distribution. Are there any outliers?

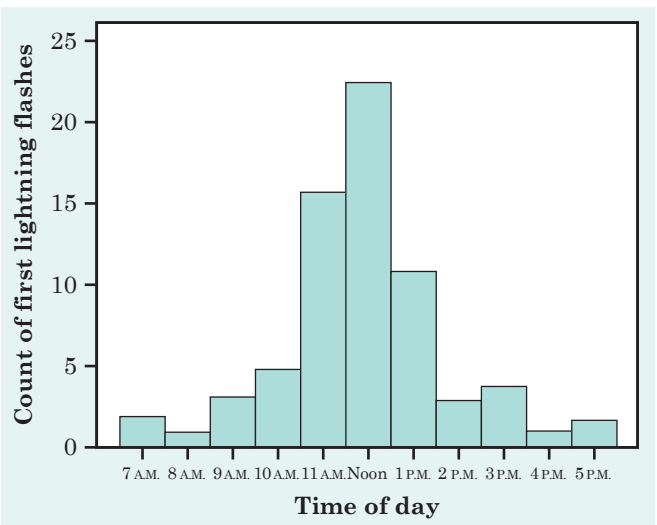

Figure 11.10 Histogram of the time of day at which the day's first lightning flash occurred (from a study in Colorado), Exercise 11.8. (Data from an episode in the Annenberg/Corporation for Public Broadcasting telecourse *Against All Odds: Inside Statistics*.)

```
19 | 9
20 | 59
21 | 13444
22 | 122255666777899
23 | 00011345556666777888
24 | 01588
25 | 69
26 |
27 |
28 | 4
```

Figure 11.11 Stemplot of the percentage of each state's residents who are 18 to 34 years old, Exercise 11.9.

11.9 Where do 18- to 34-year-olds live? Figure 11.11 is a stemplot of the percentage of residents aged 18 to 34 in each of the 50 states in July 2008. As in Figure 11.6 (page 254) for older residents, the stems are whole percents and the leaves are tenths of a percent.

(a) Utah has the largest percentage of young adults. What is the percentage for this state?

(b) Ignoring Utah, describe the shape, center, and variability of this distribution.

(c) Is the distribution for young adults more or less variable than the distribution in Figure 11.6 for older adults?

11.10 Minority students in engineering. Figure 11.12 is a histogram of the number of minority students (black, Hispanic, Native American) who earned doctorate degrees in engineering from each of 152 universities in the years 2000 through 2002. Briefly describe the shape, center, and variability of this distribution. The classes for Figure 11.12 are 1–5, 6–10, and so on.

11.11 Returns on common stocks. The total return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percentage of the beginning price. Figure 11.13 is a histogram of the distribution of total returns for

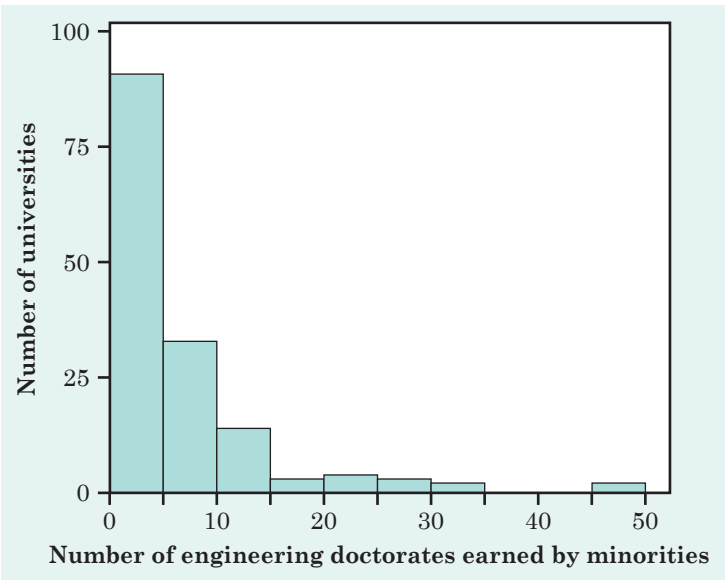

Figure 11.12 The distribution of the number of engineering doctorates earned by minority students at 152 universities, 2000–2002, Exercise 11.10. (Data from the 2003 National Science Foundation Survey of Earned Doctorates, http://webcaspar.nsf.gov/.)

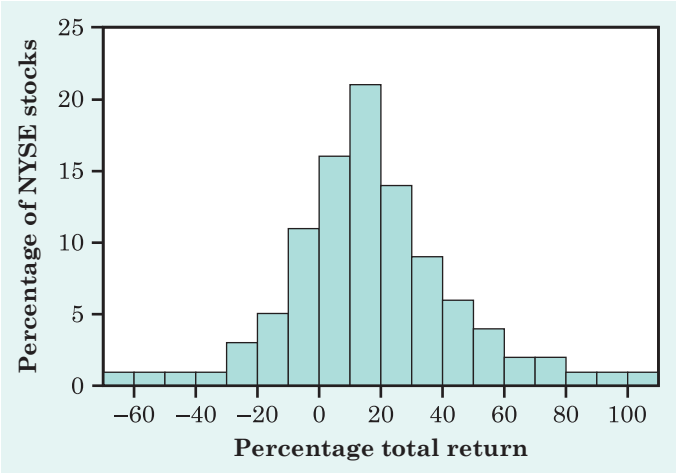

Figure 11.13 The distribution of total returns for all New York Stock Exchange common stocks in one year, Exercise 11.11. (Data from J. K. Ford, "Diversification: How Many Stocks Will Suffice?" *American Association of Individual Investors Journal*, January 1990, pp. 14–16.)

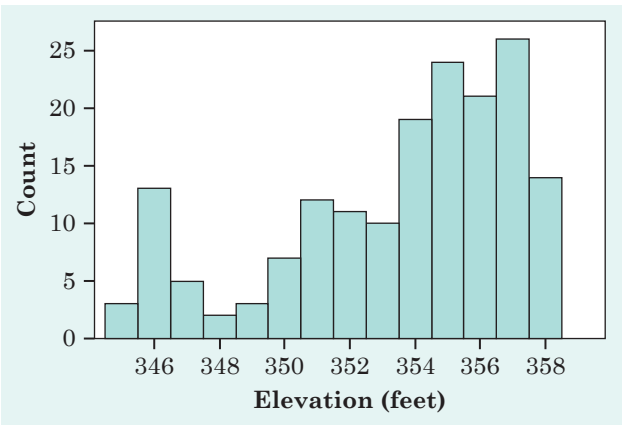

Figure 11.14 U.S. Geological Survey (USGS) data on average monthly Lake Murray elevations from November 1990 to July 2006, Exercise 11.12.

all 1528 common stocks listed on the New York Stock Exchange in one year.

(a) Describe the overall shape of the distribution of total returns.

(b) What is the approximate center of this distribution? Approximately what were the smallest and largest total returns? (This describes the variability of the distribution.)

(c) A return less than zero means that owners of the stock lost money. About what percentage of all stocks lost money?

(d) Explain why we prefer a histogram to a stemplot for describing the returns on 1528 common stocks.

11.12 More lake levels. Figure 11.14 contains average monthly lake levels

TABLE 11.2 Combined city/highway gas mileage for model year 2015 midsize cars

Model	mpg	Model	mpg
Acura RLX	24	Honda Accord Hybrid	47
Audi A8	22	Hyundai Sonata	37
BMW 528i	27	Infiniti Q40	22
Bentley Mulsanne	13	Jaguar XF	18
Buick LaCrosse	21	Kia Forte	30
Buick Regal	22	Lexus ES 350	24
Cadillac CTS	14	Lincoln MKZ	25
Cadillac CTS AWD	22	Mazda 6	30
Chevrolet Malibu	29	Mercedes-Benz B-Class	84
Chevrolet Sonic	31	Mercedez-Benz E350	23
Dodge Challenger	18	Nissan Altima	31
Ford C-max Energi Plugin hybrid	88	Nissan Leaf	114
Ford Fusion Energi Plug-in Hybrid	88	Rolls Royce Wraith	15
Ford Fusion Hybrid	42	Toyota Prius	50
Honda Accord	31	Toyota Prius Plug-in Hybrid	95
		Volvo S80	22

Source: www.fueleconomy.gov/feg/byclass/Midsize_Cars2015.shtml.

for Lake Murray in South Carolina. Describe the shape, center, and variability of the distribution of lake levels.

11.13 Automobile fuel economy. Government regulations require automakers to give the city and highway gas mileages for each model of car. Table 11.2 gives the combined highway and city mileages (miles per gallon) for 31 model year 2015 sedans. Make a stemplot of the combined gas mileages of these cars. What can you say about the overall shape of the distribution? Where is the center (the value such that half the cars have better gas mileage and half have worse gas mileage)? Some of these cars are electric and the reported milage is the electric equivalent. These cars have far higher mileage. How many electric cars are in this data set?

11.14 The obesity epidemic. Medical authorities describe the spread of obesity in the United States as an epidemic. Table 11.3 gives the percentage of adults who were obese in each of the 50 states in 2009. Display the distribution in a graph and briefly describe its shape, center, and variability.

11.15 Yankee money. Table 11.4 gives the salaries of the players on the New York Yankees baseball team for the 2015 season. Make a histogram of these data. Is the distribution

TABLE 11.3 Percentage of adult population who are obese, 2009

State	Percent	State	Percent	State	Percent
Alabama	31.0	Louisiana	33.0	Ohio	28.8
Alaska	24.8	Maine	25.8	Oklahoma	31.4
Arizona	25.5	Maryland	26.2	Oregon	23.0
Arkansas	30.5	Massachusetts	21.4	Pennsylvania	27.4
California	24.8	Michigan	29.6	Rhode Island	24.6
Colorado	18.6	Minnesota	24.6	South Carolina	29.4
Connecticut	20.6	Mississippi	34.4	South Dakota	29.6
Delaware	27.0	Missouri	30.0	Tennessee	32.3
Florida	25.2	Montana	23.2	Texas	28.7
Georgia	27.2	Nebraska	27.2	Utah	23.5
Hawaii	22.3	Nevada	25.8	Vermont	22.8
Idaho	24.5	New Hampshire	25.7	Virginia	25.0
Illinois	26.5	New Jersey	23.3	Washington	26.4
Indiana	29.5	New Mexico	25.1	West Virginia	31.1
Iowa	27.9	New York	24.2	Wisconsin	28.7
Kansas	28.1	North Carolina	29.3	Wyoming	24.6
Kentucky	31.5	North Dakota	27.9		

Source: National Centers for Disease Control and Prevention, www.cdc.gov/obesity/data/databases.html.

TABLE 11.4 Salaries of the New York Yankees, 2015

Player	Salary ($)	Player	Salary ($)
CC Sabathia	24,285,714	Nathan Eovaldi	3,300,000
Mark Teixeira	23,125,000	Ivan Nova	3,300,000
Masahiro Tanaka	22,000,000	Dustin Ackley	2,600,000
Alex Rodriguez	22,000,000	Chris Young	2,500,000
Jacoby Ellsbury	21,142,857	Brendan Ryan	2,000,000
Brian McCann	17,000,000	Adam Warren	572,600
Carlos Beltran	15,000,000	Justin Wilson	556,000
Chase Headley	13,000,000	Didi Gregorius	553,900
Brett Gardner	12,500,000	John Ryan Murphy	518,700
Andrew Miller	9,000,000	Chasen Shreve	510,275
Garrett Jones	5,000,000	Dellin Betances	507,500
Stephen Drew	5,000,000		

Source: http://espn.go.com/mlb/team/salaries/_/name/nyy/new-york-yankees.

roughly symmetric, skewed to the right, or skewed to the left? Explain.

11.16 The statistics of writing style. Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data collected by students on the percentages of words of 1 to 15 letters used in articles in *Popular Science* magazine:

Length:	1	2	3	4	5
Percent:	3.6	14.8	18.7	16.0	12.5

Length:	6	7	8	9	10
Percent:	8.2	8.1	5.9	4.4	3.6

Length:	11	12	13	14	15
Percent:	2.1	0.9	0.6	0.4	0.2

(a) Make a histogram of this distribution. Describe its shape, center, and variability.

(b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution in Figure 11.4 (page 252) for Shakespeare's plays? Look in particular at short words (two, three, and four letters) and very long words (more than 10 letters).

11.17 What's my shape? Do you expect the distribution of the total player payroll for each of the 30 teams in Major League Baseball to be roughly symmetric, clearly skewed to the right, or clearly skewed to the left? Why?

11.18 Asians in the eastern states. Here are the percentages of the population who are of Asian origin in each state east of the Mississippi River in 2008:

State	Percent	State	Percent
Alabama	1.0	New Hampshire	1.9
Connecticut	3.6	New Jersey	7.6
Delaware	2.9	New York	7.0
Florida	2.3	North Carolina	1.9
Georgia	2.9	Ohio	1.6
Illinois	4.3	Pennsylvania	2.4
Indiana	1.4	Rhode Island	2.8
Kentucky	1.0	South Carolina	1.2
Maine	0.9	Tennessee	1.3
Maryland	5.1	Vermont	1.1
Massachusetts	4.9	Virginia	4.9
Michigan	2.4	West Virginia	0.7
Mississippi	0.8	Wisconsin	2.0

Make a stemplot of these data. Describe the overall pattern of the distribution. Are there any outliers?

11.19 How many calories does a hot dog have? *Consumer Reports* magazine presented the following data on the number of calories in a hot dog for each of 17 brands of meat hot dogs:

173	191	182	190	172	147
146	139	175	136	179	153
107	195	135	140	138	

Make a stemplot of the distribution of calories in meat hot dogs and briefly describe the shape of the distribution. Most brands of meat hot dogs contain a mixture of beef and pork, with up to 15% poultry allowed by government regulations. The only brand with a different makeup was Eat Slim Veal Hot Dogs. Which point on your stemplot do you think represents this brand?

TABLE 11.5 Age distribution in the United States, 1950 and 2050 (in millions of persons)

Age group	1950	2050
Under 10 years	29.3	56.2
10 to 19 years	21.8	56.7
20 to 29 years	24.0	56.2
30 to 39 years	22.8	55.9
40 to 49 years	19.3	52.8
50 to 59 years	15.5	49.1
60 to 69 years	11.0	45.0
70 to 79 years	5.5	34.5
80 to 89 years	1.6	23.7
90 to 99 years	0.1	8.1
100 years and over	—	0.6
Total	151.1	310.6

Source: *Statistical Abstract of the United States,* www.census.gov/library/publications/2009/compendia/statab/129ed.html; and Census Bureau, www.census.gov/population/projections/.

11.20 The changing age distribution of the United States. The distribution of the ages of a nation's population has a strong influence on economic and social conditions. Table 11.5 shows the age distribution of U.S. residents in 1950 and 2050, in millions of persons. The 1950 data come from that year's census. The 2050 data are projections made by the Census Bureau.

(a) Because the total population in 2050 is much larger than the 1950 population, comparing percentages in each age group is clearer than comparing counts. Make a table of the percentage of the total population in each age group for both 1950 and 2050.

(b) Make a histogram of the 1950 age distribution (in percents). Then describe the main features of the distribution. In particular, look at the percentage of children relative to the rest of the population.

(c) Make a histogram of the projected age distribution for the year 2050. Use the same scales as in part (b) for easy comparison. What are the most important changes in the U.S. age distribution projected for the 100-year period between 1950 and 2050?

11.21 Babe Ruth's home runs. Here are the numbers of home runs that Babe Ruth hit in his 15 years with the New York Yankees, 1920 to 1934:

54 59 35 41 46 25 47 60
54 46 49 46 41 34 22

Make a stemplot of these data. Is the distribution roughly symmetric, clearly skewed, or neither? About how many home runs did Ruth hit in

a typical year? Is his famous 60 home runs in 1927 an outlier?

11.22 Back-to-back stemplot. The current major league single-season home run record is held by Barry Bonds of the San Francisco Giants. Here are Bonds's home run counts for 1986 to 2007:

16	25	24	19	33	25
34	46	37	33	42	40
37	34	49	73	46	45
45	5	26	28		

A **back-to-back stemplot** helps us compare two distributions. Write the stems as usual, but with a vertical line both to their left and to their right. On the right, put leaves for Ruth (Exercise 11.21). On the left, put leaves for Bonds. Arrange the leaves on each stem in increasing order out from the stem. Now write a brief comparison of Ruth and Bonds as home run hitters.

11.23 When it rains, it pours. On July 25 to 26, 1979, 42.00 inches of rain fell on Alvin, Texas. That's the most rain ever recorded in Texas for a 24-hour period. Table 11.6 gives the maximum precipitation ever recorded in 24 hours (through 2010) at any weather station in each state. The record amount varies a great deal from state to state—hurricanes bring extreme rains on the Atlantic

TABLE 11.6 Record 24-hour precipitation amounts (inches) by state

State	Precip.	State	Precip.	State	Precip.
Alabama	32.52	Louisiana	22.00	Ohio	10.75
Alaska	15.20	Maine	13.32	Oklahoma	15.68
Arizona	11.40	Maryland	14.75	Oregon	11.77
Arkansas	14.06	Massachusetts	18.15	Pennsylvania	13.50
California	25.83	Michigan	9.78	Rhode Island	12.13
Colorado	11.08	Minnesota	15.10	South Carolina	14.80
Connecticut	12.77	Mississippi	15.68	South Dakota	8.74
Delaware	8.50	Missouri	18.18	Tennessee	13.60
Florida	23.28	Montana	11.50	Texas	42.00
Georgia	21.10	Nebraska	13.15	Utah	5.08
Hawaii	38.00	Nevada	7.78	Vermont	9.92
Idaho	7.17	New Hampshire	11.07	Virginia	14.28
Illinois	16.91	New Jersey	14.81	Washington	14.26
Indiana	10.50	New Mexico	11.28	West Virginia	12.02
Iowa	13.18	New York	11.15	Wisconsin	11.72
Kansas	13.53	North Carolina	22.22	Wyoming	6.06
Kentucky	10.40	North Dakota	8.10		

Source: National Oceanic and Atmospheric Administration, www.noaa.gov.

coast, and the mountain West is generally dry. Make a graph to display the distribution of records for the states. Mark where your state lies in this distribution. Briefly describe the distribution.

EXPLORING THE WEB

Follow the QR code to access exercises.