

1

Picturing Distributions with Graphs



Photograph by the U.S. Census Bureau, Public Information Office (PIO)

MA 101 Ch1

In this chapter, we cover...

- 1.1 Individuals and variables
- 1.2 Categorical variables: Pie charts and bar graphs
- 1.3 Quantitative variables: Histograms
- 1.4 Interpreting histograms
- 1.5 Quantitative variables: Stemplots
- 1.6 Time plots

Statistics is the science of data. The volume of data available to us is overwhelming. For example, the U.S. Census Bureau's American Community Survey collects data from about 3,000,000 housing units each year. Astronomers work with data on tens of millions of galaxies. The checkout scanners at Walmart's more than 11,000 stores in 27 countries record hundreds of millions of transactions every week, all saved to inform both Walmart and its suppliers. The first step in dealing with such a flood of data is to organize our thinking about data. Fortunately, we can do this without looking at millions of data points.

1.1 Individuals and Variables

Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

Individuals and Variables

Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.



Data!

The documentary *Particle Fever* recreates the excitement of the Large Hadron Collider (LHC) experiment. The LHC is a 17-mile tunnel, designed to accelerate a proton to close to the speed of light and then have the protons collide to help physicists understand how the universe works. When the first collisions are recorded live in the film, American physicist Monica Dunford exclaims, "We have data. It's unbelievable how fantastic this data is."

A college's student database, for example, includes data about every currently enrolled student. The students are the individuals described by the data set. For each individual, the data contain the values of variables such as date of birth, choice of major, and grade point average (GPA). In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

1. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?
2. **What?** How many **variables** do the data contain? What are the **exact definitions** of these variables? In what **unit of measurement** is each variable recorded? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms.
3. **Where?** Student GPAs and SAT scores (or lack of them) will vary from college to college depending on many variables, including admissions "selectivity" for the college.
4. **When?** Students change from year to year, as do prices, salaries, and so forth.
5. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want answers for just these individuals or for some larger group that these individuals are supposed to represent? Are the individuals and variables suitable for the intended purpose?

Some variables, such as a person's sex or college major, simply place individuals into categories. Others, like height and GPA, take numerical values for which we can do arithmetic. It makes sense to give an average income for a company's employees, but it does not make sense to give an "average" sex. We can, however, count the numbers of female and male employees and do arithmetic with these counts.

Categorical and Quantitative Variables

A **categorical variable** places an individual into one of several groups or categories.

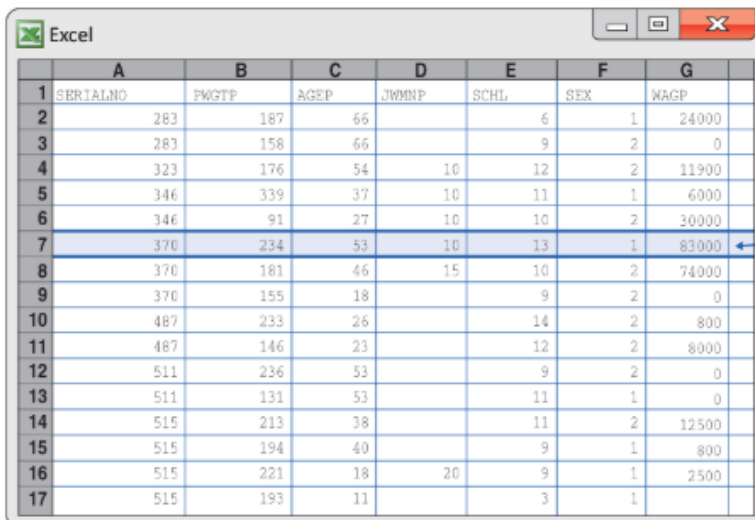
A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense. The values of a quantitative variable are usually recorded with a **unit of measurement** such as seconds or kilograms.

EXAMPLE 1.1 The American Community Survey

At the U.S. Census Bureau website, you can view the detailed data collected by the American Community Survey, though of course the identities of people and housing units are protected. If you choose the file of data on people, the *individuals* are the people living in the housing units contacted by the survey. More than 100 variables are recorded for each individual. Figure 1.1 displays a very small part of the data.

Each row records data on one individual. Each column contains the values of one *variable* for all the individuals. Translated from the U.S. Census Bureau's abbreviations, the variables are

SERIALNO	An identifying number for the household.
PWGTP	Weight in pounds.
AGEP	Age in years.
JWMNP	Travel time to work in minutes.



	A	B	C	D	E	F	G
1	SERIALNO	PWGTP	AGEP	JWMNP	SCHL	SEX	WAGP
2	283	187	66		6	1	24000
3	283	158	66		9	2	0
4	323	176	54	10	12	2	11900
5	346	339	37	10	11	1	6000
6	346	91	27	10	10	2	30000
7	370	234	53	10	13	1	83000
8	370	181	46	15	10	2	74000
9	370	155	18		9	2	0
10	487	233	26		14	2	800
11	487	146	23		12	2	8000
12	511	236	53		9	2	0
13	511	131	53		11	1	0
14	515	213	38		11	2	12500
15	515	194	40		9	1	800
16	515	221	18	20	9	1	2500
17	515	193	11		3	1	

Each row in the spreadsheet contains data on one individual.

Figure 1.1
Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.1

A spreadsheet displaying data from the American Community Survey, for [Example 1.1](#).

SCHL Highest level of education. The numbers designate categories, *not* specific grades. For example, 9 = high school graduate, 10 = some college but no degree, and 13 = bachelor's degree.

SEX Sex, designated by 1 = male and 2 = female.

WAGE Wage and salary income last year, in dollars.

Look at the highlighted row in [Figure 1.1](#). This individual is a 53-year-old man who weighs 234 pounds, travels 10 minutes to work, has a bachelor's degree, and earned \$83,000 last year.

In addition to the household serial number, there are six variables. Education and sex are categorical variables. The values for education and sex are stored as numbers, but these numbers are just labels for the categories and have no units of measurement. The other four variables are quantitative. Their values do have units. These variables are weight in pounds, age in years, travel time in minutes, and income in dollars.

The *purpose* of the American Community Survey is to collect data that represent the entire nation to guide government policy and business decisions. To do this, the households contacted are chosen at random from all households in the country. We will see in Chapter 8 why choosing at random is a good idea.

Most data tables follow this format—each row is an individual, and each column is a variable. The data set in [Figure 1.1](#) appears in a **spreadsheet** program that has rows and columns ready for your use. Spreadsheets are commonly used to enter and transmit data and to do simple calculations.

Macmillan Learning Online Resources

- The Snapshots video, *Data and Distributions*, provides a nice introduction to the ideas of this section.
- The StatClips Examples video, *Basic Principles of Exploring Data: Example A*, describes the variables collected in the American Community Survey from [Example 1.1](#).

APPLY YOUR KNOWLEDGE

- 1.1 Fuel Economy.** Here is a small part of a data set that describes the fuel economy in miles per gallon (mpg) of model year 2016 motor vehicles:

Make and Model	Vehicle Class	Transmission Type	Number of Cylinders	City mpg	Highway mpg	Annual Fuel Cost
:						
Subaru Impreza	Compact	Manual	4	25	34	\$900
Nissan Juke	Small station wagon	Manual	4	28	34	\$1,100
Hyundai Elantra GT	Midsize	Automatic	4	24	33	\$950
Chevrolet Impala	Large	Automatic	6	19	29	\$1,150
:						

The annual fuel cost is an estimate assuming 15,000 miles of travel a year (55% city and 45% highway) and an average fuel price.

- (a) What are the individuals in this data set?
- (b) For each individual, what variables are given? Which of these variables are categorical, and which are quantitative? In what units are the quantitative variables measured?
- 1.2 Students and Exercise.** You are preparing to study the exercise habits of college students. Describe two categorical variables and two quantitative variables that you might measure for each student. Give the units of measurement for the quantitative variables.

1.2 Categorical Variables: Pie Charts and Bar Graphs

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called **exploratory data analysis**. Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two principles that help us organize our exploration of a set of data.

Exploring Data

1. Begin by examining each variable by itself. Then move on to study the relationships among the variables.
2. Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. Chapters 1 through 3 present methods for describing a single variable. We study relationships among several variables in Chapters 4 through 6. In each case, we begin with graphical displays, then add numerical summaries for more complete description.

The proper choice of graph depends on the nature of the variable. To examine a single variable, we usually want to display its *distribution*.

Distribution of a Variable

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

The values of a categorical variable are labels for the categories. The **distribution of a categorical variable** lists the categories and gives either the count or the percent of individuals who fall in each category.

EXAMPLE 1.2 Which Major?



MAJORS

Approximately 1.5 million full-time, first-year students enrolled in colleges and universities in 2015. What do they plan to study? Here are data on the percents of first-year students who plan to major in several discipline areas:¹

Field of Study	Percent of Students
Arts and humanities	10.1
Biological sciences	14.9
Business	13.4
Education	4.2
Engineering	13.1
Health professions	11.3
Math and computer science	5.4
Physical sciences	2.7
Social sciences	10.8
Other majors and undeclared	13.9
Total	99.8

It's a good idea to check data for consistency. The percents should add to 100%. In fact, they add to 99.8%. What happened? Each percent is rounded to the nearest tenth. The exact percents would add to 100, but the rounded percents only come close. This is **roundoff error**. Roundoff errors don't point to mistakes in our work, just to the effect of rounding off results.

Columns of numbers take time to read. You can use a pie chart or a bar graph to display the distribution of a categorical variable more vividly. Figures 1.2 and 1.3 illustrate these displays for the distribution of intended college majors.

Pie charts show the distribution of a categorical variable as a “pie” whose slices are sized by the counts or percents for the categories. Pie charts are awkward to make by hand, but software will do the job for you. *A pie chart must include all the categories that make up a whole. Use a pie chart only when you want to emphasize each category’s relation to the whole.* We need the “Other majors and undeclared” category in Example 1.2 to complete the whole (all intended majors) and allow us to make the pie chart in Figure 1.2.



Bar graphs represent each category as a bar. The bar heights show the category counts or percents. Bar graphs are easier to make than pie charts and also easier to read. Figure 1.3 displays two bar graphs of the data on intended majors. The first orders the bars alphabetically by field of study (with “Other” at the end). It is often better to arrange the bars in order of height, as in Figure 1.3(b). This helps us immediately see which majors appear most often.

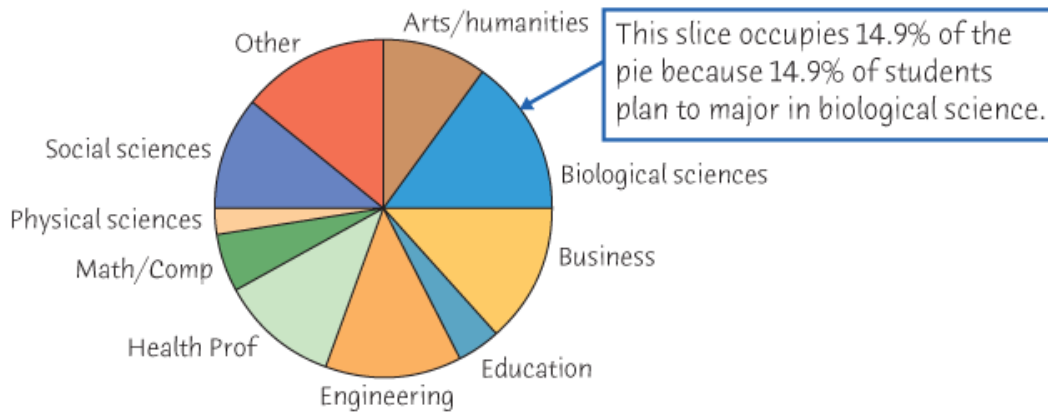


Figure 1.2

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.2

You can use a pie chart to display the distribution of a categorical variable. This pie chart, created with the Minitab 17 software package, shows the distribution of intended majors of students entering college. Statistical analysis relies heavily on statistical software, and Minitab is one of the most popular software choices both in industry and in colleges and schools of business. Computer output from other statistical packages like JMP, SPSS, and R is similar, so you can feel comfortable using any one of these packages.

Bar graphs are more flexible than pie charts. Both graphs can display the distribution of a categorical variable, but a bar graph can also compare any set of quantities that are measured in the same units.

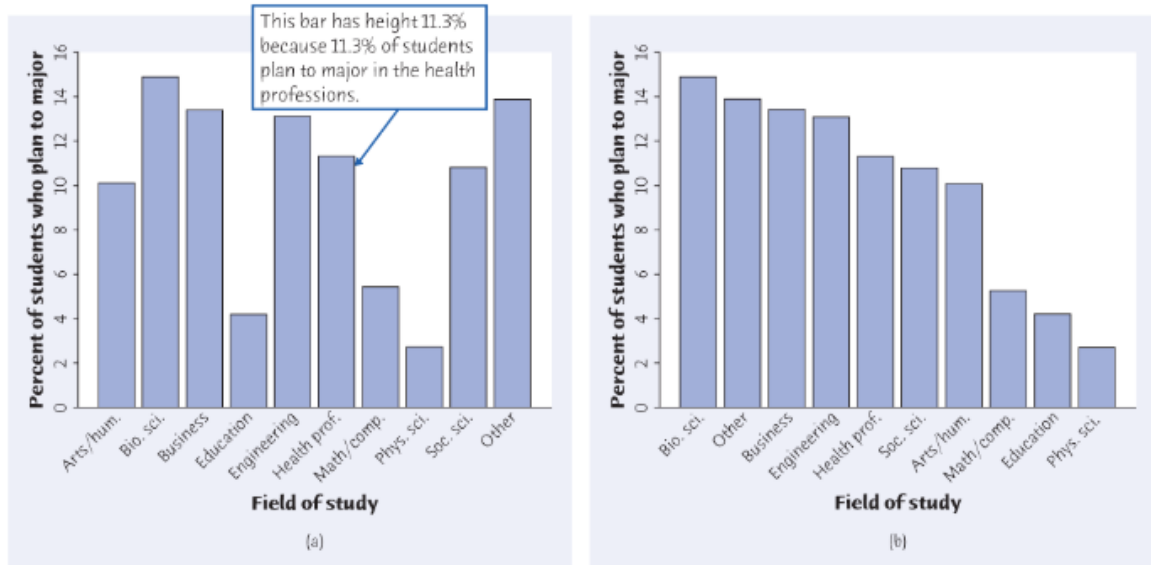


Figure 1.3
Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.3

Bar graphs of the distribution of intended majors of students entering college. In part (a), the bars follow the alphabetical order of fields of study. In part (b), the same bars appear in order of height. These figures were created with the Minitab 17 software package.

EXAMPLE 1.3 How Do 12–24s Learn about New Music?



MUSIC

What sources do Americans aged 12–24 years use to keep up-to-date and learn about music? Among those saying it was important to keep up with music, Edison Research asked which of several sources they had ever used. Here are the percents that have used each source.²

Source	Percent of 12–24s Who Have Used Each Source
AM/FM radio	57
Friends/family	77
YouTube	83
Music television channels	43
Facebook	49
Pandora	70
Apple iTunes	41
Information at local stores	37
SiriusXM Satellite Radio	24
Music blogs	23
iHeartRadio	27
Spotify	37

We can't make a pie chart to display these data. Each percent in the table refers to a different source, not to parts of a single whole. [Figure 1.4](#) is a bar graph comparing the 12 sources. We have again arranged the bars in order of height.

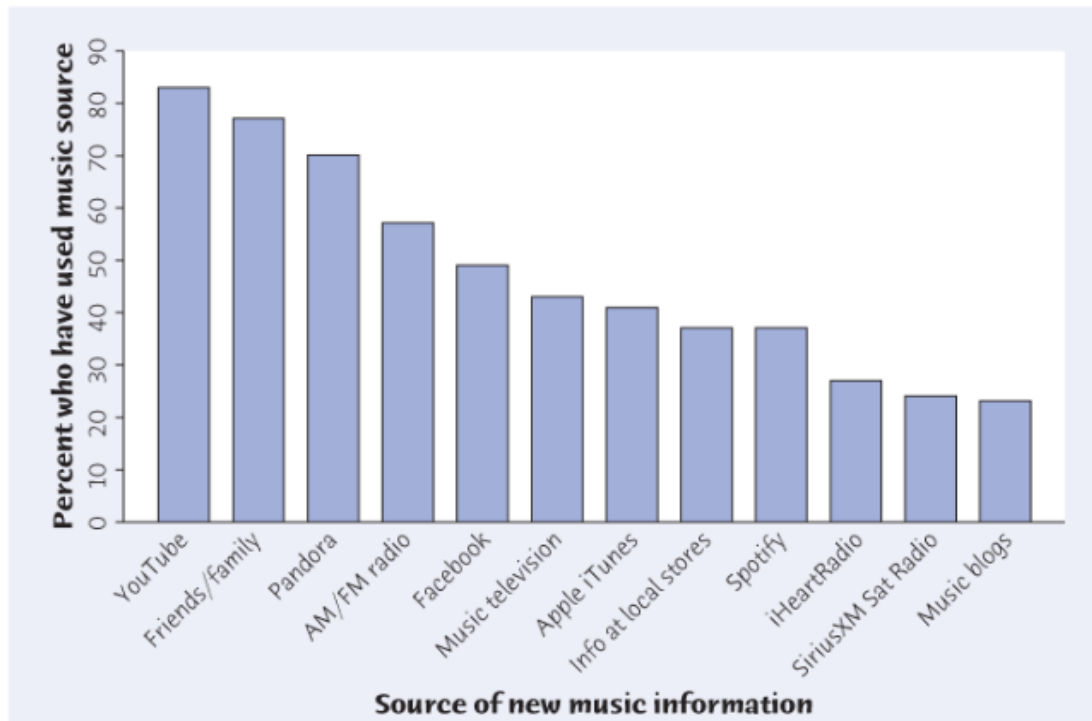


Figure 1.4

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.4

You can use a bar graph to compare quantities that are not part of a whole. This bar graph compares the percents of Americans aged 12–24 who have used each source to learn about new music, for [Example 1.3](#).

Bar graphs and pie charts are mainly tools for presenting data: they help your audience grasp data quickly. Because it is easy to understand data on a single categorical variable without a graph, bar graphs and pie charts are of limited use for data analysis. We will move on to quantitative variables, where graphs are essential tools.

Macmillan Learning Online Resources

- The Snapshots video, *Visualizing and Summarizing Categorical Data*, illustrates the ideas of both pie charts and bar graphs.
- The StatClips Examples video, *Summaries and Pictures for Categorical Data Example B*, provides the details of constructing pie charts and bar graphs through an example.

APPLY YOUR KNOWLEDGE


1.3 Social Media Preferences for Younger Audiences. Facebook remains the top choice of social media over all ages, with 65% using Facebook most often among those using social media sites. However, more visually oriented social networks such as Snapchat and Instagram continue to draw in younger audiences. When asked “Which one social networking site or service do you use most often?” here are the top sites chosen by Americans aged 12–24 who currently use any social networking site or service.³




SOCMEDIA

Social Media Site	Percentage Who Use Most Often
Facebook	43%
Instagram	18%
Snapchat	15%
Twitter	8%
Google+	4%
Pinterest	3%

- What is the sum of the percentages for these top social media sites? What percent of Americans aged 12–24 use other social media sites most often?
- Make a bar graph to display these data. Be sure to include an “Other social media site” category.
- Would it be correct to display these data in a pie chart? Why or why not?

- 1.4 How Do Students Pay for College?** The Higher Education Research Institute's Freshman Survey includes more than 200,000 first-time, full-time freshmen who entered college in 2015.⁴ The survey reports the following data on the sources students use to pay for college expenses:  EXPENSE

Source for College Expenses	Students
Family resources	80.8%
Student resources	53.4%
Aid—not to be repaid	69.0%
Aid—to be repaid	44.4%

- (a) Explain why it is *not* correct to use a pie chart to display these data.
- (b) Make a bar graph of the data. Notice that because the data contrast groups such as family and student resources, it is better to keep these bars next to each other rather than to arrange the bars in order of height.
- 1.5 Never on Sunday?** Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2014:⁵  BIRTHS

Day	Births
Sunday	7,371
Monday	11,805
Tuesday	12,630
Wednesday	12,155
Thursday	12,112
Friday	12,042
Saturday	8,344

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

1.3 Quantitative Variables: Histograms

Quantitative variables often take many values. The distribution tells us what values the variable takes and how often it takes these values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram**.



What's that Number?

You might think that numbers, unlike words, are universal. Think again. A “billion” in the United States means 1,000,000,000 (nine zeros). In Europe, a “billion” is 1,000,000,000,000 (12 zeros). OK, those are words that describe numbers. But those commas in big numbers are periods in many other languages. This is so confusing that international standards call for spaces instead, so that an American billion is written 1 000 000 000. And the decimal point of the English-speaking world is the decimal comma in many other languages, so that 3.1416 in the United States becomes 3,1416 in Europe. So what is the number 10,642.389? It depends on where you are.

EXAMPLE 1.4 Making a Histogram



GRADRATE

What percent of your home state's high school students graduate within four years? The No Child Left Behind Act of 2001 used on-time high school graduation rates as one of its monitoring requirements. However, in 2001 most states were not collecting the necessary data to compute these rates accurately. The Freshman Graduation Rate (FGR) counts the number of high school graduates in a given year for a state and divides this by the number of ninth-graders enrolled four years previously. Although the FGR can be computed from readily available data, it neglects high school students moving into and out of a state and may include students who have repeated a grade. Several alternative measures are available that partially correct for these deficiencies, but states had been free to choose their own measure, and the resulting rates could differ by more than 10%. Federal law now requires all states to use a common, more rigorous computation, the *Adjusted Cohort Graduation Rate*, that tracks individual students. The use of the Adjusted Cohort Graduation Rate was first required for 2010–2011, and this finally allowed accurate comparisons of the graduation rates among states. Table 1.1 presents the data for 2013–2014.⁶

The *individuals* in this data set are the states. The *variable* is the percent of a state's high school students who graduate within four years. The states vary quite a bit on this variable, from 61.4% in the District of Columbia to 90.5% in Iowa. It's much easier to see how your state compares with other states from a graph like a histogram than from the table. To make a histogram of the distribution of this variable, proceed as follows:

Step 1. Choose the classes. Divide the range of the data into classes of equal width. The data in Table 1.1 range from 61.4 to 90.5, so we decide to use these classes:

- percent on-time graduates between 60.0 and 65.0 (60.0 to < 65.0)
- percent on-time graduates between 65.0 and 70.0 (65.0 to < 70.0)
- ⋮
- percent on-time graduates between 90.0 and 95.0 (90.0 to < 95.0)

TABLE 1.1 Percent of state high school students graduating on time

State	Percent	Region	State	Percent	Region	State	Percent	Region
Alabama	86.3	S	Louisiana	74.6	S	Ohio	81.8	MW
Alaska	71.1	W	Maine	86.5	NE	Oklahoma	82.7	S
Arizona	75.7	W	Maryland	86.4	S	Oregon	72.0	W
Arkansas	86.9	S	Massachusetts	86.1	NE	Pennsylvania	85.5	NE
California	81.0	W	Michigan	78.6	MW	Rhode Island	80.8	NE
Colorado	77.3	W	Minnesota	81.2	MW	South Carolina	80.1	S
Connecticut	87.0	NE	Mississippi	77.6	S	South Dakota	82.7	MW
Delaware	87.0	S	Missouri	87.3	MW	Tennessee	87.2	S
Florida	76.1	S	Montana	85.4	W	Texas	88.3	S
Georgia	72.5	S	Nebraska	89.7	MW	Utah	83.9	W
Hawaii	81.8	W	Nevada	70.0	W	Vermont	87.8	NE
Idaho	77.3	W	New Hampshire	88.1	NE	Virginia	85.3	S
Illinois	86.0	MW	New Jersey	88.6	NE	Washington	78.2	W
Indiana	87.9	MW	New Mexico	68.5	W	West Virginia	84.5	S
Iowa	90.5	MW	New York	77.8	NE	Wisconsin	88.6	MW
Kansas	85.7	MW	North Carolina	83.9	S	Wyoming	78.6	W
Kentucky	87.5	S	North Dakota	87.2	MW	District of Columbia	61.4	S

It is important to specify the classes carefully so that each individual falls into exactly one class. Our notation 60 to < 65 indicates that the first class includes states with graduation rates starting at 60.0% and up to, but not including, graduation rates of 65.0%. Thus, a state with an on-time graduation rate of 65.0% falls into the second class, whereas a state with an on-time graduation rate of 64.9% falls into the first class. It is equally correct to use classes 60.0 to < 64.0, 64.0 to < 68.0, and so forth. Just be sure to specify the classes precisely so that each individual falls into exactly one class.

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

Class	Count	Class	Count
60.0 to < 65.0	1	80.0 to < 85.0	11
65.0 to < 70.0	1	85.0 to < 90.0	23
70.0 to < 75.0	5	90.0 to < 95.0	1
75.0 to < 80.0	9		

Check that the counts add to 51, the number of individuals in the data set (the 50 states and the District of Columbia).

Step 3. Draw the histogram. Mark the scale for the variable whose distribution you are displaying on the horizontal axis. That's the percent of a state's high school students who graduate within four years. The scale runs from 60.0 to 95.0 because that is the span of 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. Draw the bars with no horizontal space between them unless a class is empty so that its bar has height zero. [Figure 1.5](#) is our histogram. Remember, an observation on the boundary of the bars—say, 65.0—is counted in the bar to its right.

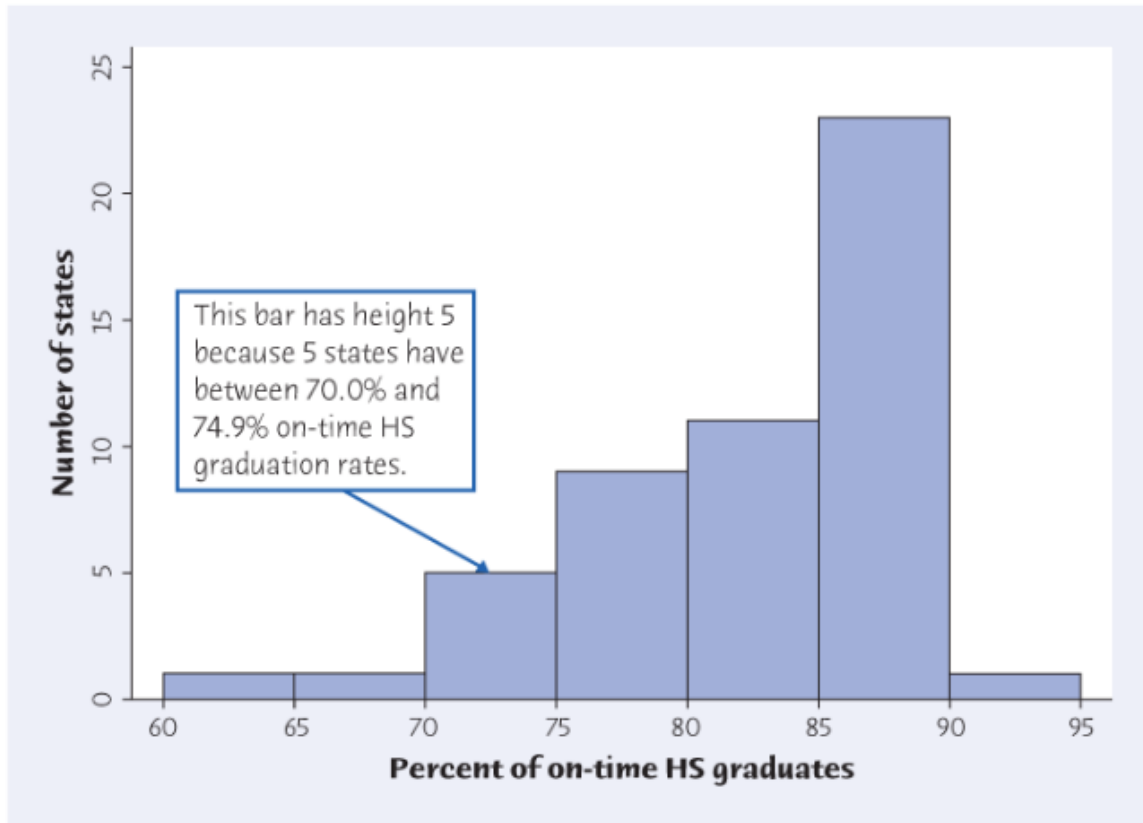


Figure 1.5

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.5

Histogram of the distribution of the percent of on-time high school graduates in 50 states and the District of Columbia, for [Example 1.4](#). This figure was created with the Minitab 17 software package.



Although histograms resemble bar graphs, their details and uses are different. A histogram displays the distribution of a quantitative variable. The horizontal axis of a histogram is marked in the units of measurement for the variable. A bar graph compares the sizes of different quantities. The horizontal axis of a bar graph simply identifies the quantities being compared and need not have any measurement scale. These quantities may be the values of a categorical variable, but they may also be unrelated, like the sources used to learn about music in [Example 1.3](#) (page 18). Draw bar graphs with blank space between the bars to separate the quantities being compared. Draw histograms with no space, to indicate that all values of the variable are covered. A gap between bars in a histogram indicates that there are no values for that class.

Our eyes respond to the *area* of the bars in a histogram.⁷ Because the classes are all the same width, area is determined by height, and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all values in a few classes with tall bars. Too many 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. The software’s choice is usually a good one, but you can change it if you want. The histogram function in the *One-Variable Statistical Calculator* applet on the text website allows you to change the number of classes by dragging with the mouse, so that it is easy to see how the choice of classes affects the histogram.



APPLY YOUR KNOWLEDGE

- 1.6 The Changing Face of America.** In 1980, approximately 20% of adults aged 18–34 were considered minorities, reporting their ethnicity as other than non-Hispanic white. By the end of 2013, that percentage had more than doubled. How are minorities between the ages of 18 and 34 distributed in the United States? In the country as a whole, 42.8% of adults aged 18–34 are considered minorities, but the states vary from 8% in Maine and Vermont to 75% in Hawaii. [Table 1.2](#) presents the data for all 50 states and the District of Columbia.⁸ Make a histogram of the percents using classes of width 10% starting at 0%. That is, the first bar covers 0% to < 10%, the second covers 10% to < 20%, and so on. (Make this histogram by hand, even if you have software, to be sure you understand the process. You may then want to compare your

histogram with your software’s choice.)



MINORITY

TABLE 1.2 Percent of state population aged 18–34 who are minorities

State	Percent	State	Percent	State	Percent
Alabama	39	Louisiana	45	Ohio	23
Alaska	40	Maine	8	Oklahoma	37
Arizona	51	Maryland	52	Oregon	27
Arkansas	31	Massachusetts	31	Pennsylvania	26
California	67	Michigan	28	Rhode Island	31
Colorado	35	Minnesota	23	South Carolina	41
Connecticut	39	Mississippi	48	South Dakota	19
Delaware	23	Missouri	23	Tennessee	30
Florida	52	Montana	15	Texas	61
Georgia	51	Nebraska	23	Utah	22
Hawaii	75	Nevada	54	Vermont	8
Idaho	20	New Hampshire	11	Virginia	41
Illinois	42	New Jersey	51	Washington	34
Indiana	23	New Mexico	67	West Virginia	9
Iowa	16	New York	48	Wisconsin	22
Kansas	27	North Carolina	41	Wyoming	18
Kentucky	17	North Dakota	15	District of Columbia	53

1.7

Choosing Classes in a Histogram. The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes the data on the percent minorities between the ages of 18 and 34 in the states from Table 1.2. Choose these data, then click on the “Histogram” tab to see a histogram.

**MINORITY**

- How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)
- Click on the graph and drag to the left. What is the smallest number of classes you can get? What are the lower and upper bounds of each class? (Click on the bar to find out.) Make a rough sketch of this histogram.
- Click and drag to the right. What is the greatest number of classes you can get? How many observations does the largest class have?
- You see that the choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram that you think best displays the distribution. How many classes did you use? Why do you think this is best?

1.4 Interpreting Histograms

Making a statistical graph is not an end in itself. *The purpose of graphs is to help us understand the data.* After you make a graph, always ask, “What do I see?” Once you have displayed a distribution, you can see its important features as follows.

Examining a Histogram

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

You can describe the overall pattern of a histogram by its **shape**, **center**, and **variability**. You will sometimes see variability referred to as **spread**.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

One way to describe the center of a distribution is by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. To find the midpoint, order the observations from smallest to largest, making sure to include repeated observations as many times as they appear in the data. First cross off the largest and smallest observations, then the largest and smallest of those remaining, and continue this process. If there were an odd number of observations initially, you will be left with a single observation, which is the midpoint. If there were an even number of observations initially, you will be left with two observations, and their average is the midpoint.

For now, we will describe the variability of a distribution by giving the *smallest and largest values*. We will learn better ways to describe center and variability in Chapter 2. The overall shape of a distribution can often be described in terms of symmetry or skewness, defined as follows.

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.

EXAMPLE 1.5 Describing a Distribution



GRADRATE

Look again at the histogram in [Figure 1.5](#). To describe the distribution, we want to look at its overall pattern and any deviations.

SHAPE: The distribution has a *single peak*, which represents states in which between 85.0% and 90.0% of students graduate high school on time. The distribution is *skewed to the left*. There is only one observation to the right of the peak, while to the left of the peak, most of the remaining states have graduation rates between 75.0% and 85.0%, but several states have much lower percents, so the graph extends quite far to the left of its peak.

CENTER: Arranging the observations from [Table 1.1](#) in order of size shows that 83.9% is the midpoint of the distribution. There are a total of 51 observations, and if we cross off the 25 highest graduation rates and the 25 lowest graduation rates, we are left with a single graduation rate of 83.9%, which we take as the center of the distribution.

VARIABILITY: The graduation rates range from 61.4% to 90.5%, which shows considerable variability in graduation rates among the states.

OUTLIERS: [Figure 1.5](#) shows no observations outside the overall single-peaked, left-skewed pattern of the distribution. [Figure 1.6](#) is another histogram of the same distribution, with classes of width 4% rather than 5%. Now the District of Columbia at 61.4%, is more clearly separated from the remaining states. Is the District of Columbia an outlier or just the smallest observations in a strongly skewed distribution? Unfortunately, there is no rule. Let's agree to call attention to only strong outliers that suggest something special about an observation—or an error such as typing 10.1 as 101. Although the District of Columbia is often included with the other 50 states in data sets, for many variables it can differ markedly from the remaining states.

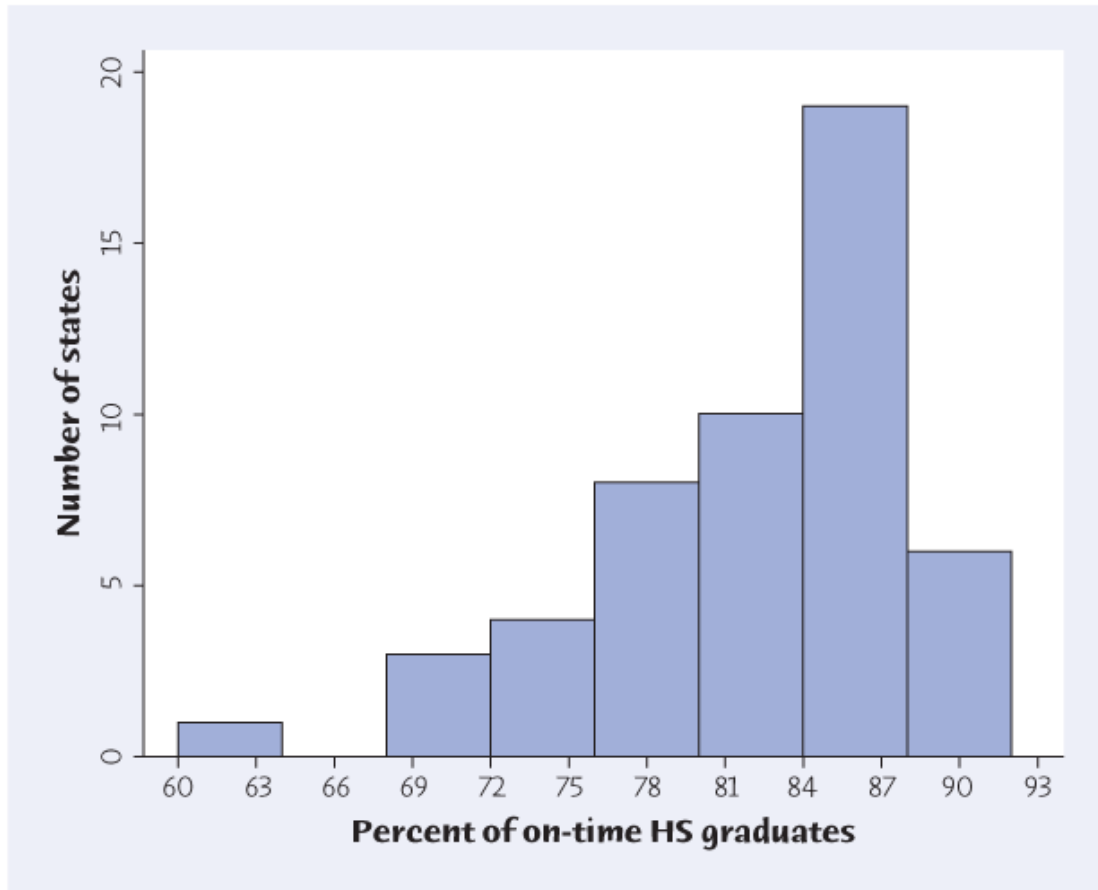


Figure 1.6

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.6

Another histogram of the distribution of the percent of on-time high school graduates, with narrower class widths than in Figure 1.5. Histograms with more classes show more detail but may have a less clear pattern.



Figures 1.5 and 1.6 remind us that interpreting graphs calls for judgment. We also see that *the choice of classes in a histogram can influence the appearance of a distribution*. Because of this, and to avoid worrying about minor details, concentrate on the main features of a distribution that persist with several choices of class intervals. Look for major peaks, not for minor ups and downs, in the bars of the histogram. When you choose a larger number of class intervals, the histogram can become more jagged, leading to the appearance of multiple peaks that are close together. Always, be sure to check for clear outliers, not just for the smallest and largest observations, and look for rough *symmetry* or clear *skewness*.

Here are more examples of describing the overall pattern of a histogram.

EXAMPLE 1.6 Iowa Tests Scores



IOWATEST

Figure 1.7 displays the scores of all 947 seventh-grade students in the public schools of Gary, Indiana, on the vocabulary part of the Iowa Tests of Basic Skills.⁹ The distribution is *single-peaked* and *symmetric*. In mathematics, the two sides of symmetric patterns are exact mirror images. Real data are almost never exactly symmetric. We are content to describe Figure 1.7 as symmetric. The center (half above, half below) is close to 7. This is seventh-grade reading level. The scores range from 2.0 (second-grade level) to 12.1 (twelfth-grade level).

Notice that the vertical scale in Figure 1.7 is not the *count* of students but the *percent* of students in each histogram class. A histogram of percents rather than counts is convenient when we want to compare several distributions. To compare Gary with Los Angeles, a much bigger city, we would use percents so that both histograms have the same vertical scale.

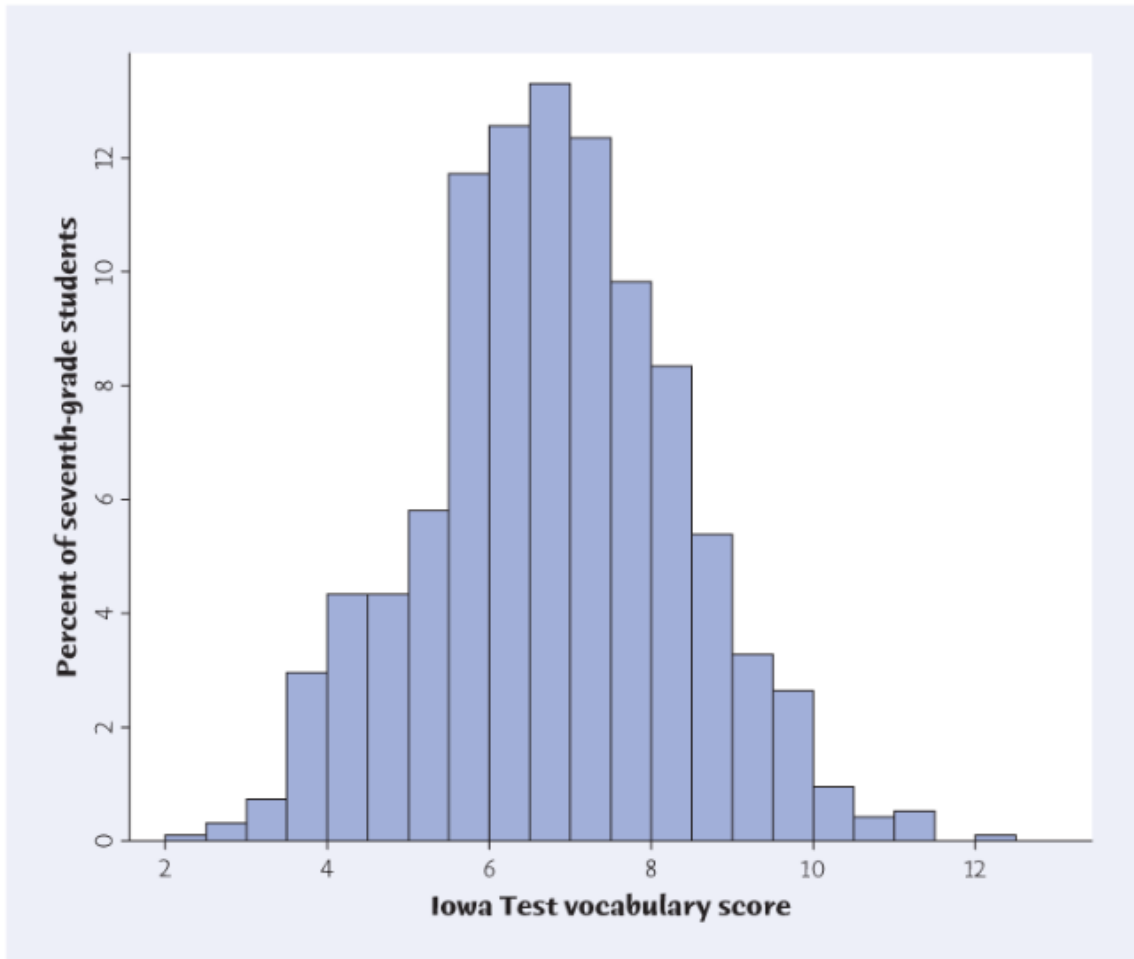


Figure 1.7

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.7

Histogram of the Iowa Tests vocabulary scores of all seventh-grade students in Gary, Indiana, for [Example 1.6](#). This distribution is single peaked and symmetric.

When describing the vertical scale of a histogram, you will sometimes see count referred to as **frequency** and percent referred to as **relative frequency**, particularly when choosing an option for the vertical scale using software.

EXAMPLE 1.7 Who Takes the SAT?



SATPCT

Depending on where you went to high school, the answer to this question may be “almost everybody” or “almost nobody.” Figure 1.8 is a histogram of the percent of high school graduates in each state who took the SAT test in 2015.¹⁰

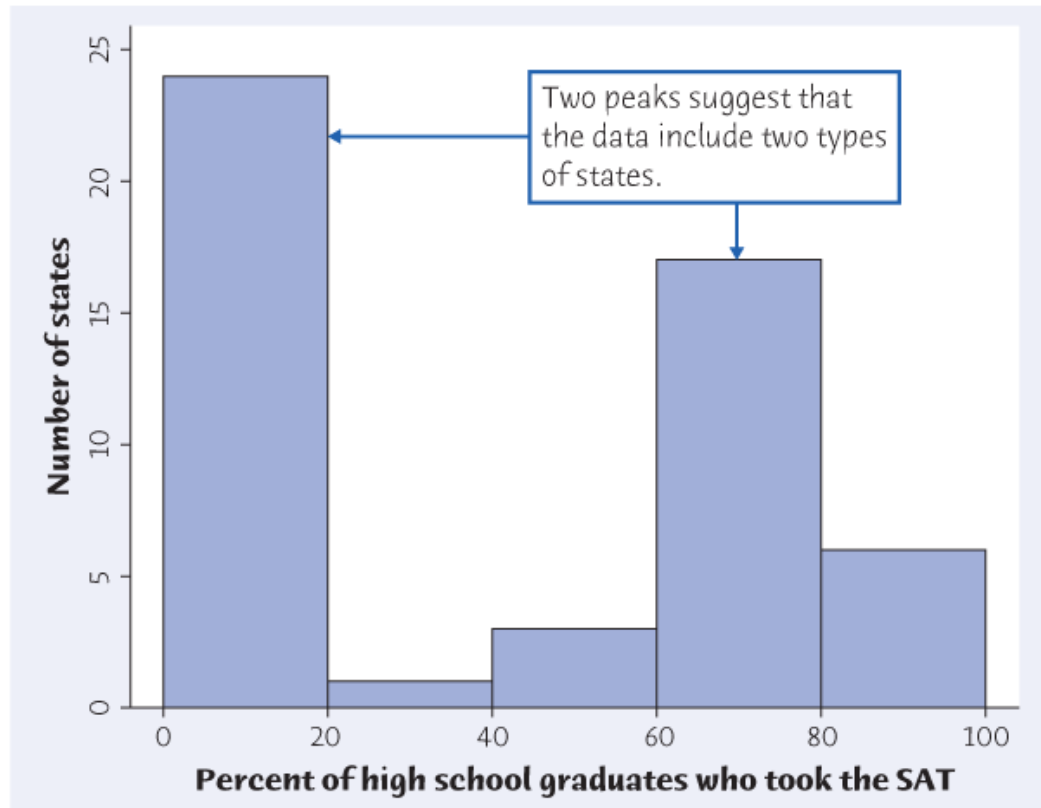


Figure 1.8

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.8

Histogram of the percent of high school graduates in each state who took the SAT Reasoning test, for [Example 1.7](#). The graph shows two groups of states: ACT states (where few students take the SAT) at the left and SAT states at the right.

The histogram shows two peaks: a high peak at the left and a lower peak in the 60% to < 80% class. The presence of more than one peak suggests that a distribution mixes several kinds of individuals. That is the case here. There are two major tests of readiness for college, the ACT and the SAT. Most states have a strong preference for one or the other. In some states, many students take the ACT exam and few take the SAT—these states form the peak on the left. In other states, many students take the SAT and few choose the ACT—these states form the lower peak at the right.

Giving the center and variability of this distribution is not very useful. The midpoint falls in the 40% to < 60% class, between the two peaks. The story told by the histogram is in the two peaks corresponding to ACT states and SAT states.

The overall shape of a distribution is important information about a variable. Some variables have distributions with predictable shapes. Many biological measurements on specimens from the same species and sex—lengths of bird bills, heights of young women—have symmetric distributions. On the other hand, data on people's incomes are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few enormous incomes. Many distributions have irregular shapes that are neither symmetric nor skewed. Some data show other patterns, such as the two peaks in [Figure 1.8](#). Use your eyes, describe the pattern you see, and then try to explain the pattern.

APPLY YOUR KNOWLEDGE

- 1.8 The Changing Face of America.** In [Exercise 1.6](#) (page 23), you made a histogram of the percent of minority residents aged 18–34 in each of the 50 states and the District of Columbia. These data are given in [Table 1.2](#). Describe the shape of the distribution. Is it closer to symmetric or skewed? What is the center (midpoint) of the data? What is the variability in terms of the smallest and largest values? Are there any states with an unusually large or small percent of minorities?



MINORITY



1.9 Lyme Disease. Lyme disease is caused by a bacteria called *Borrelia burgdorferi* and is spread through the bite of an infected black-legged tick, generally found in woods and grassy areas. There were 213,515 confirmed cases reported to the Centers for Disease Control (CDC) between 2001 and 2010, and these are broken down by age and sex in [Figure 1.9](#).¹¹

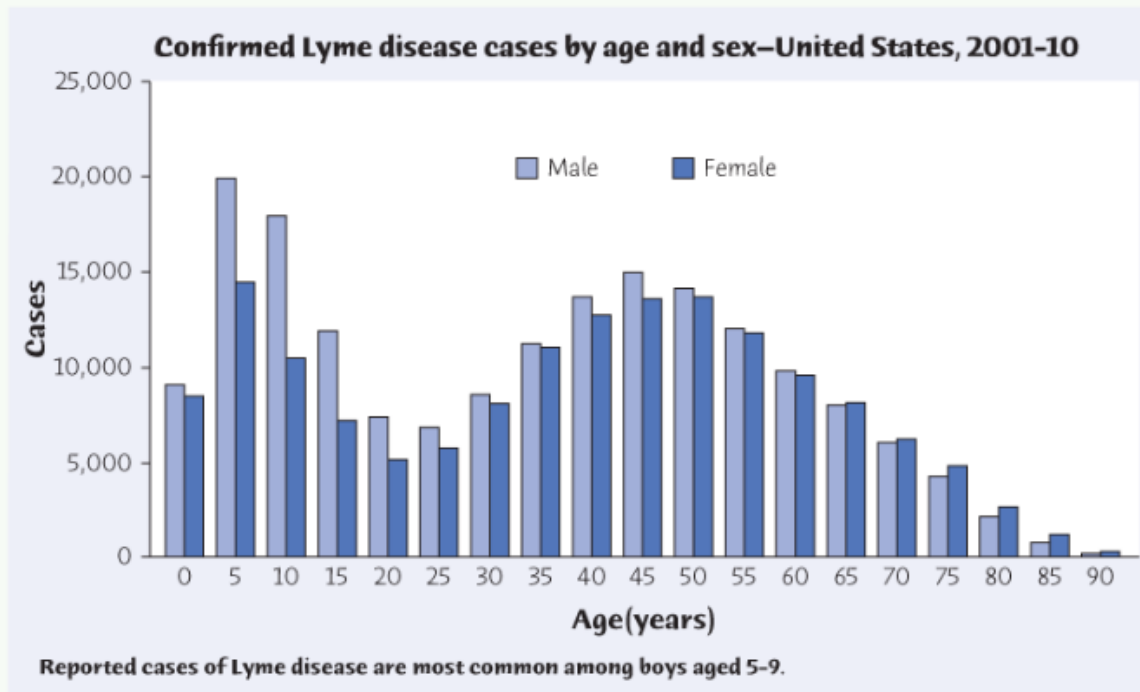


Figure 1.9

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.9

Histogram of the ages of infected individuals with Lyme disease for cases reported between 2001 and 2010 in the United States, for males and females, for [Exercise 1.9](#).

1.5 Quantitative Variables: Stemplots

Histograms are not the only graphical display of distributions. For small data sets, a *stemplot* 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.
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. Be sure to include all the stems needed to span the data, even when some stems will have no leaves.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

EXAMPLE 1.8 Making a Stemplot



MINORITY

Table 1.2 (page 24) presents the percent of adults aged 18–34 who were considered minorities in each of the states and the District of Columbia. To make a stemplot of these data, first write the percentages 8 and 9 as 08 and 09, so that all the percentages are two digit numbers. Take the tens place (leftmost digit) of the percent as the stem and the final digit (ones) as the leaf. Write stems from 0 for Maine, Vermont, and West Virginia to 7 for Hawaii. Now add leaves. Texas, 61%, has leaf 1 on the 6 stem. California and New Mexico, at 67%, each place a leaf of 7 on the same stem. These are the only observations on this stem. Arrange the leaves in order, so that 6|177 is one row in the stemplot. Figure 1.10 is the complete stemplot for the data in Table 1.2.

Here is how Figure 1.9 relates to what we have been studying. The individuals are the 213,515 people with confirmed cases, and two of the variables measured on each individual are sex and age. Considering males and females separately, we could draw a histogram of the variable age using class intervals 0 to < 5 years old, 5 to < 10 years old, and so forth. Look at the leftmost two bars in Figure 1.9. The light blue bar shows that approximately 9000 of the males were between 0 and 5 years old, and the dark blue bar shows slightly fewer females were in this age range. If we were to take all the light blue bars and put them side by side, we would have the histogram of age for males using the class intervals stated. Similarly, the dark blue bars show females. Because we are trying to display both histograms in the same graph, the bars for males and females within each class interval have been placed alongside each other for ease of comparison, with the bars for different class intervals separated by small spaces.

- (a) Describe the main features of the distribution of age for males. Why would describing this distribution in terms of only the center and variability be misleading?
- (b) Suppose that different age groups of males spend differing amounts of time outdoors. How could this fact be used to explain the pattern that you found in part (a)? Remember to use your eyes to describe the pattern you see, and then try to explain the pattern.
- (c) A 45-year-old male friend of yours looks at the histogram and tells you that he is planning on giving up hiking because this graph suggests he is in a high-risk group for getting Lyme disease. He will resume hiking when he is 65, as he will be less likely to get Lyme disease at that age. Is this a correct interpretation of the histogram?
- (d) Comparing the histograms for males and females, how are they similar? What is the main difference, and why do you think it occurs?

1.5 Quantitative Variables: Stemplots

Histograms are not the only graphical display of distributions. For small data sets, a *stemplot* 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.
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. Be sure to include all the stems needed to span the data, even when some stems will have no leaves.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

EXAMPLE 1.8 Making a Stemplot



MINORITY

Table 1.2 (page 24) presents the percent of adults aged 18–34 who were considered minorities in each of the states and the District of Columbia. To make a stemplot of these data, first write the percentages 8 and 9 as 08 and 09, so that all the percentages are two digit numbers. Take the tens place (leftmost digit) of the percent as the stem and the final digit (ones) as the leaf. Write stems from 0 for Maine, Vermont, and West Virginia to 7 for Hawaii. Now add leaves. Texas, 61%, has leaf 1 on the 6 stem. California and New Mexico, at 67%, each place a leaf of 7 on the same stem. These are the only observations on this stem. Arrange the leaves in order, so that 6|177 is one row in the stemplot. Figure 1.10 is the complete stemplot for the data in Table 1.2.

0		8 8 9
1		1 5 5 6 7 8 9
2		0 2 2 3 3 3 3 3 3 6 7 7 8
3		0 1 1 1 4 5 7 9 9
4		0 1 1 1 2 5 8 8
5		1 1 1 2 2 3 4
6		1 7 7
7		5

Figure 1.10

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.10

Stemplot of the percents of minorities aged 18–34 in the states, for [Example 1.8](#). The tens place of the percent is the stem and the ones place is the leaf.

A stemplot looks like a histogram turned on end, with the stems corresponding to the class intervals. The first stem in [Figure 1.10](#) contains all states with percents between 0% and 10%. Examine the histogram in [Figure 1.11](#), which is a histogram of the minority data using class intervals 0% to < 10%, 10% to < 20%, and so forth. Although [Figures 1.10](#) and [1.11](#) display exactly the same pattern, the stemplot, unlike the histogram, preserves the actual value of each observation.

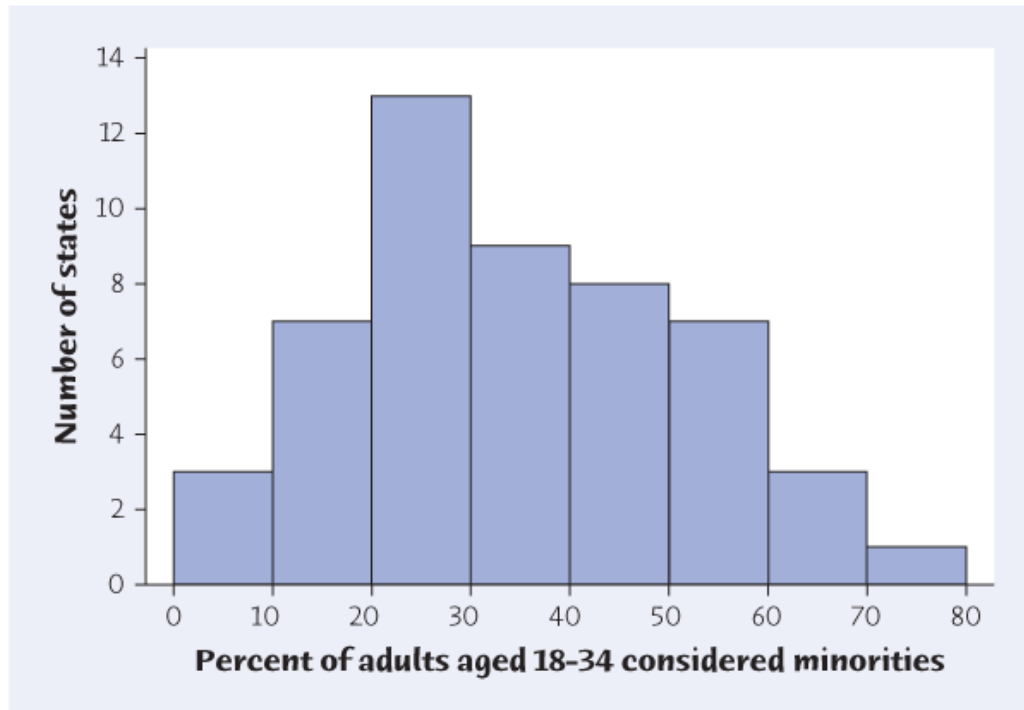


Figure 1.11

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.11

Histogram of the percents of minorities aged 18–34 in the states, for [Example 1.8](#). The class widths have been chosen to agree with the widths of the stems in the stemplot in [Figure 1.10](#).



The Vital Few

Skewed distributions can show us where to concentrate our efforts. Ten percent of the cars on the road account for half of all carbon dioxide emissions. A histogram of CO₂ emissions would show many cars with small or moderate values and a few with very high values. Cleaning up or replacing these cars would reduce pollution at a cost much lower than that of programs aimed at all cars. Statisticians who work at improving quality in industry make a principle of this: distinguish “the vital few” from “the trivial many.”

In a stemplot, the classes (the stems of a stemplot) are given to you. Histograms are more flexible than stemplots because you can choose the classes more easily. *Stemplots do not work well for large data sets, where each stem must hold a large number of leaves.* Don't try to make a stemplot of a large data set, such as the 947 Iowa Tests scores in [Figure 1.7](#).



Consider making a stemplot of the high school graduation rate data in [Table 1.1](#). If we use the tenths as the leaves, the necessary stems begin at 62 for the District of Columbia and end at 90 for Vermont, requiring a total of 39 stems. When there are too many stems as in this case, there are often no leaves or just one or two leaves on many of the stems. The number of stems can be reduced if we first **round** the data. In this example, we can round the data for each state to the nearest percent before drawing the stemplot. Here is the result:

```

6 | 1 9
7 | 0 1 2 3 5 6 6 7 7 8 8 8 9 9
8 | 0 1 1 1 2 2 3 3 4 4 5 5 5 6 6 6 6 6 6 7 7 7 7 7 7 8 8 8 8 8 9 9
9 | 0 1

```

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

Now it seems that there are too few stems. You can also **split stems** in a stemplot to double the number of stems when all the leaves would otherwise fall on just a few stems, as occurred when we rounded to the nearest percent. Each stem then appears twice. Leaves 0 to 4 go on the upper stem, and leaves 5 to 9 go on the lower stem. If you split the stems with the data rounded to the nearest percent, the stemplot becomes:

```

6 | 1
6 | 9
7 | 0 1 2 3
7 | 5 6 6 7 7 8 8 8 9 9
8 | 0 1 1 1 2 2 3 3 4 4
8 | 5 5 5 6 6 6 6 6 6 7 7 7 7 7 7 8 8 8 8 8 9 9
9 | 0 1

```

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
© 2018 W.H. Freeman and Company

which makes the left skew pattern clearer. In fact, examining [Figure 1.5](#) shows that the stems in the previous stemplot correspond to the class intervals used in the histogram of [Figure 1.5](#), although there are minor differences in the histogram and the stemplot because the data was rounded for the stemplot but not for the histogram. When drawing stemplots, some data require rounding but don't require splitting stems, some require just splitting stems, and other data require both. The *One-Variable Statistical Calculator* applet on the text website allows you to decide whether to split stems so that it is easy to see the effect.




Comparing [Figures 1.11](#) (right-skewed, [page 30](#)) and [1.5](#) (left-skewed, [page 23](#)) reminds us that *the direction of skewness is the direction of the long tail, not the direction where most observations are clustered.*

Macmillan Learning Online Resources

- The Snapshots video, *Visualizing Quantitative Data*, illustrates the ideas of both stemplots and histograms.
- The StatsBoards videos, *Creating and Interpreting a Histogram* and *Creating and Interpreting a Stemplot*, provide the details of constructing both stemplots and histograms through an example.

APPLY YOUR KNOWLEDGE

1.10 The Changing Face of America. [Figure 1.10](#) gives a stemplot of the percentages of adults aged 18–34 who were considered minorities in each of the states and the District of Columbia.  MINORITY

- Make another stemplot of this data by splitting the stems, placing leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem of the same value. Does splitting the stems give a different impression of the distribution? Explain.
- Draw a histogram of this data which uses class intervals that give the same pattern as the stemplot that you drew in part (a).

1.11 Health Care Spending. [Table 1.3](#) shows the 2013 per capita total expenditure on health in 35 countries with the highest gross domestic product in that year.¹² Health expenditure per capita is the sum of public and private health expenditure (in PPP, international \$) divided by population. Health expenditures include the provision of health services, family-planning activities, nutrition activities, and emergency aid designated for health but exclude the provision of water and sanitation. Make a stemplot of the data after rounding to the nearest \$100 (so that stems are thousands of dollars and leaves are hundreds of dollars). Split the stems, placing leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem of the same value. Describe the shape, center, and variability of the distribution. Which country is the high outlier?



TABLE 1.3 Per capita total expenditure on health (international dollars)

Country	Dollars	Country	Dollars	Country	Dollars
Argentina	1725	Indonesia	293	Saudi Arabia	1681
Australia	4191	Iran	1414	South Africa	1121
Austria	4885	Italy	3126	Spain	2846
Belgium	4526	Japan	3741	Sweden	4244
Brazil	1454	Korea, South	2398	Switzerland	6187
Canada	4759	Malaysia	938	Thailand	658
China	646	Mexico	1061	Turkey	1053
Colombia	843	Netherlands	5601	United Arab Emirates	2233
Denmark	4552	Nigeria	207	United Kingdom	3311
France	4334	Norway	6308	United States	9146
Germany	4812	Poland	1551	Venezuela	656
India	215	Russia	1587		

1.6 Time Plots

Many variables are measured at intervals over time. We might, for example, measure the height of a growing child or the price of a stock at the end of each month. In these examples, our main interest is change over time. To display change over time, make a *time plot*.

Time Plot

A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

EXAMPLE 1.9 Water Levels in the Everglades



WATERLEV

Water levels in Everglades National Park are critical to the survival of this unique region. The photo shows a water-monitoring station in Shark River Slough, the main path for surface water moving through the “river of grass” that is the Everglades. Each day the mean gauge height, the height in feet of the water surface above the gauge datum, is measured at the Shark River Slough monitoring station. (The gauge datum is a vertical control measure established in 1929 and is used as a reference for establishing varying elevations. It establishes a zero point from which to measure the gauge height.) [Figure 1.12](#) is a time plot of mean daily gauge height at this station from January 2, 2000, to June 30, 2016.¹³



When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. Figure 1.12 shows strong **cycles**, regular up-and-down movements in water level. The cycles show the effects of Florida's wet season (roughly June to November) and dry season (roughly December to May). Water levels are highest in late fall. If you look closely, you can see the year-to-year variation. The dry season in 2003 ended early, with the first-ever April tropical storm. In consequence, the dry-season water level in 2003 did not dip as low as in other years. The drought in the southeastern portion of the country in 2008 and 2009 shows up in the steep drop in the mean gauge height in 2009, whereas the lower peaks in 2006 and 2007 reflect lower water levels during the wet seasons in these years. Finally, in 2011, an extra-long dry season and a slow start to the 2011 rainy season compounded into the worst drought in the southwest Florida area in 80 years, which shows up as the steep drop in the mean gauge height in 2011.

Another common overall pattern in a time plot is a **trend**, a long-term upward or downward movement over time. Many economic variables show an upward trend. Incomes, house prices, and (alas) college tuitions tend to move generally upward over time.

Histograms and time plots give different kinds of information about a variable. The time plot in Figure 1.12 presents **time series data** that show the change in water level at one location over time. A histogram displays **cross-sectional data**, such as water levels at many locations in the Everglades at the same time.

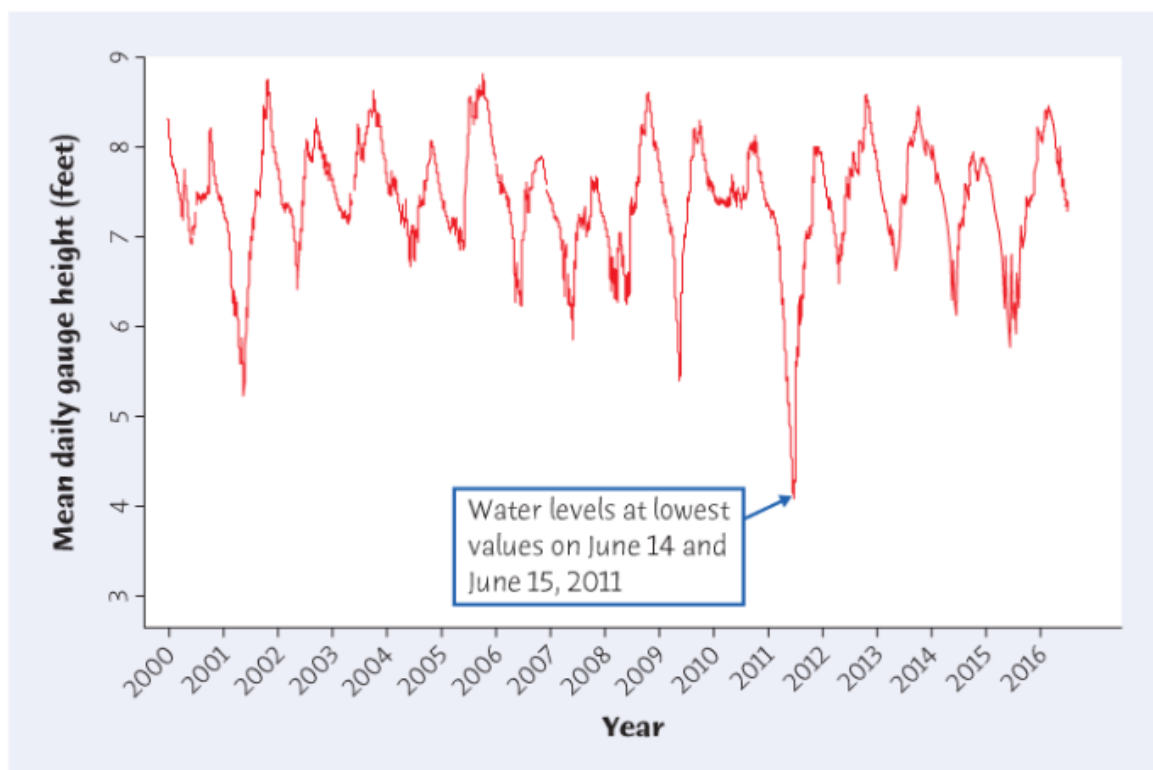


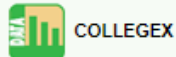
Figure 1.12

FIGURE 1.12

Time plot of average gauge height at a monitoring station in Everglades National Park over a 16-year period, for [Example 1.9](#). The yearly cycles reflect Florida's wet and dry seasons. This figure was created with the Minitab 17 software package.

APPLY YOUR KNOWLEDGE

1.12 The Cost of College. Here are data on the average tuition and fees charged to in-state students by public four-year colleges and universities for the 1980 to 2015 academic years. Because almost any variable measured in dollars increases over time due to inflation (the falling buying power of a dollar), the values are given in “constant dollars,” adjusted to have the same buying power that a dollar had in 2015.¹⁴



Year	Tuition	Year	Tuition	Year	Tuition	Year	Tuition
1980	\$2323	1990	\$3498	2000	\$4854	2010	\$8351
1981	\$2371	1991	\$3700	2001	\$5074	2011	\$8742
1982	\$2529	1992	\$3971	2002	\$5440	2012	\$9006
1983	\$2747	1993	\$4197	2003	\$6028	2013	\$9077
1984	\$2821	1994	\$4359	2004	\$6459	2014	\$9161
1985	\$2923	1995	\$4408	2005	\$6708	2015	\$9410
1986	\$3088	1996	\$4530	2006	\$6807		
1987	\$3120	1997	\$4634	2007	\$7093		
1988	\$3184	1998	\$4757	2008	\$7160		
1989	\$3260	1999	\$4821	2009	\$7838		

- Make a time plot of average tuition and fees.
- What overall pattern does your plot show?
- Some possible deviations from the overall pattern are outliers, periods when charges went down (in 2015 dollars) and periods of particularly rapid increase. Which are present in your plot, and during which years?
- In looking for patterns, do you think that it would be better to study a time series of the tuition for each year or the percent increase for each year? Why?

CHAPTER 1 SUMMARY

Chapter Specifics

- 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, sex, or salary.
- Some variables are **categorical**, and others are **quantitative**. A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that describe some characteristic of each individual using a **unit of measurement**, such as height in centimeters or salary in dollars.
- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.
- After you understand the background of your data (individuals, variables, units of measurement), almost always the first thing to do is **plot your data**.
- The **distribution** of a variable describes what values the variable takes and how often it takes these values. **Pie charts** and **bar graphs** display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units. **Histograms** and **stemplots** graph the distribution of a quantitative variable.
- When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.
- **Shape, center, and variability** describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as **symmetric** or **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.
- **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers, and try to explain them.
- When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends, cycles**, or other changes over time.

Link It

Practical statistics uses data to draw conclusions about some broader universe. You should reread [Example 1.1](#) (page 14), as it will help you understand this basic idea. For the American Community Survey described in the example, the data are the responses from those households responding to the survey, although the broader universe of interest is the entire nation.

In our study of practical statistics, we will divide the subject into three main areas. In exploratory data analysis, graphs and numerical summaries are used for exploring, organizing, and describing data so that the patterns become apparent. Data production concerns where the data come from and helps us to understand whether what we learn from our data can be generalized to a wider universe. And statistical inference provides tools for generalizing what we learn to a wider universe.

In this chapter, we have begun to learn about data analysis. A data set can consist of hundreds of observations on many variables. Even if we consider only one variable at a time, it is difficult to see what the data have to say by scanning a list containing many data values. Graphs provide a visual tool for organizing and identifying patterns in data and are a good starting point in the exploration of the distribution of a variable.

Pie charts and bar graphs can summarize the information in a categorical variable by giving us the percent of the distribution in the various categories. Although a table containing the categories and percent gives the same information as a bar graph, a substantial advantage of the bar graph over a tabular presentation is that the bar graph allows us to visually compare percents among all categories simultaneously by means of the heights of the bars.

Histograms and stemplots are graphical tools for summarizing the information provided by a quantitative variable. The overall pattern in a histogram or stemplot illustrates some of the important features of the distribution of a variable that will be of interest as we continue our study of practical statistics. The center of the histogram tells us about the value of a “typical” observation on this variable, whereas the variability gives us a sense of how close most of the observations are to this value. Other interesting features are the presence of outliers and the general shape of the plot. For data collected over time, time plots can show patterns such as seasonal variation and trends in the variable. In the next chapter, we will see how the information about the distribution of a variable can also be described using numerical summaries.

Macmillan Learning Online Resources

If you are having difficulty with any of the sections of this chapter, these online resources should help prepare you to solve the exercises at the end of this chapter.

- StatTutor starts with a video review of each section and asks a series of questions to check your understanding.
- LearningCurve provides you with a series of questions about the chapter that adjust to your level of understanding.

There are also online resources to help you use technology.

- The statistical software packages CrunchIt! and JMP are available online.
- The Video Technology Manuals for Minitab, the TI graphing calculator, Excel, JMP, CrunchIt!, R, and SPSS provide explicit instructions for producing graphical output similar to that provided in Chapter 1 for the particular technology you are using.

CHECK YOUR SKILLS

The multiple-choice exercises in *Check Your Skills* ask straightforward questions about basic facts from the chapter. Answers to all odd-numbered exercises appear in the back of the book. You should expect almost all your answers to be correct.

1.13 Here are the first lines of a professor's data set at the end of a statistics course:

Name	Major	Total Points	Grade
ADVANI, SURA	COMM	397	B
BARTON, DAVID	HIST	323	C
BROWN, ANNETTE	BIOL	446	A
CHIU, SUN	PSYC	405	B
CORTEZ, MARIA	PSYC	461	A

The individuals in these data are

- (a) the students.
 - (b) the total points.
 - (c) the course grades.
- 1.14** According to the National Household Survey on Drug Use and Health, when asked in 2012, 41% of those aged 18 to 24 years used cigarettes in the past year, 9% used smokeless tobacco, 36.3% used illicit drugs, and 10.4% used pain relievers or sedatives.¹⁵ To display this data, it would be correct to use
- (a) either a pie chart or a bar graph.
 - (b) a pie chart provided a category for other is added to get to 100%.
 - (c) a bar graph but not a pie chart.
- 1.15** A description of different houses on the market includes the variables square footage of the house and the average monthly gas bill.
- (a) Square footage and average monthly gas bill are both categorical variables.
 - (b) Square footage and average monthly gas bill are both quantitative variables.
 - (c) Square footage is a categorical variable, and average monthly gas bill is a quantitative variable.
- 1.16** A political party's data bank includes the zip codes of past donors, such as
47906 34236 53075 10010 90210 75204 30304 99709
- Zip code is a
- (a) quantitative variable.
 - (b) categorical variable.
 - (c) unit of measurement.

1.17 Figure 1.6 (page 26) is a histogram of the percent of on-time high school graduates in each state. The rightmost bar in the histogram covers percents of on-time high school graduates ranging from about

- (a) 87% to 90%.
- (b) 88% to 92%.
- (c) 90% to 93%.

1.18 Here are the exam scores of 10 students in a statistics class:

50 35 41 97 76 69 94 91 23 65

To make a stemplot of these data, you would use stems

- (a) 2, 3, 4, 5, 6, 7, 9.
- (b) 2, 3, 4, 5, 6, 7, 8, 9.
- (c) 20, 30, 40, 50, 60, 70, 80, 90.

1.19 Where do students go to school? Although 80.4% of first-time first-year students attended college in the state in which they lived, this percent varied considerably over the states. Here is a stemplot of the percent of first-year students in each of the 50 states who were from the state where they enrolled. The

stems are 10s and the leaves are 1s. The stems have been split in the plot.¹⁶



INSTATE

3	4								
3	8								
4	0								
4	6								
5	0								
5	6								
6	0 0 1 2 2 3 4								
6	5 9								
7	2 2 2 4 4 4								
7	5 6 6 6 6 7 7 8 8 9								
8	0 0 1 2 2 3								
8	5 5 7 7 8 9								
9	0 0 1 2 3 3 4								

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

The midpoint of this distribution is

- (a) 60%.
- (b) 76%.
- (c) 80%.

1.20 The shape of the distribution in [Exercise 1.19](#) is

- (a) skewed to the left.
- (b) skewed upward.
- (c) skewed to the right.

1.21 The state with the smallest percent of first-year students enrolled in the state has

- (a) 0.34% enrolled.
- (b) 3.4% enrolled.
- (c) 34% enrolled.

1.22 You look at real estate ads for houses in Naples, Florida. There are many houses ranging from \$200,000 to \$500,000 in price. The few houses on the water, however, have prices up to \$15 million. The distribution of house prices will be

- (a) skewed to the left.
- (b) roughly symmetric.
- (c) skewed to the right.

CHAPTER 1 EXERCISES

- 1.23 Medical students.** Students who have finished medical school are assigned to residencies in hospitals to receive further training in a medical specialty. Here is part of a hypothetical database of students seeking residency positions. USMLE is the student's score on Step 1 of the national medical licensing examination.

Name	Medical School	Sex	Age	USMLE	Specialty Sought
Abrams, Laurie	Florida	F	28	238	Family medicine
Brown, Gordon	Meharry	M	25	205	Radiology
Cabrera, Maria	Tufts	F	26	191	Pediatrics
Ismael, Miranda	Indiana	F	32	245	Internal medicine

- (a) What individuals does this data set describe?
- (b) In addition to the student's name, how many variables does the data set contain? Which of these variables are categorical, and which are quantitative? If a variable is quantitative, what units is it measured in?
- 1.24 Buying a refrigerator.** *Consumer Reports* will have an article comparing refrigerators in the next issue. Some of the characteristics to be included in the report are the brand name and model; whether it has a top, bottom, or side-by-side freezer; the estimated energy consumption per year (kilowatts); whether or not it is Energy Star compliant; the width, depth, and height in inches; and both the freezer and refrigerator net capacity in cubic feet. Which of these variables are categorical, and which are quantitative? Give the units for the quantitative variables and the categories for the categorical variables. What are the individuals in the report?
- 1.25 What color is your car?** The most popular colors for cars and light trucks vary with region, type of vehicle, and over time. In North America, silver and gray are the most popular choices for midsize cars, black and red for sports cars, and white for light trucks. Despite this variation, overall white remains the top choice worldwide for the fifth consecutive year, increasing its lead by 7% over the previous year.

Here is the distribution of the top colors for vehicles sold globally in 2015:¹⁷



CARCOLOR

Color	Popularity
White	35%
Black	17%
Silver	12%
Gray	11%
Red	8%
Beige, brown	8%
Blue	7%
Other colors	

Fill in the percent of vehicles that are in other colors. Make a graph to display the distribution of color popularity.

- 1.26 High school tobacco use.** Despite the intense anti-smoking campaigns sponsored by both federal and private agencies, smoking continues to be the single-biggest cause of preventable death in the United States. How has the tobacco use of high school students changed over the past few years? For each of several tobacco products, high school students were asked whether they had used each of them in the past 30 days. Here are some of the results:¹⁸

Product	Year				
	2011	2012	2013	2014	2015
Any tobacco product	24.3	23.3	22.9	24.6	25.3
Cigarettes	15.8	14.0	12.7	9.2	9.3
Cigars	12.6	11.6	11.9	8.2	8.6
Pipes	4.5	4.0	4.1	1.5	1.0
Smokeless tobacco	7.3	6.4	5.7	5.5	6.0
E-cigarettes	1.5	2.8	4.5	13.4	16.0

The first row of the table gives the percentages of high school students who had used any tobacco product including cigarettes, pipes, cigars, smokeless tobacco, e-cigarettes, hookahs, snus, bidis or dissolvable tobacco in the last 30 days for the years 2011–2015. The remaining rows give the percentage of high school students using the most common tobacco products in each of these years.

- (a) Using the information in the first row of the table, draw a bar chart that shows the change in the use of any tobacco product between 2011 and 2015. How would you describe the pattern of change in this usage?

- (b) Draw a bar chart that illustrates the change in usage in these years for the individual tobacco products. If your software allows it, give a single bar chart that contains the information for all products. Otherwise, provide a separate bar chart for each product.
- (c) Using the bar charts in parts (a) and (b), give a simple description of the changes in the use of tobacco products by high school students between 2011 and 2015.

1.27 Deaths among young people. Among persons aged 15–24 years in the United States, there were 28,486 deaths in 2013. The leading causes of death and number of deaths were: accidents, 11,619; suicide, 4878; homicide, 4329; cancer, 1496; heart disease, 1170; congenital defects, 362.¹⁹

(a) Make a bar graph to display these data.

(b) Can you make a pie chart using the information given? Explain carefully why or why not.

1.28 Student debt. At the end of 2014, the average outstanding student debt for graduate and undergraduate study combined was \$26,700, yielding a total outstanding student debt of \$1.16 trillion. Figure 1.13 is a pie chart to show the distribution of outstanding education debt.²⁰ About what percent of students had an outstanding debt between \$25,000 and \$49,999? \$150,000 or more? You see that it is hard to determine numbers from a pie chart. Bar graphs are much easier to use. (Many agencies include the percents in their pie charts to aid in interpretation.)

Pie Chart of the Distribution of Outstanding Education Debt, 2014

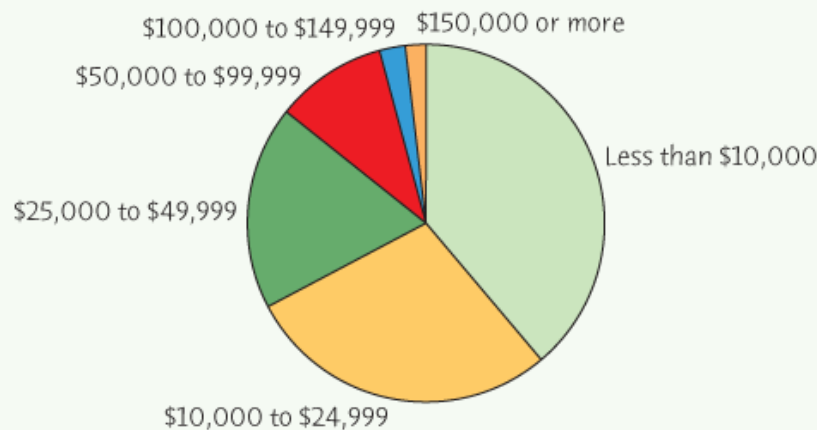


Figure 1.13

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
© 2018 W.H. Freeman and Company

FIGURE 1.13

Pie chart of the distribution of outstanding education debt, for Exercise 1.28.

1.29 Time spent on mobile apps. Social media and entertainment account for two-thirds of the time spent on mobile apps, with those between 18 and 34 spending approximately two hours a day using social media and entertainment apps. Social media apps have become the primary platform for users to keep informed on news, culture, and their family and friends, while entertainment apps fill pockets of free time throughout the day. Here is the average monthly time spent per mobile app user on social and entertainment categories by age group.²¹



MOBILAPP

Age Group	Hours Using Social Apps	Hours Using Entertainment Apps
18 to 34 years	29.6	29.2
35 to 54 years	25.4	17.4
Over 54 years	18.3	10.0

- If your software allows it, draw a bar graph with adjacent bars for social and entertainment hours for each of the three age categories, allowing easy comparison of social and entertainment hours within each age category. If your software does not allow this, draw two bar charts, one for the social hours and the second for the entertainment hours.
- Describe the main differences in social and entertainment usage by age group.
- Explain carefully why it is not correct to make a pie chart for the social usage hours or the entertainment usage hours.

1.30 Do adolescent girls eat fruit? We all know that fruit is good for us. Many of us don't eat enough. [Figure 1.14](#) is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.²² Describe the shape, center, and variability of this distribution. Are there any outliers? What percent of these girls ate six or more servings per day? How many of these girls ate fewer than two servings per day?

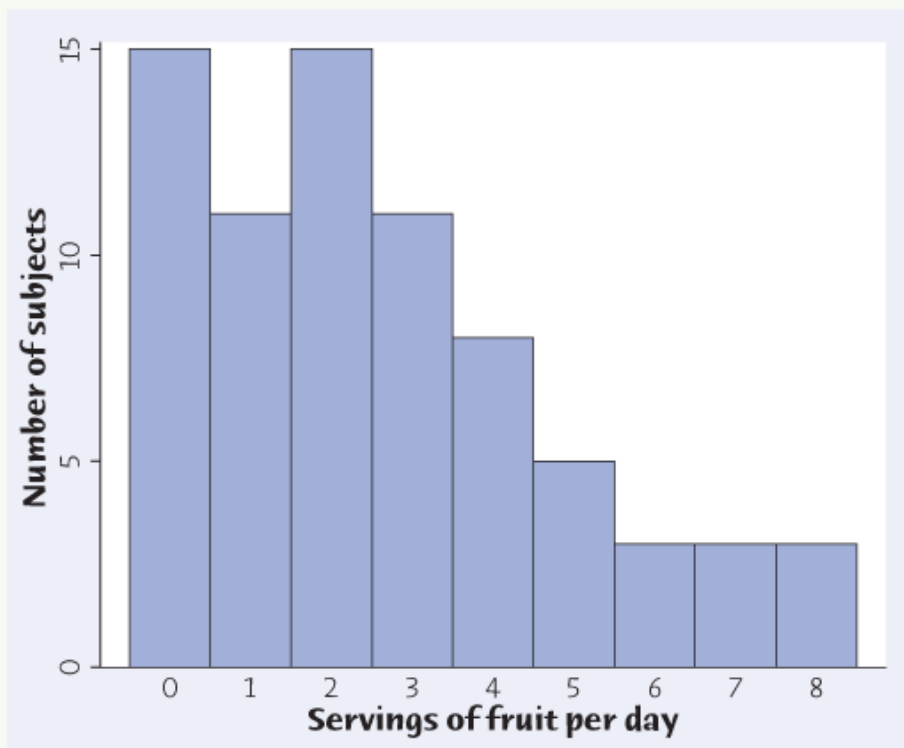



Figure 1.14

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
© 2018 W.H. Freeman and Company

FIGURE 1.14

The distribution of fruit consumption in a sample of 74 seventeen-year-old girls, for [Exercise 1.30](#).

1.31 IQ test scores. [Figure 1.15](#) is a stemplot of the IQ test scores of 78 seventh-grade students in a rural midwestern school.²³  IQ


- Four students had low scores that might be considered outliers. Ignoring these, describe the shape, center, and variability of the remainder of the distribution.
- We often read that IQ scores for large populations are centered at 100. What percent of these 78 students have scores above 100?

7		2 4
7		7 9
8		
8		6 9
9		0 1 3 3
9		6 7 7 8
10		0 0 2 2 3 3 3 3 4 4
10		5 5 5 6 6 6 7 7 7 7 8 9
11		0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 4 4 4 4
11		5 5 6 8 8 9 9 9
12		0 0 3 3 4 4
12		6 7 7 8 8 8
13		0 2
13		6

Figure 1.15
 Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
 © 2018 W.H. Freeman and Company

FIGURE 1.15

The distribution of IQ scores for 78 seventh-grade students, for [Exercise 1.31](#).

1.32 Returns on common stocks. The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. [Figure 1.16](#) is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets from January 1985 to December 2015 (372 months).²⁴ The extreme low outlier is the market crash of October 1987, when stocks lost 23% of their value in one month. The other two low outliers are 16% during August 1998, a month when the Dow Jones Industrial Average experienced its second-largest drop in history to that time, and the financial crisis in October 2008, when stocks lost 17% of their value.  STOCKRET

- Ignoring the outliers, describe the overall shape of the distribution of monthly returns.
- What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the months having lower returns and half having higher returns.)
- Approximately what were the smallest and largest monthly returns, leaving out the outliers? (This is one way to describe the variability of the distribution.)
- A return less than zero means that stocks lost value in that month. About what percent of all months had returns less than zero?

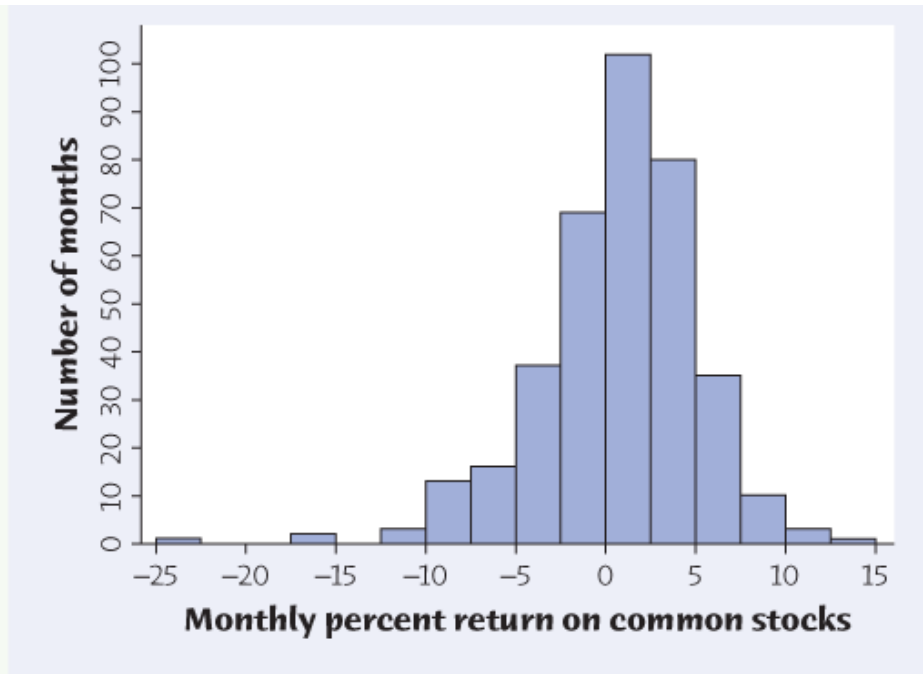


Figure 1.16

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
© 2018 W.H. Freeman and Company

FIGURE 1.16

The distribution of monthly percent returns on U.S. common stocks from January 1985 to December 2015, for [Exercise 1.32](#).

1.33 Name that variable. A survey of a large college class asked the following questions:

1. Are you female or male? (In the data, male = 0, female = 1.)
2. Are you right-handed or left-handed? (In the data, right = 0, left = 1.)
3. What is your height in inches?
4. How many minutes do you study on a typical weeknight? [Figure 1.17](#) shows histograms of the student responses, in scrambled order and without scale markings. Which graph goes with each variable? Explain your reasoning.

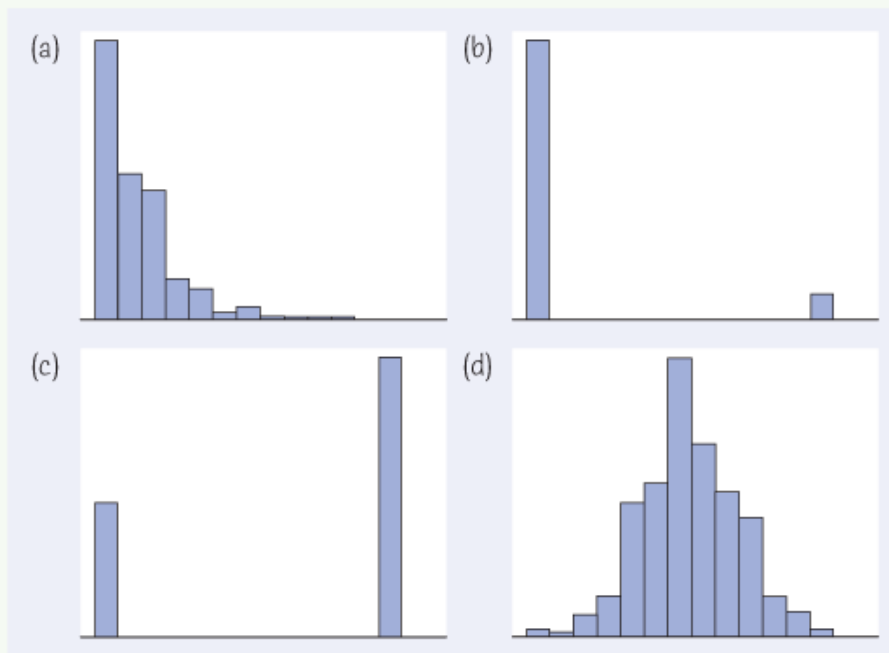


Figure 1.17

Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
© 2018 W.H. Freeman and Company

FIGURE 1.17

Histograms of four distributions, for [Exercise 1.33](#).

1.34 Food oils and health. Fatty acids, despite their unpleasant name, are necessary for human health. Two types of essential fatty acids, called omega-3 and omega-6, are not produced by our bodies and so must be obtained from our food. Food oils, widely used in food processing and cooking, are major sources of these compounds. There is some evidence that a healthy diet should have more omega-3 than omega-6. [Table 1.4](#) gives the ratio of omega-3 to omega-6 in some common food oils.²⁵ Values greater than 1 show that an oil has more omega-3 than omega-6.



FOODOILS

- Make a histogram of these data, using classes bounded by the whole numbers from 0 to 6.
- What is the shape of the distribution? How many of the 30 food oils have more omega-3 than omega-6? What does this distribution suggest about the possible health effects of modern food oils?
- [Table 1.4](#) contains entries for several fish oils (cod, herring, menhaden, salmon, sardine). How do these values support the idea that eating fish is healthy?

FIGURE 1.17Histograms of four distributions, for [Exercise 1.33](#).

1.34 Food oils and health. Fatty acids, despite their unpleasant name, are necessary for human health. Two types of essential fatty acids, called omega-3 and omega-6, are not produced by our bodies and so must be obtained from our food. Food oils, widely used in food processing and cooking, are major sources of these compounds. There is some evidence that a healthy diet should have more omega-3 than omega-6. [Table 1.4](#) gives the ratio of omega-3 to omega-6 in some common food oils.²⁵ Values greater than 1 show that an oil has more omega-3 than omega-6.



FOODOILS

- Make a histogram of these data, using classes bounded by the whole numbers from 0 to 6.
- What is the shape of the distribution? How many of the 30 food oils have more omega-3 than omega-6? What does this distribution suggest about the possible health effects of modern food oils?
- [Table 1.4](#) contains entries for several fish oils (cod, herring, menhaden, salmon, sardine). How do these values support the idea that eating fish is healthy?

TABLE 1.4 Omega-3 fatty acids as a fraction of omega-6 fatty acids in food oils

Oil	Ratio	Oil	Ratio
Perilla	5.33	Flaxseed	3.56
Walnut	0.20	Canola	0.46
Wheat germ	0.13	Soybean	0.13
Mustard	0.38	Grape seed	0.00
Sardine	2.16	Menhaden	1.96
Salmon	2.50	Herring	2.67
Mayonnaise	0.06	Soybean, hydrogenated	0.07
Cod liver	2.00	Rice bran	0.05
Shortening (household)	0.11	Butter	0.64
Shortening (industrial)	0.06	Sunflower	0.03
Margarine	0.05	Corn	0.01
Olive	0.08	Sesame	0.01
Shea nut	0.06	Cottonseed	0.00
Sunflower (oleic)	0.05	Palm	0.02
Sunflower (linoleic)	0.00	Cocoa butter	0.04

1.35 Where are the nurses? Table 1.5 gives the number of active nurses per 100,000 people in each state.²⁶



- Why is the number of nurses per 100,000 people a better measure of the availability of nurses than a simple count of the number of nurses in a state?
- Make a stemplot that displays the distribution of nurses per 100,000 people. The data will first need to be rounded (see page 30). What units are you going to use for the stems? The leaves? You should round the data to the units you are planning to use for the leaves before drawing the stemplot. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?
- Do you think it would be useful to split the stems when drawing the stemplot for these data? Explain your reason.

TABLE 1.5 Nurses per 100,000 people, by state

State	Nurses	State	Nurses	State	Nurses
Alabama	911	Louisiana	881	Ohio	1021
Alaska	717	Maine	1093	Oklahoma	742
Arizona	585	Maryland	906	Oregon	803
Arkansas	798	Massachusetts	1260	Pennsylvania	1030
California	630	Michigan	849	Rhode Island	1104
Colorado	831	Minnesota	1093	South Carolina	834
Connecticut	1017	Mississippi	950	South Dakota	1296
Delaware	1155	Missouri	1038	Tennessee	984
Florida	814	Montana	855	Texas	678
Georgia	665	Nebraska	1054	Utah	635
Hawaii	689	Nevada	609	Vermont	914
Idaho	682	New Hampshire	1006	Virginia	764
Illinois	901	New Jersey	858	Washington	814
Indiana	901	New Mexico	614	West Virginia	953
Iowa	1022	New York	848	Wisconsin	946
Kansas	934	North Carolina	940	Wyoming	864
Kentucky	1003	North Dakota	968	District of Columbia	1483

1.36 Child mortality rates. Although child mortality rates have dropped by more than 50% since 1990, in 2015 it was still the case that 16,000 children under five years old died each day. The mortality rates for children under five varied from 1.9 per 1000 in Luxembourg to 156.9 per 1000 in Angola. Although the data set includes 213 countries, the child mortality rates of 22 country were not available on the World Health Organization database. The data set is too large to print here, but here are the data for the first five countries.²⁷



MORTALTY

Country	Child Mortality Rate (per 1000)
Aruba	—
Andorra	2.800
Afghanistan	91.100
Angola	156.900
Albania	14.000

- Why do you think that mortality rates are measured as the number of deaths per 1000 children under age five rather than simply the number of deaths?
- Make a histogram that displays the distribution of child mortality rates. Describe the shape, center, and variability of the distribution. Do any countries appear to be obvious outliers in the histogram?

1.37 Fur seals on St. Paul Island. Every year, hundreds of thousands of northern fur seals return to their haul-outs in the Pribilof Islands in Alaska to breed, give birth, and teach their pups to swim, hunt, and survive in the Bering Sea. U.S. commercial fur sealing operations continued on St. Paul until 1984, but despite a reduction in harvest, the population of fur seals has continued to decline. Possible reasons include climate shifts in the North Pacific, changes in the availability of prey, and new or increased interaction with commercial fisheries that increase mortality. Here are data on the estimated number of fur seal pups born on St. Paul Island (in thousands) from 1979 to 2014, where a dash indicates a year in which no data were collected.²⁸



FURSEALS



Year	Pups Born (thousands)	Year	Pups Born (thousands)	Year	Pups Born (thousands)	Year	Pups Born (thousands)
1979	245.93	1988	202.23	1997	—	2006	109.96
1980	203.82	1989	171.53	1998	179.15	2007	—
1981	179.44	1990	201.30	1999	—	2008	102.67
1982	203.58	1991	—	2000	158.74	2009	—
1983	165.94	1992	182.44	2001	—	2010	94.50
1984	173.27	1993	—	2002	145.72	2011	—
1985	182.26	1994	192.10	2003	—	2012	96.83
1986	167.66	1995	—	2004	122.82	2013	—
1987	171.61	1996	170.12	2005	—	2014	91.74

Make a stemplot to display the distribution of pups born per year. Describe the shape, center, and variability of the distribution. Are there any outliers?

- 1.38 Nintendo and laparoscopic skills.** In laparoscopic surgery, a video camera and several thin instruments are inserted into the patient's abdominal cavity. The surgeon uses the image from the video camera positioned inside the patient's body to perform the procedure by manipulating the instruments that have been inserted. It has been found that the Nintendo Wii™ reproduces the movements required in laparoscopic surgery more closely than other video games with its motion-sensing interface. If training with a Nintendo Wii can improve laparoscopic skills, it can complement the more expensive training on a laparoscopic simulator. Forty-two medical residents were chosen, and all were tested on a set of basic laparoscopic skills. Twenty-one were selected at random to undergo systematic Nintendo Wii training for one hour a day, five days a week, for four weeks. The remaining 21 residents were given no Nintendo Wii training and asked to refrain from video games during this period. At the end of four weeks, all 42 residents were tested again on the same set of laparoscopic skills. One of the skills involved a virtual gall bladder removal, with several performance measures including time to complete the task recorded. Here are the improvement (before–after) times in seconds after four weeks for the two groups:²⁹




NINTENDO

Treatment						Control					
281	134	186	128	84	243	21	66	54	85	229	92
212	121	134	221	59	244	43	27	77	-29	-14	88
79	333	-13	-16	71	-16	145	110	32	90	46	-81
71	77	144				68	61	44			

- (a) In the context of this study, what do the negative values in the data set mean?
- (b) **Back-to-back stemplots** can be used to compare the two samples. That is, use one set of stems with two sets of leaves, one to the right and one to the left of the stems. (Draw a line on either side of the stems to separate stems and leaves.) Order both sets of leaves from smallest at the stem to largest away from the stem. Complete the back-to-back stemplot started below. The data have been rounded to the nearest 10, with stems being 100s and leaves being 10s. The stems have been split. The first control observation corresponds to -80 and the next two to -30 and -10 .
- (c) Report the approximate midpoints of both groups. Does it appear that the treatment has resulted in a greater improvement in times than seen in the control group? (To better understand the magnitude of the improvements, note that the median time to complete the task on the first occasion was 11 minutes and 40 seconds, using the times of all 42 residents.)

Treatment		Control
	-0	8
1 2 2	-0	3 1
	0	
	0	
	1	
	1	
	2	
	2	
	3	

Moore/Notz/Fligner, *The Basic Practice of Statistics*,
8e, © 2018 W.H. Freeman and Company

1.39 Fur seals on St. Paul Island. Make a time plot of the number of fur seals born per year from [Exercise 1.37](#). What does the time plot show that your stemplot in [Exercise 1.37](#) did not show? When you have data collected over time, a time plot is often needed to understand what is happening.  FURSEALS

1.40 Marijuana and traffic accidents. Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents, and they asked the drivers about marijuana use. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same age.³⁰

	Marijuana Use per Year			
	Never	1–10 Times	11–50 Times	51+ Times
Accidents caused	59	36	15	50
Drivers	452	229	70	156

- Explain carefully why a useful graph must compare *rates* (accidents per driver) rather than *counts* of accidents in the four marijuana use classes.
- Compute the accident rates in the four marijuana use classes. After you have done this, make a graph that displays the accident rate for each class. What do you conclude? (You can't conclude that marijuana use *causes* accidents because risk takers are more likely both to drive aggressively and to use marijuana.)

1.41 Accessing digital media in the U.S., Canada, and the U.K. Most mobile usage occurs via apps, particularly on smartphones, with time on smartphone apps now surpassing time spent on desktops in the United States. Digital media can be accessed on mobile platforms by using apps on a smartphone or tablet or by using a browser on a smartphone or tablet, or it can be accessed on the more traditional desktop platform (including laptops). Here is the breakdown of the share of digital media time by access method in the United States, Canada, and the United Kingdom.³¹



DIGMEDIA

Access Method	Canada	U.S.	U.K.
Desktop	48%	39%	44%
Smartphone (apps)	31%	43%	34%
Smartphone (browsers)	5%	6%	7%
Tablet (apps)	14%	10%	12%
Tablet (browsers)	2%	2%	3%

- Draw a bar graph for the distribution of the share of digital time by access method for Canada. Do the same for the United States and United Kingdom, using the same scale for the percent axis.
- Describe the most important difference in the distributions of the share of digital time by access method for the three countries. How does this difference show up in the bar graphs?
- Explain why it *is* appropriate to use a pie chart to display any of these distributions. Draw a pie chart for each distribution. Do you think it is easier to compare the three distributions with bar graphs or pie charts? Explain your reasoning.

1.42 She sounds tall! Presented with recordings of a pair of people of the same sex speaking the same phrase, can a listener determine which speaker is taller simply from the sound of their voice? Twenty-four young adults at Washington University listened to 100 pairs of speakers and, within each pair, were asked to indicate which of the two speakers was the taller. Here are the number correct (out of 100) for each of the 24 participants.³²

65 61 67 59 58 62 56 67 61 67 63 53
68 49 66 58 69 70 65 56 68 56 58 70

Researchers believe that the key to correct discrimination is contained in a particular type of sound produced in the lower airways or the lungs, known as subglottal resonances, whose frequency is lower for taller people. Despite the masking of these resonances by other voice sounds, researchers wondered whether the information they contained could still be heard by listeners and used to identify the taller person.

- Make two stemplots, with and without splitting the stems. Which plot do you prefer and why?
- Describe the shape, center, and variability of the distribution. Are there any outliers?
- If the experimental subjects are just guessing which speaker is taller, they should correctly identify the taller person about 50% of the time. Does this data support the researchers' conjecture that there is information in a person's voice to help identify the taller person? Why or why not?

1.43 Watch those scales! Figures 1.18(a) and 1.18(b) both show time plots of tuition charged to in-state students from 1980 through 2015.³³

- Which graph appears to show the biggest increase in tuition between 2000 and 2015?
- Read the graphs and compute the actual increase in tuition between 2000 and 2015 in each graph. Do you think these graphs are for the same or different data sets? Why? The impression that a time plot gives depends on the scales you use on the two axes. Changing the scales can make tuition appear to increase very rapidly or to have only a gentle increase. The moral of this exercise is: always pay close attention to the scales when you look at a time plot.



Figure 1.18
Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e, © 2018 W.H. Freeman and Company

FIGURE 1.18

Time plots of in-state tuition between 1980 and 2015, for Exercise 1.43.

1.44 The value of a four-year degree! The big economic news of 2007 was a severe downturn in housing that began in mid-2006. This was followed by the financial crisis in 2008. How did these economic events affect the unemployment rate, and were all segments of the population affected similarly? The data are the monthly unemployment rates for those over 25 years of age with a high school diploma and no college and those over 25 year of age with a four-year college degree from January 1992 through December 2015. The data set is too large to print here, but here are the data for the unemployment rates for both groups for the first five months:³⁴



UNEMPLOY

Month	Four-Year College Degree	HS Degree Only
January 1992	3.1	6.8
February 1992	3.2	7.0
March 1992	2.9	6.9
April 1992	3.2	6.9
May 1992	3.2	6.9

- Make a time plot of the monthly unemployment rates for those over 25 years of age with a high school diploma and no college and those over 25 year of age with a four-year college degree. If your software allows it, make both time plots on the same set of axes. Otherwise, make separate time plots for each group but use the same scale for both plots for ease of comparison. Are the patterns in the two time plots similar? What is the primary difference between the two time plots?
- How are economic events described reflected in the time plots of the unemployment rates? Since the end of 2009, how would you describe the behavior of the unemployment rate for both groups?
- Are there any other periods during which there were patterns in the unemployment rate? Describe them.

1.45 Housing starts. Figure 1.19 is a time plot of the number of single-family homes started by builders each month from January 1990 through June 2016.³⁵ The counts are in thousands of homes.



HOUSING

- The most notable pattern in this time plot is yearly up-and-down cycles. At what season of the year are housing starts highest? Lowest? The cycles are explained by the weather in the northern part of the country.
- Is there a longer-term trend visible in addition to the cycles? If so, describe it.
- The big economic news of 2007 was a severe downturn in housing that began in mid-2006. This was followed by the financial crisis in 2008. How are these economic events reflected in the time plot?
- How would you describe the behavior of the time plot since January 2011?

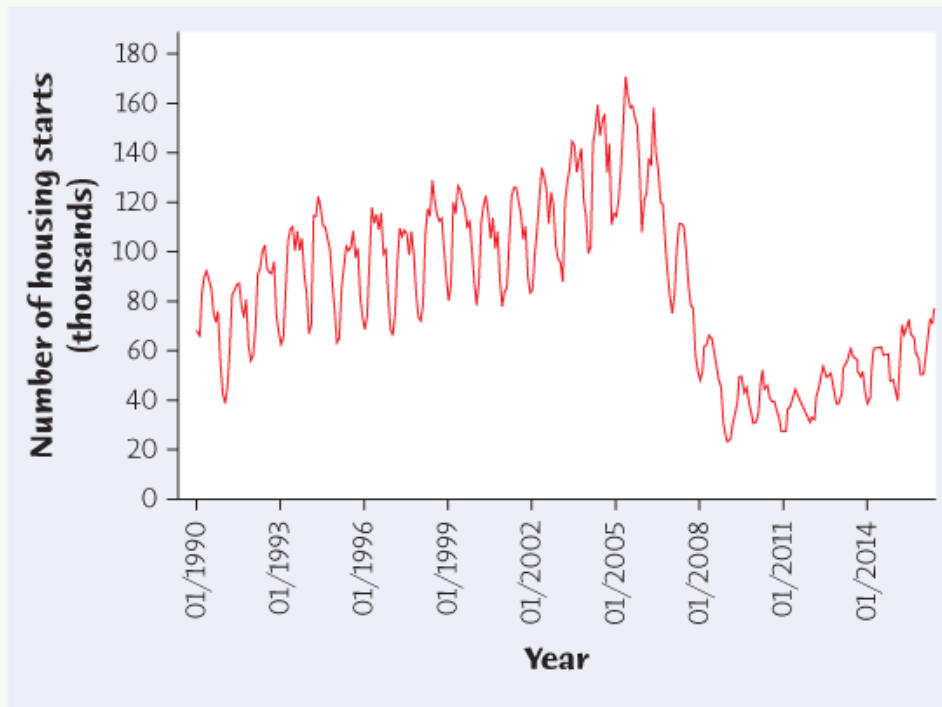


Figure 1.19
 Moore/Notz/Fligner, *The Basic Practice of Statistics*, 8e,
 © 2018 W.H. Freeman and Company

FIGURE 1.19

Time plot of the monthly count of new single-family houses started (in thousands) between January 1990 and June 2016, for [Exercise 1.45](#).

1.46 Choosing class intervals. Student engineers learn that, although handbooks give the strength of a material as a single number, in fact the strength varies from piece to piece. A vital lesson in all fields of study is that “variation is everywhere.” Here are data from a typical student laboratory exercise: the load in pounds needed to pull apart pieces of Douglas fir 4 inches long and 1.5 inches square:



33,190 31,860 32,590 26,520 33,280
 32,320 33,020 32,030 30,460 32,700
 23,040 30,930 32,720 33,650 32,340
 24,050 30,170 31,300 28,730 31,920

The data sets in the *One-Variable Statistical Calculator* applet on the text website include the “pulling wood apart” data given in this exercise. How many class intervals does the applet choose when drawing the histogram? Use the applet to make several histograms with a larger number of class intervals. Are there any important features of the data that are revealed using a larger number of class intervals? Which histogram do you prefer? Explain your choice.