



# KING SCHOOL

**2017 Summer Assignment**

**MAT 510: AP Statistics**

## **Summer Math Review Packet**

### **MAT 510: AP Statistics**

1. Read chapter 1 from the textbook and take notes on important concepts (vocabulary, types of graphs, examples, rules, etc).
1. Complete the following problems:
  - a. Introduction p.7-8 #1-7 (odd), 8
  - b. Section 1.1 p.22-26 #10,12,14,18,20,22,24,27-32 (all)
  - c. Section 1.2 p.42-49 #38,40,46,48,50,52,54,60,68-74
  - d. Section 1.3 p.70-74  
#80,82,84,86,88,90,94,96,98,100,102,104,106-110
2. Show ALL of your work on a separate sheet of paper and be prepared for a test on this material when you return to school.



**Introduction**

**In the Introduction, you'll learn about:**

- Individuals and variables
- From data analysis to inference

Data analysis

# Data Analysis: Making Sense of Data

Statistics is the science of data. The volume of data available to us is overwhelming. For example, the Census Bureau's American Community Survey collects data from 3,000,000 housing units each year. Astronomers work with data on tens of millions of galaxies. The checkout scanners at Walmart's 6500 stores in 15 countries record hundreds of millions of transactions every week. In all these cases, the data are trying to tell us a story—about U.S. households, objects in space, or Walmart shoppers. To hear what the data are saying, we need to help them speak by organizing, displaying, summarizing, and asking questions. That's **data analysis**.

## Individuals and Variables

Any set of data contains information about some group of **individuals**. The characteristics we measure on each individual are called **variables**.

**DEFINITION: Individuals and variables**

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

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

A high school's student data base, 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 age, gender, grade point average, homeroom, and grade level. In practice, any set of data is accompanied by background information that helps us understand the data. When you first meet a new data set, ask yourself the following questions:

1. Who are the individuals described by the data? How many individuals are there?
2. What are the variables? In what *units* is each variable recorded? Weights, for example, might be recorded in grams, pounds, thousands of pounds, or kilograms.

We could follow a newspaper reporter's lead and extend our list of questions to include *Why*, *When*, *Where*, and *How* were the data produced? For now, we'll focus on the first two questions.

Some variables, like gender and grade level, simply place individuals into categories. Others, like age and grade point average (GPA), take numerical values for which we can do arithmetic. It makes sense to give an average GPA for a group of students, but it doesn't make sense to give an "average" gender.

**AP EXAM TIP** If you learn to distinguish categorical from quantitative variables now, it will pay big rewards later. The type of data determines what kinds of graphs and which numerical summaries are appropriate. You will be expected to analyze categorical and quantitative data effectively on the AP exam.

### DEFINITION: Categorical variable and quantitative variable

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

A **quantitative variable** takes numerical values for which it makes sense to find an average.



Not every variable that takes number values is quantitative. Zip code is one example. Although zip codes are numbers, it doesn't make sense to talk about the average zip code. In fact, zip codes place individuals (people or dwellings) into categories based on location. Some variables—such as gender, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. For instance, we could classify people in a data set by age: 0–9, 10–19, 20–29, and so on.

The proper method of analysis for a variable depends on whether it is categorical or quantitative. As a result, it is important to be able to distinguish these two types of variables.

## EXAMPLE

### Census at School

#### Data, individuals, and variables

CensusAtSchool is an international project that collects data about primary and secondary school students using surveys. Hundreds of thousands of students from Australia, Canada, New Zealand, South Africa, and the United Kingdom have taken part in the project since 2000. Data from the surveys are available at the project's Web site ([www.censusatschool.com](http://www.censusatschool.com)). We used the site's "Random Data Selector" to choose 10 Canadian students who completed the survey in a recent year. The table below displays the data.

Province	Gender	Languages spoken	Handed	Height (cm)	Wrist circum. (mm)	Preferred communication	Travel to school (min)
Ontario	Male	1	Right	175	175	Internet chat or MSN	25
Alberta	Female	3	Right	147	140	MySpace/Facebook	20
Ontario	Male	1	Right	165	170	Internet chat	4
British Columbia	Female	1	Right	155	145	In person	10
New Brunswick	Male	9	Left	130.5	130	Other	40
Ontario	Male	2	Right	170	165	In person	7
Ontario	Male	3	Left	150	100	Internet chat	10
New Brunswick	Male	2	Both	167.5	220	Internet chat	30
Ontario	Female	1	Right	161	104	Text messaging	10
Ontario	Male	6	Right	190.5	180	Internet chat	10

**PROBLEM:**

- Who are the individuals in this data set?
- What variables were measured? Identify each as categorical or quantitative. In what units were the quantitative variables measured?
- Describe the individual in the highlighted row.

**SOLUTION:**

- The individuals are the 10 randomly selected Canadian students.
- The eight variables measured are province where student lives (categorical), gender (categorical), number of languages spoken (quantitative, in whole numbers), dominant hand (categorical), height (quantitative, in centimeters), wrist circumference (quantitative, in millimeters), preferred communication method (categorical), and travel time to school (quantitative, in minutes).
- This student lives in Ontario, is male, speaks three languages, is left-handed, is 150 cm tall (about 59 inches), has a wrist circumference of 100 mm (about 4 inches), prefers to communicate via Internet chat, and travels 10 minutes to get to school.

 For Practice Try Exercise 3

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

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

Most data tables follow the format shown in the example – each row is an individual, and each column is a variable. Sometimes the individuals are called *cases*.

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

**DEFINITION: Distribution**

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



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

**How to Explore Data**

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

### CHECK YOUR UNDERSTANDING



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

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

## From Data Analysis to Inference

### Inference

Sometimes, we're interested in drawing conclusions that go beyond the data at hand. That's the idea of **inference**. In the CensusAtSchool example, 7 of the 10 randomly selected Canadian students are right-handed. That's 70% of the *sample*. Can we conclude that 70% of the *population* of Canadian students who participated in CensusAtSchool are right-handed? No. If another random sample of 10 students was selected, the percent who are right-handed would probably not be exactly 70%. Can we at least say that the actual population value is "close" to 70%? That depends on what we mean by "close."

The following Activity gives you an idea of how statistical inference works.

### ACTIVITY Hiring discrimination—it just won't fly!

MATERIALS: Deck of cards for each student or pair of students



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

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

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

1. Separate the cards into piles by suit. Use ten cards from one suit to represent the female pilots. To represent the 15 male pilots, you'll need all 13 cards of another suit plus two extra cards from a third suit. A second student or group can use the leftover cards from the deck to set up their simulation in a similar way.

2. Shuffle your stack of 25 cards thoroughly and deal 8 cards. Count the number of female pilots selected. Record this value in a table like the one shown below.

Trial:	1	2	3	4	5
No. of females:					

3. Return the 8 cards to your stack. Shuffle and deal four more times so that you have a total of five simulated lottery results.
4. Your teacher will draw and label axes for a class dotplot. Each student should plot the number of females obtained in each of the five simulation trials on the graph.
5. Discuss the results with your classmates. Does it seem believable that airline managers carried out a fair lottery? What advice would you give the male pilot who contacted you?
6. Would your advice change if the lottery had chosen 6 female (and 2 male) pilots? Explain.

Our ability to do inference is determined by how the data are produced. Chapter 4 discusses the two primary methods of data production—sampling and experiments—and the types of conclusions that can be drawn from each. As the Activity illustrates, the logic of inference rests on asking, “What are the chances?” *Probability*, the study of chance behavior, is the topic of Chapters 5 through 7. We’ll introduce the most common inference techniques in Chapters 8 through 12.

## INTRODUCTION

## Summary

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

## INTRODUCTION

## Exercises

- 1.** Protecting wood How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables recorded: type of wood (yellow poplar, pine, cedar); type of water repellent (solvent-based, water-based); paint thickness (millimeters); paint color (white, gray, light blue); weathering time (months). Identify each variable as categorical or quantitative.

- 2.** Medical study variables Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Here are some of the variables recorded: gender (female or male); age (years); race (Asian, black, white, or other); smoker (yes or no); systolic blood pressure (millimeters of mercury); level of calcium in the blood (micrograms per milliliter). Identify each as categorical or quantitative.

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

pg 3

Gender	Hand	Height (in)	Homework time (min)	Favorite music	Pocket change (cents)
F	L	65	200	Hip-hop	50
M	L	72	30	Country	35
M	R	62	95	Rock	35
F	L	64	120	Alternative	0
M	R	63	220	Hip-hop	0
F	R	58	60	Alternative	75
F	R	67	150	Rock	215

- (a) What individuals does this data set describe?  
(b) Clearly identify each of the variables. Which are quantitative? In what units are they measured?  
(c) Describe the individual in the highlighted row.

- 4.** Coaster craze Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The table below displays data on several roller coasters that were opened in 2009.<sup>2</sup>

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (s)
Wild mouse	Steel	49.3	Sit down	28	70
Terminator	Wood	95	Sit down	50.1	180
Mania	Steel	140	Flying	56	155
Prowler	Wood	102.3	Sit down	51.2	150
Diamondback	Steel	230	Sit down	80	180

- (a) What individuals does this data set describe?  
(b) Clearly identify each of the variables. Which are quantitative? In what units are they measured?  
(c) Describe the individual in the highlighted row.

- 5.** Ranking colleges Popular magazines rank colleges and universities on their “academic quality” in serving undergraduate students. Describe two categorical variables and two quantitative variables that you might record for each institution. Give the units of measurement for the quantitative variables.

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

**Multiple choice: Select the best answer.**

Exercises 7 and 8 refer to the following setting. At the Census Bureau Web site, you can view detailed data collected by the American Community Survey. The table below includes data for 10 people chosen at random from the more than one million people in households contacted by the survey. “School” gives the highest level of education completed.

Weight (lb)	Age (yr)	Travel to work (min)	Travel		Income last year (\$)
			School	Gender	
187	66	0	Ninth grade	1	24,000
158	66	n/a	High school grad	2	0
176	54	10	Assoc. degree	2	11,900
339	37	10	Assoc. degree	1	6,000
91	27	10	Some college	2	30,000
155	18	n/a	High school grad	2	0
213	38	15	Master's degree	2	125,000
194	40	0	High school grad	1	800
221	18	20	High school grad	1	2,500
193	11	n/a	Fifth grade	1	0

7. The individuals in this data set are
- households.
  - people.
  - adults.
  - 120 variables.
  - columns.

8. This data set contains
- 7 variables, 2 of which are categorical.
  - 7 variables, 1 of which is categorical.
  - 6 variables, 2 of which are categorical.
  - 6 variables, 1 of which is categorical.
  - None of these.

## 1.1

**In Section 1.1,  
you'll learn about:**

- Bar graphs and pie charts
- Graphs: Good and bad
- Two-way tables and marginal distributions
- Relationships between categorical variables: Conditional distributions
- Organizing a statistical problem
- Simpson's paradox\*

**EXAMPLE****Radio Station Formats****Distribution of a categorical variable**

The radio audience rating service Arbitron places the country's 13,838 radio stations into categories that describe the kinds of programs they broadcast. Here are two different tables showing the distribution of station formats.<sup>3</sup>

Frequency table	
Format	Count of stations
Adult contemporary	1,556
Adult standards	1,196
Contemporary hit	569
Country	2,066
News/Talk/Information	2,179
Oldies	1,060
Religious	2,014
Rock	869
Spanish language	750
Other formats	1,579
Total	13,838

Relative frequency table	
Format	Percent of stations
Adult contemporary	11.2
Adult standards	8.6
Contemporary hit	4.1
Country	14.9
News/Talk/Information	15.7
Oldies	7.7
Religious	14.6
Rock	6.3
Spanish language	5.4
Other formats	11.4
Total	99.9

Frequency table  
Relative frequency  
table

In this case, the *individuals* are the radio stations and the *variable* being measured is the kind of programming that each station broadcasts. The table on the left, which we call a **frequency table**, displays the counts (*frequencies*) of stations in each format category. On the right, we see a **relative frequency table** of the data that shows the percents (*relative frequencies*) of stations in each format category.

\*This is an interesting topic, but it is not required for the AP Statistics exam.

Roundoff error

It's a good idea to check data for consistency. The counts should add to 13,838, the total number of stations. They do. The percents should add to 100%. In fact, they add to 99.9%. 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.



Pie chart  
Bar graph

## Bar Graphs and Pie Charts

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. Figure 1.1 illustrates both displays for the distribution of radio stations by format.

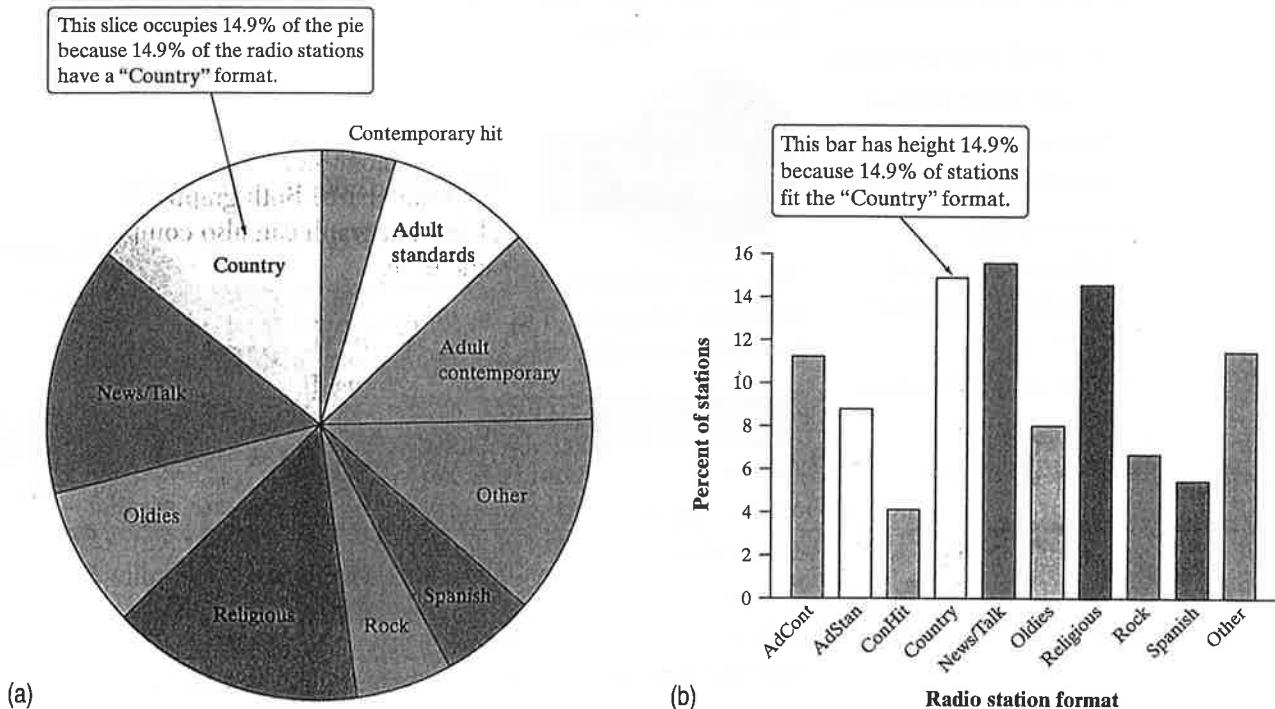
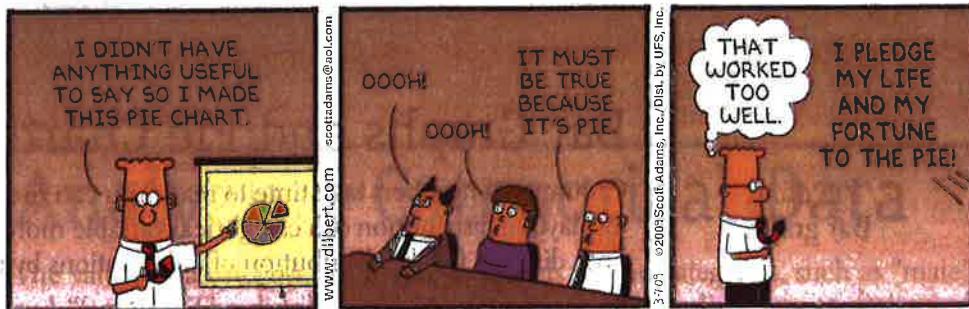


FIGURE 1.1 (a) Pie chart and (b) bar graph of U.S. radio stations by format.

THINK  
ABOUT  
IT

Do the data tell you what you want to know? Let's say that you plan to buy radio time to advertise your Web site for downloading MP3 music files. How helpful are the data in Figure 1.1? Not very. You are not interested in counting *stations*, but in counting *listeners*. For example, 14.6% of all stations are religious, but they have only a 5.5% share of the radio audience, according to Arbitron. In fact, you aren't even interested in the entire radio audience, because MP3 users are mostly young people. You really want to know what kinds of radio stations reach the largest numbers of young people. *Always think about whether the data you have help answer your questions.*

Pie charts show the distribution of a categorical variable as a “pie” whose slices are sized by the counts or percents for the categories. A pie chart must include all the categories that make up a whole. In the radio station example, we needed the “Other formats” category to complete the whole (all radio stations) and allow us to make a pie chart. Use a pie chart only when you want to emphasize each category’s relation to the whole. Pie charts are awkward to make by hand, but technology will do the job for you.



**Bar graphs are also called bar charts.**

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 are also easier to read. To convince yourself, try to use the pie chart in Figure 1.1 to estimate the percent of radio stations that have an “Oldies” format. Now look at the bar graph—it’s easy to see that the answer is about 8%.

Bar graphs are also 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.

## EXAMPLE

### *Who Owns an MP3 Player?*

#### Choosing the best graph to display the data

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player, according to an Arbitron survey of 1112 randomly selected people.<sup>1</sup>

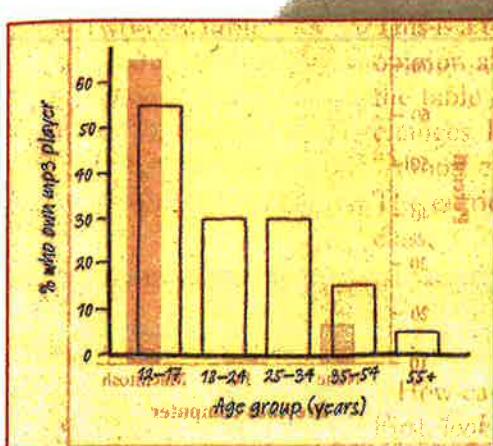
Age group (years)	Percent owning an MP3 player
12 to 17	54
18 to 24	30
25 to 34	30
35 to 54	13
55 and older	5

#### PROBLEM:

- Make a well-labeled bar graph to display the data. Describe what you see.
- Would it be appropriate to make a pie chart for these data? Why or why not?

#### SOLUTION:

- We start by labeling the axes: age group goes on the horizontal axis, and percent who own an MP3 player goes on the vertical axis. For the vertical scale, which is measured in percents, we’ll start at 0



**FIGURE 1.2** Bar graph comparing the percents of several age groups who own portable MP3 players.

and go up to 60, with tick marks for every 10. Then for each age category, we draw a bar with height corresponding to the percent of survey respondents who said they have an MP3 player. Figure 1.2 shows the completed bar graph. It appears that MP3 players are more popular among young people and that their popularity generally decreases as the age category increases.

- (b) Making a pie chart to display these data is not appropriate because each percent in the table refers to a different age group, not to parts of a single whole.

### For Practice Try Exercise 15

## Graphs: Good and Bad

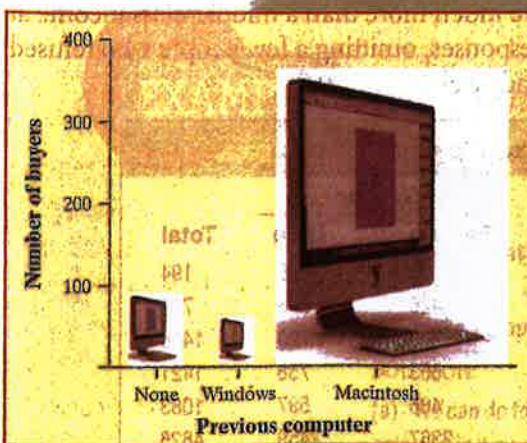
Bar graphs compare several quantities by comparing the heights of bars that represent the quantities. Our eyes, however, react to the *area* of the bars as well as to their height. When all bars have the same width, the area (width  $\times$  height) varies in proportion to the height, and our eyes receive the right impression. When you draw a bar graph, make the bars equally wide. Artistically speaking, bar graphs are a bit dull. It is tempting to replace the bars with pictures for greater eye appeal. Don't do it! The following example shows why.

### EXAMPLE

#### Who Buys iMacs?

##### Beware the pictograph!

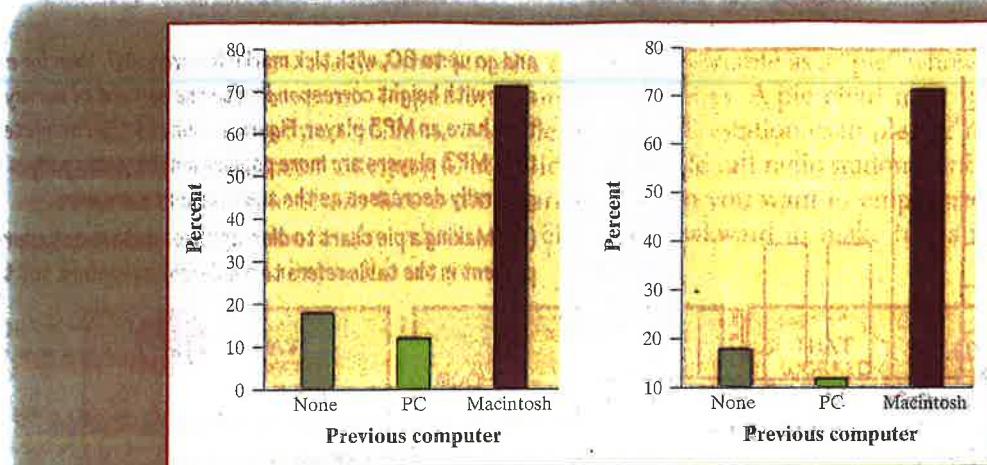
When Apple, Inc., introduced the iMac, the company wanted to know whether this new computer was expanding Apple's market share. Was the iMac mainly being bought by previous Macintosh owners, or was it being purchased by first-time computer buyers and by previous PC users who were switching over? To find out, Apple hired a firm to conduct a survey of 500 iMac customers. Each customer was categorized as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey results.<sup>5</sup>



Previous ownership	Count	Percent
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	500	100.0

##### PROBLEM

- (a) Here's a clever graph of the data that uses pictures instead of the more traditional bars. How is this graph misleading?
- (b) Two possible bar graphs of the data are shown on the next page. Which one could be considered deceptive? Why?

**SOLUTION:**

(a) Although the heights of the pictures are accurate, our eyes respond to the area of the pictures. The pictograph makes it seem like the percent of iMac buyers who are former Mac owners is at least ten times higher than either of the other two categories, which isn't the case.

(b) The bar graph on the right is misleading. By starting the vertical scale at 10 instead of 0, it looks like the percent of iMac buyers who previously owned a PC is less than half the percent who are first-time computer buyers. We get a distorted impression of the relative percents in the three categories.

For Practice Try Exercise 17

There are two important lessons to be learned from this example: (1) beware the pictograph, and (2) watch those scales.

## Two-Way Tables and Marginal Distributions

We have learned some techniques for analyzing the distribution of a single categorical variable. What do we do when a data set involves two categorical variables? We begin by examining the counts or percents in various categories for one of the variables. Here's an example to show what we mean.

### EXAMPLE

#### I'm Gonna Be Rich!

#### Relationship between two categorical variables

A survey of 4826 randomly selected young adults (aged 19 to 25) asked, "What do you think are the chances you will have much more than a middle-class income at age 30?" The table below shows the responses, omitting a few people who refused to respond or who said they were already rich.<sup>6</sup>

Opinion	Gender		
	Female	Male	Total
Almost no chance	96	98	194
Some chance but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
<b>Total</b>	<b>2367</b>	<b>2459</b>	<b>4826</b>



Two-way table

This is a **two-way table** because it describes two categorical variables, gender and opinion about becoming rich. Opinion is the *row variable* because each row in the table describes young adults who held one of the five opinions about their chances. Because the opinions have a natural order from "Almost no chance" to "Almost certain," the rows are also in this order. Gender is the *column variable*. The entries in the table are the counts of individuals in each opinion-by-gender class.

How can we best grasp the information contained in the two-way table above? First, look at the distribution of each variable separately. The distribution of a categorical variable says how often each outcome occurred. The "Total" column at the right of the table contains the totals for each of the rows. These row totals give the distribution of opinions about becoming rich in the entire group of 4826 young adults: 194 felt that they had almost no chance, 712 thought they had just some chance, and so on. (If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them.) The distributions of opinion alone and gender alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

### DEFINITION: Marginal distribution

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

Percents are often more informative than counts, especially when we are comparing groups of different sizes. We can display the marginal distribution of opinions in percents by dividing each row total by the table total and converting to a percent. For instance, the percent of these young adults who think they are almost certain to be rich by age 30 is

$$\frac{\text{almost certain total}}{\text{table total}} = \frac{1083}{4826} = 0.224 = 22.4\%$$

### EXAMPLE

#### *I'm Gonna Be Rich!*

#### Examining a marginal distribution

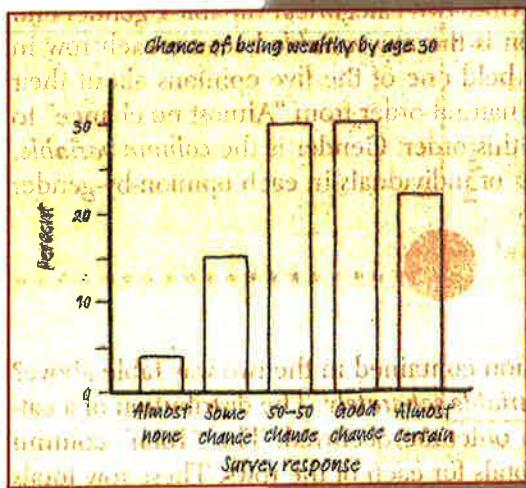
##### PROBLEM

- (a) Use the data in the two-way table to calculate the marginal distribution (in percents) of opinions.

- (b) Make a graph to display the marginal distribution. Describe what you see.

##### SOLUTION

- (a) We can do four more calculations like the one shown above to obtain the marginal distribution of opinions in percents. Here is the complete distribution.



**FIGURE 1.3** Bar graph showing the marginal distribution of opinion about chance of being rich by age 30.

Response	Percent
Almost no chance	$\frac{194}{4826} = 4.0\%$
Some chance	$\frac{712}{4826} = 14.8\%$
A 50-50 chance	$\frac{1416}{4826} = 29.3\%$
A good chance	$\frac{1421}{4826} = 29.4\%$
Almost certain	$\frac{1083}{4826} = 22.4\%$

(b) Figure 1.3 is a bar graph of the distribution of opinion among these young adults. It seems that many young adults are optimistic about their future income. Over 50% of those who responded to the survey felt that they had "a good chance" or were "almost certain" to be rich by age 30.

#### For Practice Try Exercise 19

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw earlier, we can use a bar graph or a pie chart to display such a distribution.

#### CHECK YOUR UNDERSTANDING

- 1. Use the data in the two-way table on page 12 to calculate the marginal distribution (in percents) of gender.
- 2. Make a graph to display the marginal distribution. Describe what you see.

## Relationships between Categorical Variables: Conditional Distributions

The two-way table contains much more information than the two marginal distributions of opinion alone and gender alone. *Marginal distributions tell us nothing about the relationship between two variables.* To describe a relationship between two categorical variables, we must calculate some well-chosen percents from the counts given in the body of the table.

Young adults by gender and chance of getting rich			
Opinion	Gender		Total
	Female	Male	
Almost no chance	96	98	194
Some chance but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Conditional distribution of opinion among women	
Response	Female
Almost no chance	$\frac{96}{2367} = 4.1\%$
Some chance	$\frac{426}{2367} = 18.0\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$
A good chance	$\frac{663}{2367} = 28.0\%$
Almost certain	$\frac{486}{2367} = 20.5\%$

We can study the opinions of women alone by looking only at the “Female” column in the two-way table. To find the percent of young women who think they are almost certain to be rich by age 30, divide the count of such women by the total number of women, the column total:

$$\frac{\text{women who are almost certain}}{\text{column total}} = \frac{486}{2367} = 0.205 = 20.5\%$$

Doing this for all five entries in the “Female” column gives the **conditional distribution** of opinion among women. See the table in the margin. We use the term “conditional” because this distribution describes only young adults who satisfy the condition that they are female.

### DEFINITION: Conditional distribution

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable. There is a separate conditional distribution for each value of the other variable.

Now let's examine the men's opinions.

### EXAMPLE

#### I'm Gonna Be Rich!

#### Calculating a conditional distribution

**PROBLEM** Calculate the conditional distribution of opinion among the men.

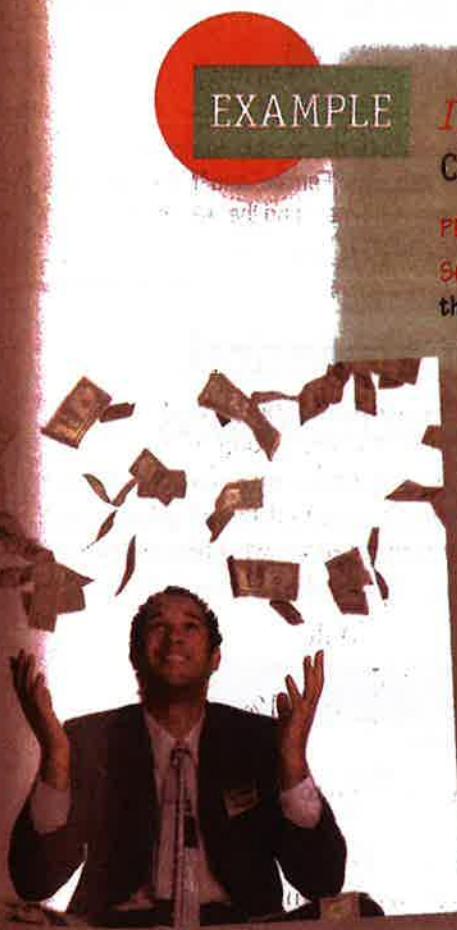
**SOLUTION** To find the percent of men who think they are almost certain to be rich by age 30, divide the count of such men by the total number of men, the column total:

$$\frac{\text{men who are almost certain}}{\text{column total}} = \frac{597}{2459} = 24.3\%$$

If we do this for all five entries in the “Male” column, we get the conditional distribution shown in the table.

### Conditional distribution of opinion among men

Response	Male
Almost no chance	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{597}{2459} = 24.3\%$



step process you can follow

to solve it.

For Practice Try Exercise 21

Software will calculate conditional distributions for you. Most programs allow you to choose which conditional distributions you want to compute.

### TECHNOLOGY CORNER Analyzing two-way tables

Figure 1.4 presents the two conditional distributions of opinion, for women and for men, and also the marginal distribution of opinion for all of the young adults. The distributions agree (up to rounding) with the results in the last two examples.

	Female	Male	All
A: Almost no chance	96 4.06	98 3.99	194 4.02
B: Some chance but probably not	626 18.00	286 11.63	712 14.75
C: A 50-50 chance	696 29.40	720 29.28	1416 29.34
D: A good chance	663 28.01	758 30.83	1421 29.44
E: Almost certain	486 20.53	597 24.28	1083 22.44
All	2367 100.00	2459 100.00	4826 100.00
Cell Contents:	Count % of Column		

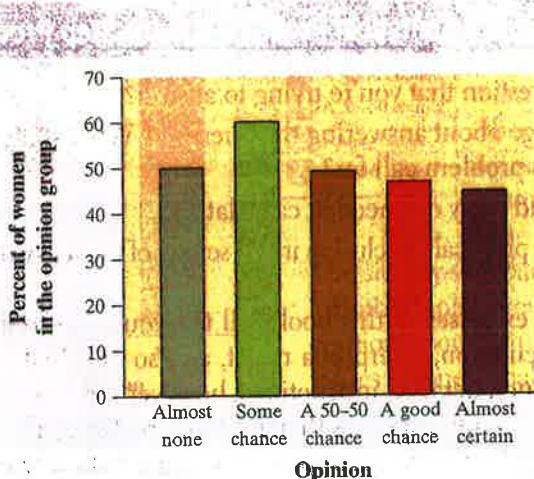
FIGURE 1.4 Minitab output for the two-way table of young adults by gender and chance of being rich, along with each entry as a percent of its column total. The “Female” and “Male” columns give the conditional distributions of opinion for women and men, and the “All” column shows the marginal distribution of opinion for all these young adults.

There are *two sets* of conditional distributions for any two-way table. So far, we have looked at the conditional distributions of opinion for the two genders. We could also examine the five conditional distributions of gender, one for each of the five opinions, by looking separately at the rows in the original two-way table. For instance, the conditional distribution of gender among those who responded “Almost certain” is

Female	Male
$\frac{486}{1083} = 44.9\%$	$\frac{597}{1083} = 55.1\%$

That is, of the young adults who said they were almost certain to be rich by age 30, 44.9% were female and 55.1% were male.

Because the variable “gender” has only two categories, comparing the five conditional distributions amounts to comparing the percents of women among young adults who hold each opinion. Figure 1.5 makes this comparison in a bar graph. The bar heights do *not* add to 100%, because each bar represents a different group of people.



**FIGURE 1.5** Bar graph comparing the percents of females among those who hold each opinion about their chance of being rich by age 30.

Segmented bar graph



An alternative to the bar graph in Figure 1.5 is a segmented bar graph, like the one shown in Figure 1.6. For each opinion category, there is a single bar with “segments” that correspond to the different genders. The height of each segment is determined by the percent of young adults having that opinion who were of each gender. We can see the two percents we calculated earlier displayed in the “Almost certain” bar—female 44.9% and male 55.1%. Notice that each bar has a total height of 100%.

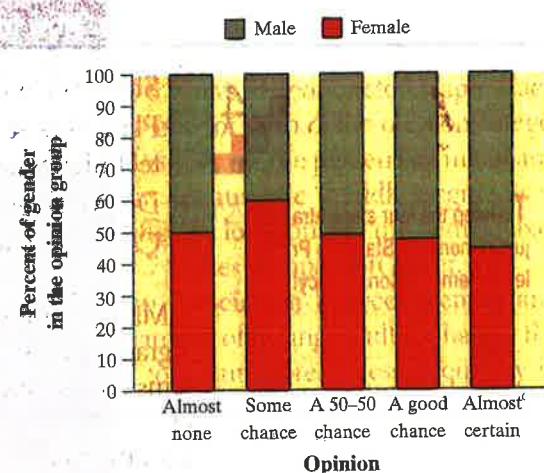
**Which conditional distributions should we compare?** Our goal all along has been to analyze the relationship between gender and opinion about chances of becoming rich for these young adults. We started by examining the conditional distributions of opinion for males and females. Then we looked at the conditional distributions of gender for each of the five opinion categories. Which of these two gives us the information we want? Here’s a hint: think about whether changes in one variable might help explain changes in the other. In this case, it seems reasonable to think that gender might influence young adults’ opinions about their chances of getting rich. To see whether the data support this idea, we should compare the conditional distributions of opinion for women and men.

### CHECK YOUR UNDERSTANDING

- Find the conditional distributions of gender among each of the other four opinion categories (we did “Almost certain” earlier). Use Figure 1.5 or Figure 1.6 to check that your answers are approximately correct.
- Make a revised version of Figure 1.4 that includes your results from Question 1.

## Organizing a Statistical Problem

As you learn more about statistics, you will be asked to solve more complex problems. Although no single strategy will work on every problem, it might be helpful to have a general framework for organizing your thinking. Here is a four-step process you can follow.



**FIGURE 1.6** Segmented bar graph showing the conditional distribution of gender for each opinion category.

### How to Organize a Statistical Problem: A Four-Step Process



To keep the four steps straight, just remember: Statistics Problems Demand Consistency!

**State:** What's the question that you're trying to answer?

**Plan:** How will you go about answering the question? What statistical techniques does this problem call for?

**Do:** Make graphs and carry out needed calculations.

**Conclude:** Give your practical conclusion in the setting of the real-world problem.

Many examples and exercises in this book will tell you what to do—construct a graph, perform a calculation, interpret a result, and so on. Real statistics problems don't come with such detailed instructions, however. From now on, you will encounter some examples and exercises that are more realistic. They are marked with the four-step icon. Use the four-step process as a guide to solving these problems, as the following example illustrates.

### EXAMPLE



#### Women's and Men's Opinions

##### Conditional distributions and relationships

Based on the survey data, can we conclude that young men and women differ in their opinions about the likelihood of future wealth? Give appropriate evidence to support your answer. Follow the four-step process.

**STATE:** What is the relationship between gender and responses to the question "What do you think are the chances you will have much more than a middle-class income at age 30?"

**PLAN:** We suspect that gender might influence a young adult's opinion about the chance of getting rich. So we'll compare the conditional distributions of response for men alone and for women alone.

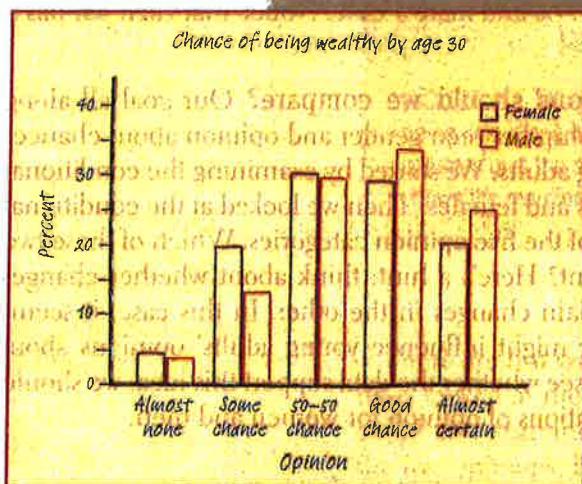


FIGURE 1.7 Side-by-side bar graph comparing the opinions of males and females.

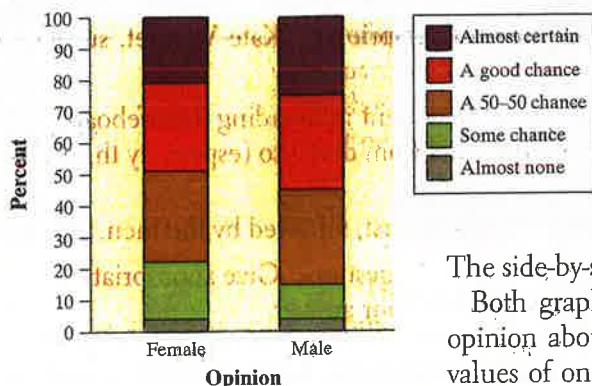
Response	Female	Male
Almost no chance	$\frac{96}{2367} = 4.1\%$	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{426}{2367} = 18.0\%$	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{663}{2367} = 28.0\%$	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{486}{2367} = 20.5\%$	$\frac{597}{2459} = 24.3\%$

Side-by-side bar graph

**DO:** We'll make a side-by-side bar graph to compare the opinions of males and females. Figure 1.7 displays the completed graph.

**CONCLUDE:** Based on the sample data, men seem somewhat more optimistic about their future income than women. Men were less likely to say that they have "some chance but probably not" than women (11.6% vs. 18.0%). Men were more likely to say that they have "a good chance" (30.8% vs. 28.0%) or are "almost certain" (24.3% vs. 20.5%) to have much more than a middle-class income by age 30 than women were.

For Practice Try Exercise 25



**FIGURE 1.8** Segmented bar graph comparing the opinions of males and females.

We could have used a segmented bar graph to compare the distributions of male and female responses in the previous example. Figure 1.8 shows the completed graph. Each bar has five segments—one for each of the opinion categories. It's fairly difficult to compare the percents of males and females in each category because the “middle” segments in the two bars start at different locations on the vertical axis.

The side-by-side bar graph in Figure 1.7 makes comparison easier.

Both graphs provide evidence of an **association** between gender and opinion about future wealth in this sample of young adults. That is, the values of one variable (opinion) tend to occur more or less frequently in combination with specific values of the other variable (gender). Men more often rated their chances of becoming rich in the two highest categories; women said “some chance but probably not” much more frequently. Can we say that there is an association between gender and opinion in the *population* of young adults? Making this determination requires formal inference, which will have to wait a few chapters.

#### DEFINITION: Association

We say that there is an **association** between two variables if specific values of one variable tend to occur in common with specific values of the other.

There's one caution that we need to offer: *even a strong association between two categorical variables can be influenced by other variables lurking in the background*. The Data Exploration that follows gives you a chance to explore this idea using a famous (or infamous) data set.



## DATA EXPLORATION A Titanic disaster

In 1912 the luxury liner *Titanic*, on its first voyage across the Atlantic, struck an iceberg and sank. Some passengers got off the ship in lifeboats, but many died. The two-way table below gives information about adult passengers who lived and who died, by class of travel.

Class of Travel	Survival Status	
	Survived	Died
First class	197	122
Second class	94	167
Third class	151	476

Here's another table that displays data on survival status by gender and class of travel.

Class of Travel	Gender		Survived	Died
	Female	Male		
First class	140	4	57	118
Second class	80	13	14	154
Third class	76	89	75	387



The movie *Titanic*, starring Leonardo DiCaprio and Kate Winslet, suggested the following:

- First-class passengers received special treatment in boarding the lifeboats, while some other passengers were prevented from doing so (especially third-class passengers).
  - Women and children boarded the lifeboats first, followed by the men.
1. What do the data tell us about these two suggestions? Give appropriate graphical and numerical evidence to support your answer.
  2. How does gender affect the relationship between class of travel and survival status? Explain.

## \* Not on the AP Test **Simpson's Paradox\***

In the most extreme cases, it is possible for an association between two categorical variables to be “reversed” when we consider a third variable. Here is an example that demonstrates the surprises that can await the unsuspecting user of data.

### EXAMPLE

#### *Do Medical Helicopters Save Lives?*

##### Reversing an association

Accident victims are sometimes taken by helicopter from the accident scene to a hospital. Helicopters save time. Do they also save lives? Let's compare the percents of accident victims who die with helicopter evacuation and with the usual transport to a hospital by road. Here are hypothetical data that illustrate a practical difficulty:<sup>\*</sup>

	Helicopter	Road
Victim died	64	260
Victim survived	136	840
Total	200	1100

We see that 32% (64 out of 200) of helicopter patients died, but only 24% (260 out of 1100) of the others did. That seems discouraging.

The explanation is that the helicopter is sent mostly to serious accidents, so that the victims transported by helicopter are more often seriously injured. They are more likely to die with or without helicopter evacuation. Here are the same data broken down by the seriousness of the accident:

Serious Accidents		
	Helicopter	Road
Died	48	60
Survived	52	40
Total	100	100

Less Serious Accidents		
	Helicopter	Road
Died	16	200
Survived	84	800
Total	100	1000

Inspect these tables to convince yourself that they describe the same 1300 accident victims as the original two-way table. For example, 200 (100 + 100) were moved by helicopter, and 64 (48 + 16) of these died.

\*This is an interesting topic, but it is not required for the AP Statistics exam.

Among victims of serious accidents, the helicopter saves 52% (52 out of 100) compared with 40% for road transport. If we look only at less serious accidents, 84% of those transported by helicopter survive, versus 80% of those transported by road. Both groups of victims have a higher survival rate when evacuated by helicopter.

How can it happen that the helicopter does better for both groups of victims but worse when all victims are lumped together? Examining the data makes the explanation clear. Half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients. So the helicopter carries patients who are more likely to die. The seriousness of the accident was a “lurking variable” that, until we uncovered it, hid the true relationship between survival and mode of transport to a hospital. This example illustrates Simpson’s paradox.

#### DEFINITION: Simpson's paradox

An association between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This reversal is called Simpson's paradox.

## SECTION 1.1

# Summary

- The distribution of a categorical variable lists the categories and gives the count (**frequency table**) or percent (**relative frequency table**) of individuals that fall in each category.
- 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. When examining any graph, ask yourself, “What do I see?”
- A **two-way table** of counts organizes data about two categorical variables. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.
- The row totals and column totals in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- There are two sets of **conditional distributions** for a two-way table: the distributions of the row variable for each value of the column variable, and the distributions of the column variable for each value of the row variable. You may want to use a **side-by-side bar graph** (or possibly a **segmented bar graph**) to display conditional distributions.



- A statistical problem has a real-world setting. You can organize many problems using the four steps state, plan, do, and conclude.
- To describe the **association** between the row and column variables, compare an appropriate set of conditional distributions. Remember that even a strong association between two categorical variables can be influenced by other variables lurking in the background.
- An association between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is **Simpson's paradox**.

**1.1 TECHNOLOGY CORNER**

Analyzing two-way tables ..... page 16

## SECTION 1.1 Exercises

9. Cool car colors The most popular colors for cars and light trucks change over time. Silver passed green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. Here is the distribution of colors for vehicles sold in North America in 2008.<sup>8</sup>

Color	Percent of vehicles
White	20
Black	17
Silver	17
Blue	13
Gray	12
Red	11
Beige/brown	5
Green	3
Yellow/gold	2

- (a) What percent of vehicles had colors other than those listed?
- (b) Display these data in a bar graph. Be sure to label your axes and title your graph.
- (c) Would it be appropriate to make a pie chart of these data? Explain.

10. Spam Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:<sup>9</sup>

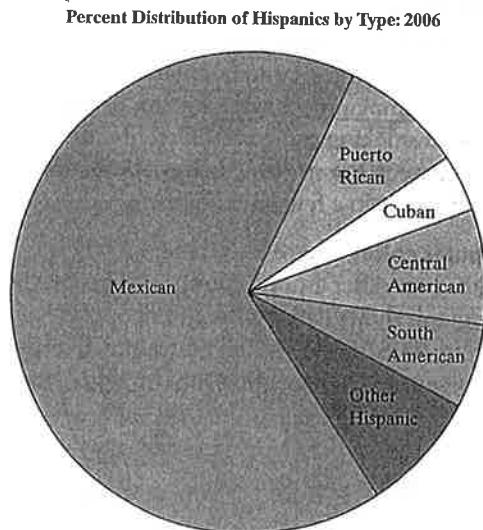
Type of spam	Percent
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

- (a) What percent of spam would fall in the "Other" category?
- (b) Display these data in a bar graph. Be sure to label your axes and title your graph.
- (c) Would it be appropriate to make a pie chart of these data? Explain.
11. Birth days Births are not evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in the United States in a recent year.<sup>10</sup>

Day	Births
Sunday	7,374
Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8,459

- (a) Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart?
- (b) Suggest some possible reasons why there are fewer births on weekends.
12. Deaths among young people Among persons aged 15 to 24 years in the United States, the leading causes of death and number of deaths in a recent year were as follows: accidents, 15,567; homicide, 5359; suicide, 4139; cancer, 1717; heart disease, 1067; congenital defects, 483.<sup>11</sup>
- (a) Make a bar graph to display these data.
- (b) To make a pie chart, you need one additional piece of information. What is it?

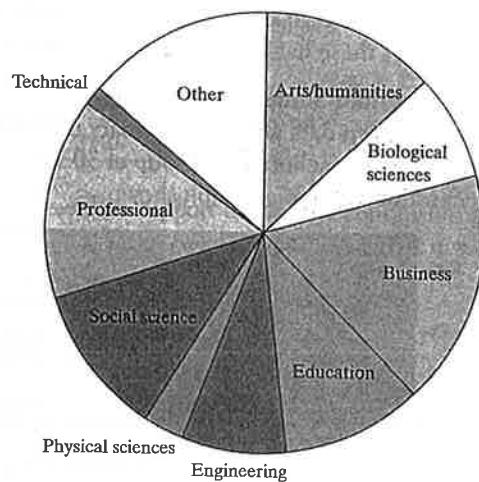
13. Hispanic origins Below is a pie chart prepared by the Census Bureau to show the origin of the more than 43 million Hispanics in the United States in 2006.<sup>12</sup> About what percent of Hispanics are Mexican? Puerto Rican?



*Comment:* You see that it is hard to determine numbers from a pie chart. Bar graphs are much easier to use. (The Census Bureau did include the percents in its pie chart.)

14. Which major? About 1.6 million first-year students enroll in colleges and universities each year. What do they plan to study? The pie chart displays data on the percents of first-year students who plan to major

in several discipline areas.<sup>13</sup> About what percent of first-year students plan to major in business? In social science?



- pg 10  
15. Buying music online Young people are more likely than older folk to buy music online. Here are the percents of people in several age groups who bought music online in 2006.<sup>14</sup>

Age group	Bought music online
12 to 17 years	24%
18 to 24 years	21%
25 to 34 years	20%
35 to 44 years	16%
45 to 54 years	10%
55 to 64 years	3%
65 years and over	1%

- (a) Explain why it is *not* correct to use a pie chart to display these data.
- (b) Make a bar graph of the data. Be sure to label your axes and title your graph.

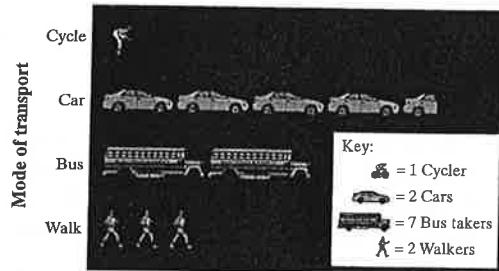
16. The audience for movies Here are data on the percent of people in several age groups who attended a movie in the past 12 months.<sup>15</sup>

Age group	Movie attendance
18 to 24 years	83%
25 to 34 years	73%
35 to 44 years	68%
45 to 54 years	60%
55 to 64 years	47%
65 to 74 years	32%
75 years and over	20%

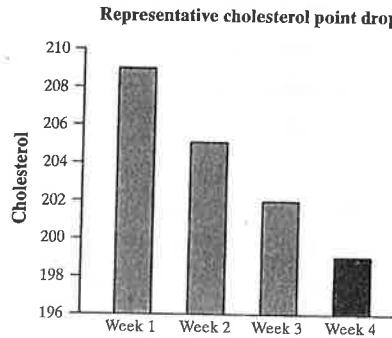
- (a) Display these data in a bar graph. Describe what you see.

- (b) Would it be correct to make a pie chart of these data? Why or why not?  
 (c) A movie studio wants to know what percent of the total audience for movies is 18 to 24 years old. Explain why these data do not answer this question.

- 17. Going to school** Students in a high school statistics class were given data about the primary method of transportation to school for a group of 30 students. They produced the pictograph shown.



- pg 11** (a) How is this graph misleading?  
 (b) Make a new graph that isn't misleading.  
**18. Oatmeal and cholesterol** Does eating oatmeal reduce cholesterol? An advertisement included the following graph as evidence that the answer is "Yes."



- (a) How is this graph misleading?  
 (b) Make a new graph that isn't misleading. What do you conclude about the effect of eating oatmeal on cholesterol reduction?

- pg 13** **19. Attitudes toward recycled products** Recycling is supposed to save resources. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. People who actually use a recycled product may have different opinions from those who don't use it. Here are data on attitudes toward coffee filters made of recycled paper among people who do and don't buy these filters:<sup>16</sup>

Think the quality of the recycled product is:

	Higher	The same	Lower
Buyers	20	7	9
Nonbuyers	29	25	43

- (a) How many people does this table describe? How many of these were buyers of coffee filters made of recycled paper?

- (b) Give the marginal distribution of opinion about the quality of recycled filters. What percent think the quality of the recycled product is the same or higher than the quality of other filters?  
**20. Smoking by students and parents** Here are data from a survey conducted at eight high schools on smoking among students and their parents.<sup>17</sup>

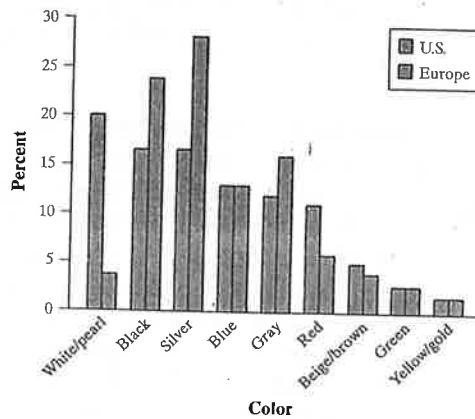
	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

- (a) How many students are described in the two-way table? What percent of these students smoke?  
 (b) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.

- pg 15** **21. Attitudes toward recycled products** Exercise 19 gives data on the opinions of people who have and have not bought coffee filters made from recycled paper. To see the relationship between opinion and experience with the product, find the conditional distributions of opinion (the response variable) for buyers and nonbuyers. What do you conclude?

- 22. Smoking by students and parents** Refer to Exercise 20. Calculate three conditional distributions of students' smoking behavior: one for each of the three parental smoking categories. Describe the relationship between the smoking behaviors of students and their parents in a few sentences.

- 23. Popular colors—here and there** Favorite vehicle colors may differ among countries. The side-by-side bar graph shows data on the most popular colors of cars in 2008 for the United States and Europe. Write a few sentences comparing the two distributions.



24. Comparing car colors Favorite vehicle colors may differ among types of vehicle. Here are data on the most popular colors in 2008 for luxury cars and for SUVs, trucks, and vans.

Color	Luxury cars (%)	SUVs, trucks, vans (%)
Black	22	13
Silver	16	16
White pearl	14	1
Gray	12	13
White	11	25
Blue	7	10
Red	7	11
Yellow/gold	6	1
Green	3	4
Beige/brown	2	6

- (a) Make a graph to compare colors by vehicle type.  
 (b) Write a few sentences describing what you see.

25. Snowmobiles in the park Yellowstone National Park surveyed a random sample of 1526 winter visitors to the park. They asked each person whether they owned, rented, or had never used a snowmobile. Respondents were also asked whether they belonged to an environmental organization (like the Sierra Club). The two-way table summarizes the survey responses.

Environmental Clubs			
	No	Yes	Total
Never used	445	212	657
Snowmobile renter	497	77	574
Snowmobile owner	279	16	295
<b>Total</b>	<b>1221</b>	<b>305</b>	<b>1526</b>

Do these data provide convincing evidence of an association between environmental club membership and snowmobile use for the population of visitors to Yellowstone National Park? Follow the four-step process.

26. Angry people and heart disease People who get angry easily tend to have more heart disease. That's the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure. CHD stands for "coronary heart disease."

This includes people who had heart attacks and those who needed medical treatment for heart disease.

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	190
No CHD	3057	4621	606	8284
<b>Total</b>	<b>3110</b>	<b>4731</b>	<b>633</b>	<b>8474</b>

Do these data support the study's conclusion about the relationship between anger and heart disease? Follow the four-step process.

**Multiple choice: Select the best answer.**

Exercises 27 to 32 refer to the following setting. The National Survey of Adolescent Health interviewed several thousand teens (grades 7 to 12). One question asked was "What do you think are the chances you will be married in the next ten years?" Here is a two-way table of the responses by gender:<sup>18</sup>

	Female	Male
Almost no chance	119	103
Some chance, but probably not	150	171
A 50-50 chance	447	512
A good chance	735	710
Almost certain	1174	756

27. The percent of females among the respondents was  
 (a) 2625. (c) about 46%. (e) None of these.  
 (b) 4877. (d) about 54%.
28. Your percent from the previous exercise is part of  
 (a) the marginal distribution of females.  
 (b) the marginal distribution of gender.  
 (c) the marginal distribution of opinion about marriage.  
 (d) the conditional distribution of gender among adolescents with a given opinion.  
 (e) the conditional distribution of opinion among adolescents of a given gender.
29. What percent of females thought that they were almost certain to be married in the next ten years?  
 (a) About 16% (c) About 40% (e) About 61%  
 (b) About 24% (d) About 45%
30. Your percent from the previous exercise is part of  
 (a) the marginal distribution of gender.  
 (b) the marginal distribution of opinion about marriage.  
 (c) the conditional distribution of gender among adolescents with a given opinion.

pg 18

STEP 4

STEP 4

## 26 CHAPTER 1 EXPLORING DATA

- (d) the conditional distribution of opinion among adolescents of a given gender.  
 (e) the conditional distribution of “Almost certain” among females.
31. What percent of those who thought they were almost certain to be married were female?  
 (a) About 16% (c) About 40% (e) About 61%  
 (b) About 24% (d) About 45%
32. Your percent from the previous exercise is part of  
 (a) the marginal distribution of gender.  
 (b) the marginal distribution of opinion about marriage.  
 (c) the conditional distribution of gender among adolescents with a given opinion.  
 (d) the conditional distribution of opinion among adolescents of a given gender.  
 (e) the conditional distribution of females among those who said “Almost certain.”
33. Marginal distributions aren’t the whole story Here are the row and column totals for a two-way table with two rows and two columns:

<i>a</i>	<i>b</i>	50
<i>c</i>	<i>d</i>	50
60	40	100

Find two different sets of counts *a*, *b*, *c*, and *d* for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

- 34.\* Baseball paradox Most baseball hitters perform differently against right-handed and left-handed pitching. Consider two players, Joe and Moe, both of whom bat right-handed. The table below records their performance against right-handed and left-handed pitchers:

Player	Pitcher	Hits	At-bats
Joe	Right	40	100
	Left	80	400
Moe	Right	120	400
	Left	10	100

- (a) Use these data to make a two-way table of player (Joe or Moe) versus outcome (hit or no hit).

- (b) Show that Simpson’s paradox holds: one player has a higher overall batting average, but the other player hits better against both left-handed and right-handed pitching.

- (c) The manager doesn’t believe that one player can hit better against both left-handers and right-handers yet have a lower overall batting average. Explain in simple language why this happens to Joe and Moe.

- 35.\* Race and the death penalty Whether a convicted murderer gets the death penalty seems to be influenced by the race of the victim. Here are data on 326 cases in which the defendant was convicted of murder:<sup>19</sup>

White Defendant		Black Defendant			
White victim	Black victim	White victim	Black victim		
Death	19	0	Death	11	6
Not	132	9	Not	52	97

- (a) Use these data to make a two-way table of defendant’s race (white or black) versus death penalty (yes or no).  
 (b) Show that Simpson’s paradox holds: a higher percent of white defendants are sentenced to death overall, but for both black and white victims a higher percent of black defendants are sentenced to death.  
 (c) Use the data to explain why the paradox holds in language that a judge could understand..

36. Fuel economy (Introduction) Here is a small part of a data set that describes the fuel economy (in miles per gallon) of model year 2009 motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City mpg	Highway mpg
Aston Martin Vantage	Two-seater	Manual	8	12	19
Honda Civic	Subcompact	Automatic	4	25	36
Toyota Prius	Midsize	Automatic	4	48	45
Chevrolet Impala	Large	Automatic	6	18	29

- (a) What are the individuals in this data set?  
 (b) What variables were measured? Identify each as categorical or quantitative.

\*These exercises relate to the optional content on Simpson’s paradox.

## 1.2

In Section 1.2,  
you'll learn about:

- Dotplots
- Describing shape
- Comparing distributions
- Stemplots
- Histograms
- Using histograms wisely

Dotplot

## EXAMPLE

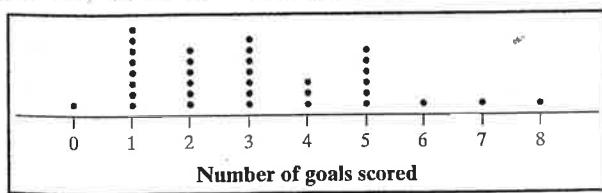


FIGURE 1.9 A dotplot of goals scored by the U.S. women's soccer team in 2004.

# Displaying Quantitative Data with Graphs

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

## Dotplots

One of the simplest graphs to construct and interpret is a **dotplot**. Each data value is shown as a dot above its location on a number line. We'll show how to make a dotplot using some sports data.

### Gooooaaaalllll!

#### How to make a dotplot

How good was the 2004 U.S. women's soccer team? With players like Brandi Chastain, Mia Hamm, and Briana Scurry, the team put on an impressive showing en route to winning the gold medal at the 2004 Olympics in Athens. Here are data on the number of goals scored by the team in 34 games played during the 2004 season:<sup>20</sup>

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

Here are the steps in making a dotplot:

- *Draw a horizontal axis (a number line) and label it with the variable name.* In this case, the variable is number of goals scored.
- *Scale the axis.* Start by looking at the minimum and maximum values of the variable. For these data, the minimum number of goals scored was 0, and the maximum was 8. So we mark our scale from 0 to 8, with tick marks at every whole-number value.
- *Mark a dot above the location on the horizontal axis corresponding to each data value.* Figure 1.9 displays a completed dotplot for the soccer data.



Making a graph is not an end in itself. The purpose of a graph is to help us understand the data. After you make a graph, always ask, "What do I see?" Here is a general strategy for interpreting graphs of quantitative data.

### How to Examine the Distribution of a Quantitative Variable

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

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

We'll learn more formal ways of describing shape, center, and spread and identifying outliers shortly. For now, let's use our informal understanding of these ideas to examine the graph in Figure 1.9.

Mode

**Shape:** The dotplot has a peak at 1. This indicates that the team's most frequent number of goals scored in games that season (known as the **mode**) was 1. In most of its games, the U.S. women's soccer team scored between 1 and 5 goals. However, the distribution has a long tail to the right. (Later, we will describe the shape of Figure 1.9 as *skewed to the right*.)

Range

**Center:** We can describe the center by finding a value that divides the observations so that about half take larger values and about half take smaller values. This value is called the **median** of the distribution. In Figure 1.9, the median is 3. That is, in a typical game during the 2004 season, the U.S. women's soccer team scored about 3 goals. Of course, we could also summarize the center of the distribution by calculating the average (**mean**) number of goals scored per game. For the 2004 season, the team's mean was 3.06 goals.

When describing a distribution of quantitative data, don't forget your SOCS (shape, outliers, center, spread)!

**Spread:** The spread of a distribution tells us how much **variability** there is in the data. One way to describe the variability is to give the smallest and largest values. The spread in Figure 1.9 is from 0 goals to 8 goals scored. Alternatively, we can compute the **range** of the distribution by subtracting the smallest value from the largest value. For these data, the range is  $8 - 0 = 8$  goals.

**Outliers:** Was the game in which the women's team scored 8 goals an outlier? How about the team's 7-goal game? These values differ somewhat from the overall pattern. However, they don't clearly stand apart from the rest of the distribution. For now, let's agree to call attention only to potential outliers that suggest something special about an observation. In Section 1.3, we'll establish a procedure for determining whether a particular data value is an outlier.

### EXAMPLE

#### Are You Driving a Gas Guzzler? Interpreting a dotplot



The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy ratings for cars (think of those large window stickers on a new car). For years, consumers complained that their actual gas mileages were noticeably lower than the values reported by the EPA. It seems that the EPA's tests—all of which are done on computerized devices to ensure consistency—did not consider things like outdoor temperature, use of the air conditioner, or realistic accel-

eration and braking by drivers. In 2008, the EPA changed the method for measuring a vehicle's fuel economy to try to give more accurate estimates.

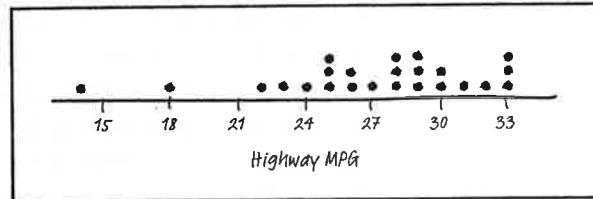
The table below displays the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 24 model year 2009 midsize cars.

Model	Mpg	Model	Mpg	Model	Mpg
Acura RL	22	Dodge Avenger	30	Mercury Milan	29
Audi A6 Quattro	23	Hyundai Elantra	33	Mitsubishi Galant	27
Bentley Arnage	14	Jaguar XF	25	Nissan Maxima	26
BMW 528i	28	Kia Optima	32	Rolls Royce Phantom	18
Buick Lacrosse	28	Lexus GS 350	26	Saturn Aura	33
Cadillac CTS	25	Lincoln MKZ	28	Toyota Camry	31
Chevrolet Malibu	33	Mazda 6	29	Volkswagen Passat	29
Chrysler Sebring	30	Mercedes-Benz E350	24	Volvo S80	25

Source: 2009 Fuel Economy Guide, from the U.S. Environmental Protection Agency's Web site at [www.fueleconomy.gov](http://www.fueleconomy.gov).



Here is a dotplot of the data:



**PROBLEM:** Describe the shape, center, and spread of the distribution. Are there any outliers?

**SOLUTION:** Don't forget your SOCS (shape, outliers, center, spread)! **Shape:** In the dotplot, we can see three clusters of values: cars that get around 25 mpg, cars that get about 28 to 30 mpg, and cars that get around 33 mpg. We can also see large gaps between the Acura RL at 22 mpg, the Rolls Royce Phantom at 18 mpg, and the Bentley Arnage at 14 mpg. **Center:** The median is 28. So a "typical" model year 2009 midsize car got about 28 miles per gallon on the highway. **Spread:** The highest value is 33 mpg and the lowest value is 14 mpg. The range is  $33 - 14 = 19$  mpg. **Outliers:** We see two midsize cars with unusually low gas mileage ratings—the Bentley Arnage (14 mpg) and the Rolls Royce Phantom (18 mpg). These cars are potential outliers.



For Practice Try Exercise 39

## Describing Shape

When you describe a distribution's shape, concentrate on the main features. Look for major peaks, not for minor ups and downs in the graph. Look for clusters of values and obvious gaps. Look for potential outliers, not just for the smallest and largest observations. Look for rough symmetry or clear skewness.

### DEFINITION: Symmetric and skewed distributions

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

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

For brevity, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed to the right." We could also describe a distribution with a long tail to the left as "skewed toward negative values" or "negatively skewed" and a distribution with a long right tail as "positively skewed."



The direction of skewness is the direction of the long tail, not the direction where most observations are clustered. See the drawing in the margin for a cute but corny way to help you keep this straight.



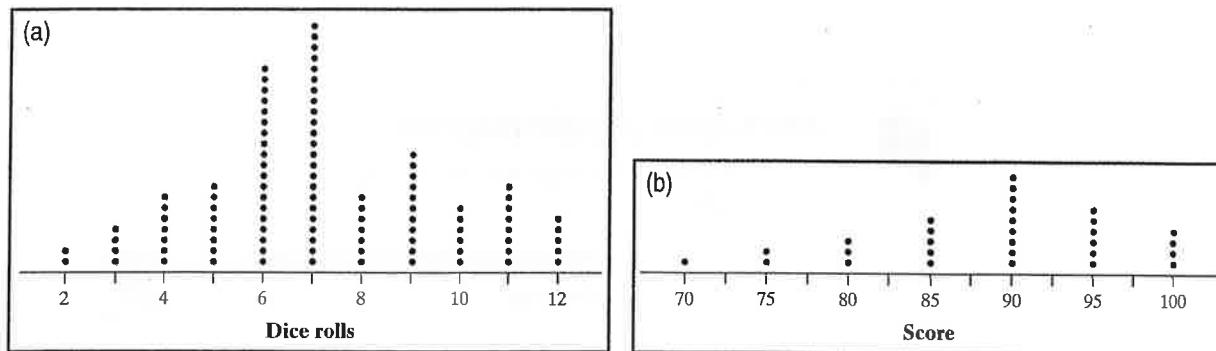
For his own safety, which way should Mr. Starnes go “skewing”?

## EXAMPLE

### Die Rolls and Quiz Scores

#### Describing shape

Figure 1.10 displays dotplots for two different sets of quantitative data. Let’s practice describing the shapes of these distributions. Figure 1.10(a) shows the results of rolling a pair of fair, six-sided dice and finding the sum of the up-faces 100 times. This distribution is roughly symmetric. The dotplot in Figure 1.10(b) shows the scores on an AP Statistics class’s first quiz. This distribution is skewed to the left.



**FIGURE 1.10** Dotplots displaying different shapes: (a) roughly symmetric; (b) skewed to the left.

Unimodal

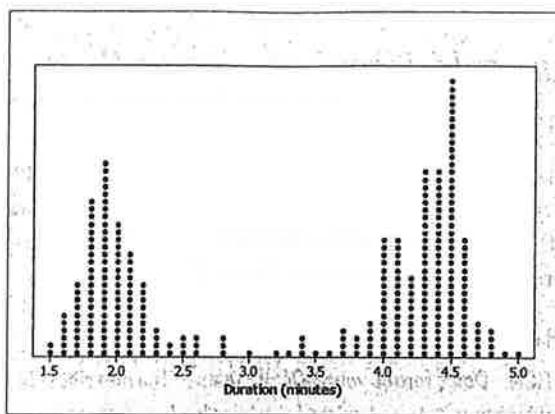
Bimodal

Multimodal

**THINK  
ABOUT  
IT**

Although the dotplots in the previous example have different shapes, they do have something in common. Both are **unimodal**, that is, they have a single peak: the graph of dice rolls at 7 and the graph of quiz scores at 90. (We don’t count minor ups and downs in a graph, like the “bumps” at 9 and 11 in the dice rolls dotplot, as “peaks.”) Figure 1.11 is a dotplot of the duration (in minutes) of 220 eruptions of the Old Faithful geyser. We would describe this distribution’s shape as **bimodal** since it has two clear peaks: one near 2 minutes and the other near 4.5 minutes. (Although we could continue the pattern with “trimodal” for three peaks and so on, it’s more common to refer to distributions with more than two clear peaks as **multimodal**.)

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



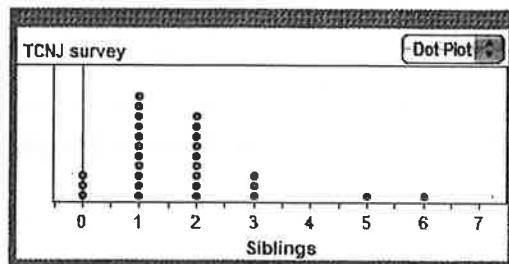
**FIGURE 1.11** Dotplot displaying duration (in minutes) of Old Faithful eruptions. This graph has a bimodal shape.

strong right-skew. Many distributions have irregular shapes that are neither symmetric nor skewed. Some data show other patterns, such as the two peaks in Figure 1.11. Use your eyes, describe the pattern you see, and then try to explain the pattern.



### CHECK YOUR UNDERSTANDING

The Fathom dotplot displays data on the number of siblings reported by each student in a statistics class.



1. Describe the shape of the distribution.
2. Describe the center of the distribution.
3. Describe the spread of the distribution.
4. Identify any potential outliers.

## Comparing Distributions

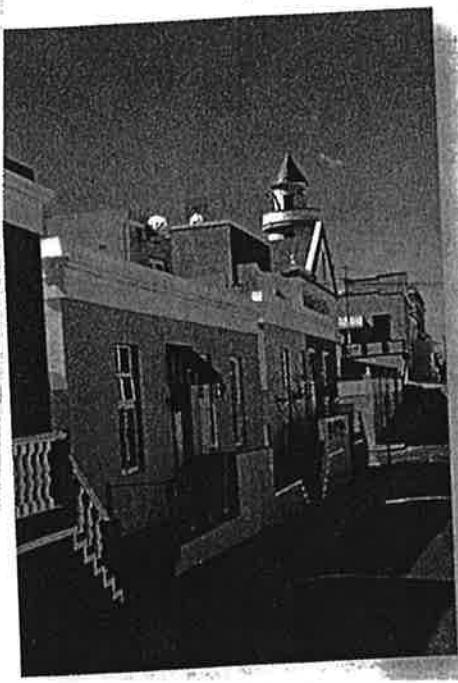
Some of the most interesting statistics questions involve comparing two or more groups. Which of two popular diets leads to greater long-term weight loss? Who texts more—males or females? Does the number of people living in a household differ among countries? As the following example suggests, you should always discuss shape, center, spread, and possible outliers whenever you compare distributions of a quantitative variable.

**EXAMPLE*****Household Size: U.K. versus South Africa*****Comparing distributions**

How do the numbers of people living in households in the United Kingdom (U.K.) and South Africa compare? To help answer this question, we used CensusAtSchool's "Random Data Selector" to choose 50 students from each country. Figure 1.12 is a dotplot of the household sizes reported by the survey respondents.

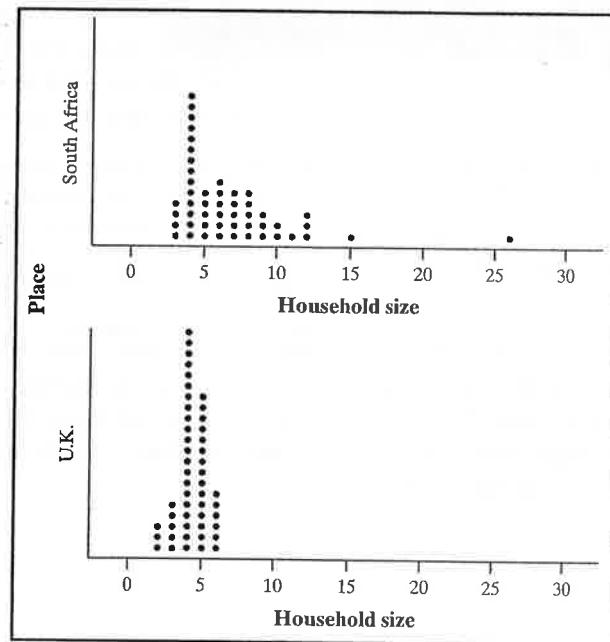
**PROBLEM:** Compare the distributions of household size for these two countries.

**SOLUTION:** Don't forget your SOCS! **Shape:** The distribution of household size for the U.K. sample is roughly symmetric and unimodal, while the distribution for the South Africa sample is skewed to the right and unimodal. **Center:** Household sizes for the South African students tended to be larger than for the U.K. students. The median household sizes for the two groups are 6 people and 4 people, respectively. **Spread:** There is more variability (greater spread) in the household sizes for the South African students than for the U.K. students. The range for the South African data is  $26 - 3 = 23$  people, while the range for the U.K. data is  $6 - 2 = 4$  people. **Outliers:** There don't appear to be any potential outliers in the U.K. distribution. The South African distribution has two potential outliers in the right tail of the distribution—students who reported living in households with 15 and 26 people. (The U.K. households with 2 people actually will be classified as outliers when we introduce a procedure in the next section.)



**AP EXAM TIP** When comparing distributions of quantitative data, it's not enough just to list values for the center and spread of each distribution. You have to explicitly *compare* these values, using words like "greater than," "less than," or "about the same as."

**FIGURE 1.12** Dotplot of household size for random samples of 50 students from the United Kingdom and South Africa.



For Practice Try Exercise 43

Notice that we discussed the distributions of household size only for the two *samples* of 50 students in the previous example. We might be interested in whether the sample data give us convincing evidence of a difference in the *population* distributions of household size for South Africa and the United Kingdom. We'll have to wait a few chapters to decide whether we can reach such a conclusion, but our ability to make such an inference later will be helped by the fact that the students in our samples were chosen at random.

## Stemplots

Another simple graphical display for fairly small data sets is a **stemplot** (also called a stem-and-leaf plot). Stemplots give us a quick picture of the shape of a distribution while including the actual numerical values in the graph. Here's an example that shows how to make a stemplot.

### EXAMPLE

#### How Many Shoes?

##### Making a stemplot



How many pairs of shoes does a typical teenager have? To find out, a group of AP Statistics students conducted a survey. They selected a random sample of 20 female students from their school. Then they recorded the number of pairs of shoes that each respondent reported having. Here are the data:

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

Here are the steps in making a stemplot. Figure 1.13 displays the process.

- Separate each observation into a **stem**, consisting of all but the final digit, and a **leaf**, the final digit. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column. Do not skip any stems, even if there is no data value for a particular stem. For

these data, the tens digits are the stems, and the ones digits are the leaves. The stems run from 1 to 5.

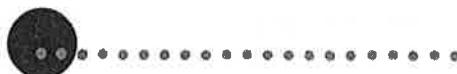
- Write each leaf in the row to the right of its stem. For example, the female student with 50 pairs of shoes would have stem 5 and leaf 0, while the student with 31 pairs of shoes would have stem 3 and leaf 1.

- Arrange the leaves in increasing order out from the stem.

- Provide a key that explains in context what the stems and leaves represent.

1	1	93335	1	33359	Key: 4 9 represents a female student who reported having 49 pairs of shoes.	
2	2	664233	2	233466		
3	3	1840	3	0148		
4	4	9	4	9		
5	5	0701	5	0017		

FIGURE 1.13 Making a stemplot of the shoe data. (1) Write the stems. (2) Go through the data and write each leaf on the proper stem. (3) Arrange the leaves on each stem in order out from the stem. (4) Add a key.

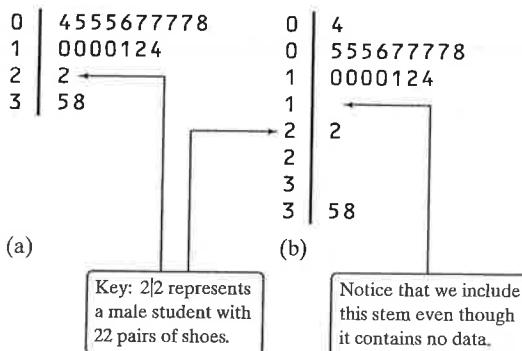


The AP Statistics students in the previous example also collected data from a random sample of 20 male students at their school. Here are the numbers of pairs of shoes reported by each male in the sample:

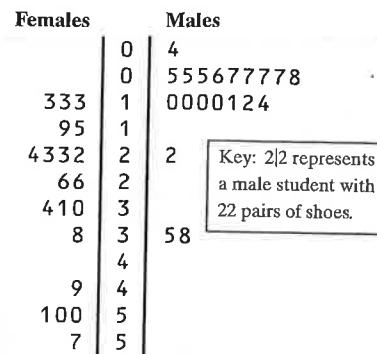
14	7	6	5	12	38	8	7	10	10
10	11	4	5	22	7	5	10	35	7

Splitting stems

What would happen if we tried the same approach as before: using the first digits as stems and the last digits as leaves? The completed stemplot is shown in Figure 1.14(a). What shape does this distribution have? It is difficult to tell with so few stems. We can get a better picture of male shoe ownership by **splitting stems**.



**FIGURE 1.14** Two stemplots showing the male shoe data. Figure 1.14(b) improves on the stemplot of Figure 1.14(a) by splitting stems.



**FIGURE 1.15** Back-to-back stemplot comparing numbers of pairs of shoes for male and female students at a school.

In Figure 1.14(a), the values from 0 to 9 are placed on the “0” stem. Figure 1.14(b) shows another stemplot of the same data. This time, values having leaves 0 through 4 are placed on one stem, while values ending in 5 through 9 are placed on another stem. Now we can see the single peak, the cluster of values between 4 and 14, and the large gap between 22 and 35 more clearly.

What if we want to compare the number of pairs of shoes that males and females have? That calls for a **back-to-back stemplot** with common stems. The leaves on each side are ordered out from the common stem. Figure 1.15 is a back-to-back stemplot for the male and female shoe data. Note that we have used the split stems from Figure 1.14(b) as the common stems. The values on the right are the male data from Figure 1.14(b). The values on the left are the female data, ordered out from the stem from right to left. We’ll ask you to compare these two distributions shortly.

Here are a few tips to consider before making a stemplot:

- Stemplots do not work well for large data sets, where each stem must hold a large number of leaves.
- There is no magic number of stems to use, but five is a good minimum. Too few or too many stems will make it difficult to see the distribution’s shape.
- If you split stems, be sure that each stem is assigned an equal number of possible leaf digits (two stems, each with five possible leaves; or five stems, each with two possible leaves).
- You can get more flexibility by rounding the data so that the final digit after rounding is suitable as a leaf. Do this when the data have too many digits. For example, in reporting teachers’ salaries, using all five digits (for example, \$42,549) would be unreasonable. It would be better to round to the nearest thousand and use 4 as a stem and 3 as a leaf.

#### Back-to-back stemplot

Instead of rounding, you can also *truncate* (remove one or more digits) when data have too many digits. The teacher’s salary of \$42,549 would truncate to \$42,000.

#### CHECK YOUR UNDERSTANDING

1. Use the back-to-back stemplot in Figure 1.15 to write a few sentences comparing the number of pairs of shoes owned by males and females. Be sure to address shape, center, spread, and outliers.

*Multiple choice: Select the best answer for Questions 2 through 4.*

Here is a stemplot of the percents of residents aged 65 and older in the 50 states and the District of Columbia. The stems are whole percents and the leaves are tenths of a percent.

6	8
7	
8	8
9	79
10	08
11	15566
12	012223444457888999
13	0123333444899
14	02666
15	23
16	8

Key: 8|8 represents a state in which 8.8% of residents are 65 and older.

Histogram

## EXAMPLE



2. The low outlier is Alaska. What percent of Alaska residents are 65 or older?  
(a) 0.68    (b) 6.8    (c) 8.8    (d) 16.8    (e) 68
3. Ignoring the outlier, the shape of the distribution is  
(a) skewed to the right    (c) skewed to the left.    (e) skewed to the middle.  
(b) roughly symmetric    (d) bimodal.
4. The center of the distribution is close to  
(a) 13.3%.    (b) 12.8%.    (c) 12.0%.    (d) 11.6%.    (e) 6.8% to 16.8%.

## Histograms

Quantitative variables often take many 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**. Let's look at how to make a histogram using data on foreign-born residents in the United States.

### Foreign-Born Residents

#### Making a histogram

What percent of your home state's residents were born outside the United States? The country as a whole has 12.5% foreign-born residents, but the states vary from 1.2% in West Virginia to 27.2% in California. The table below presents the data for all 50 states.<sup>21</sup> The *individuals* in this data set are the states. The *variable* is the percent of a state's residents who are foreign-born. It's much easier to see from a graph than from the table how your state compares with other states.

State	Percent	State	Percent	State	Percent
Alabama	2.8	Louisiana	2.9	Ohio	3.6
Alaska	7.0	Maine	3.2	Oklahoma	4.9
Arizona	15.1	Maryland	12.2	Oregon	9.7
Arkansas	3.8	Massachusetts	14.1	Pennsylvania	5.1
California	27.2	Michigan	5.9	Rhode Island	12.6
Colorado	10.3	Minnesota	6.6	South Carolina	4.1
Connecticut	12.9	Mississippi	1.8	South Dakota	2.2
Delaware	8.1	Missouri	3.3	Tennessee	3.9
Florida	18.9	Montana	1.9	Texas	15.9
Georgia	9.2	Nebraska	5.6	Utah	8.3
Hawaii	16.3	Nevada	19.1	Vermont	3.9
Idaho	5.6	New Hampshire	5.4	Virginia	10.1
Illinois	13.8	New Jersey	20.1	Washington	12.4
Indiana	4.2	New Mexico	10.1	West Virginia	1.2
Iowa	3.8	New York	21.6	Wisconsin	4.4
Kansas	6.3	North Carolina	6.9	Wyoming	2.7
Kentucky	2.7	North Dakota	2.1		

Here are the steps in making a histogram:

- Divide the range of the data into classes of equal width. The data in the table vary from 1.2 to 27.2, so we might choose to use classes of width 5, beginning at 0:

0–5   5–10   10–15   15–20   20–25   25–30

But we need to specify the classes so that each individual falls into exactly one class. For instance, what if a state had exactly 5.0% of its residents born outside the United States? Since a value of 0.0% would go in the 0–5 class, we'll agree to place a value of 5.0% in the 5–10 class, a value of 10.0% in the 10–15 class, and so on. In reality, then, our classes for the percent of foreign-born residents in the states are

0 to <5      5 to <10      10 to <15      15 to <20      20 to <25      25 to <30

- Find the count (frequency) or percent (relative frequency) of individuals in each class. Here is a frequency table and a relative frequency table for these data:

Notice that the frequencies add to 50, the number of individuals (states) in the data, and that the relative frequencies add to 100%.

Frequency table		Relative frequency table	
Class	Count	Class	Percent
0 to < 5	20	0 to < 5	40
5 to < 10	13	5 to < 10	26
10 to < 15	9	10 to < 15	18
15 to < 20	5	15 to < 20	10
20 to < 25	2	20 to < 25	4
25 to < 30	1	25 to < 30	2
Total	50	Total	100

- Label and scale your axes and draw the histogram. Label the horizontal axis with the variable whose distribution you are displaying. That's the percent of a state's residents who are foreign-born. The scale on the horizontal axis runs from 0 to 30 because that is the span of the classes we chose. The vertical axis contains the scale of counts or percents. Each bar represents a class. The base of the bar covers the class, and the bar height is the class frequency or relative frequency. Draw the bars with no horizontal space between them unless a class is empty, so that its bar has height zero.

Figure 1.16(a) shows a completed frequency histogram; Figure 1.16(b) shows a completed relative frequency histogram. The two graphs look identical except for the vertical scales.

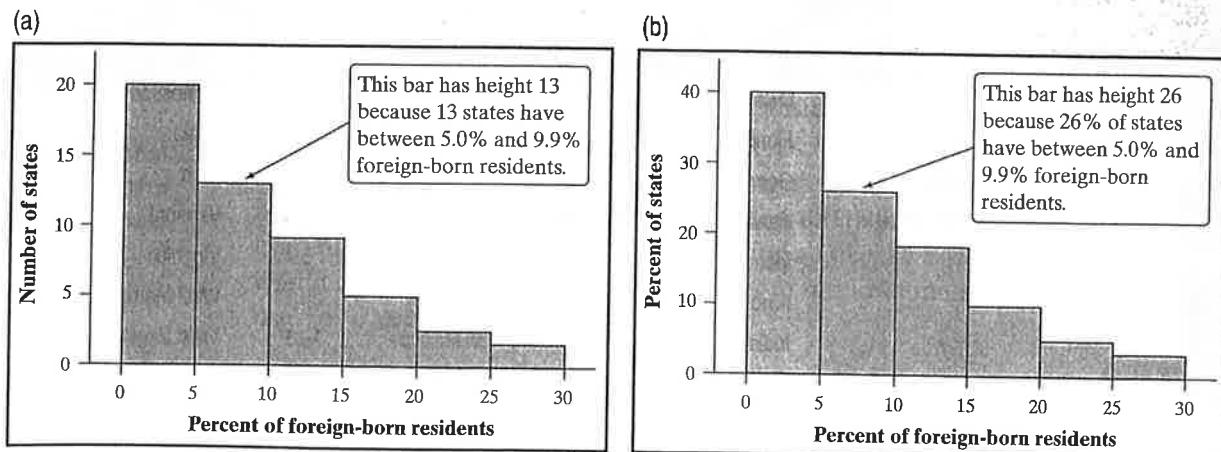


FIGURE 1.16 (a) Frequency histogram and (b) relative frequency histogram of the distribution of the percent of foreign-born residents in the 50 states.

What do the histograms in Figure 1.16 tell us about the percent of foreign-born residents in the states? To find out, we follow our familiar routine: describe the pattern and look for any departures from the pattern.

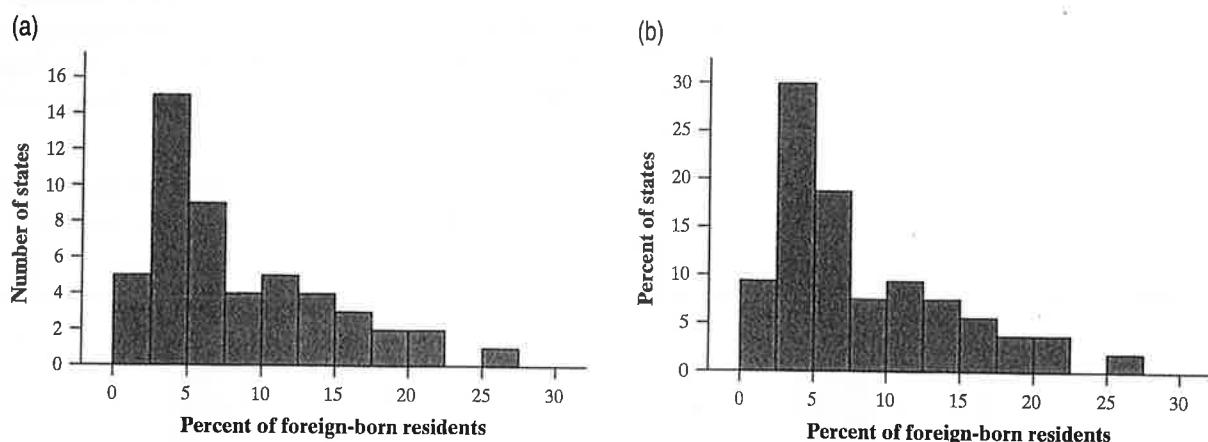
**Shape:** The distribution is skewed to the right. A majority of states have fewer than 10% foreign-born residents, but several states have much higher percents, so that the graph extends quite far to the right of its peak. The distribution has a *single peak* at the left, which represents states in which between 0% and 4.9% of residents are foreign-born.

**Center:** From the graph, we see that the midpoint (median) would fall somewhere in the 5.0% to 9.9% class. Remember that we're looking for the value having 25 states with smaller percents foreign-born and 25 with larger. (Arranging the observations from the table in order of size shows that the median is 6.1%.)

**Spread:** The histogram shows that the percent of foreign-born residents in the states varies from less than 5% to over 25%. (Using the data in the table, we see that the range is  $27.2\% - 1.2\% = 26.0\%$ .)

**Outliers:** We don't see any observations outside the overall single-peaked, right-skewed pattern of the distribution.

Figure 1.17 shows (a) a frequency histogram and (b) a relative frequency histogram of the same distribution, with classes half as wide. The new classes are 0–2.4, 2.5–4.9, etc. Now California, at 27.2%, stands out as a potential outlier in the right tail. The choice of classes in a histogram can influence the appearance of a distribution. Histograms with more classes show more detail but may have a less clear pattern.



**FIGURE 1.17** (a) Frequency histogram and (b) relative frequency histogram of the distribution of the percent of foreign-born residents in the 50 states, with classes half as wide as in Figure 1.16.

Statistical software and graphing calculators will choose the classes for you. The default choice is a good starting point, but you should adjust the classes to suit your needs. To see what we're talking about, launch the *One-Variable Statistical Calculator* applet at the book's Web site, [www.whfreeman.com/tps4e](http://www.whfreeman.com/tps4e). Select the "Percent of foreign-born residents" data set, and then click on the "Histogram" tab. You can change the number of classes by dragging the horizontal axis with your mouse or pointing device. By doing so, it's easy to see how the choice of classes affects the histogram. *Bottom line: Use your judgment in choosing classes to display the shape.*



## TECHNOLOGY CORNER Histograms on the calculator

TI-83/84

TI-89

1. Enter the data for the percent of state residents born outside the United States in your Statistics/List Editor.

- Press **STAT** and choose **1:Edit...**

- Type the values into list L1.

L1	L2	L3	L4
2.8			
7			
15.1			
3.8			
27.2			
10.3			
12.9			
<b>L1(0)=2.8</b>			

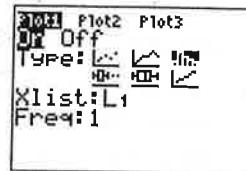
- Press **APPS** and select **Stats/List Editor**.

- Type the values into list1.

2. Set up a histogram in the Statistics Plots menu.

- Press **2nd Y =** (STAT PLOT).

- Press **ENTER** or **1** to go into Plot1.



- Adjust the settings as shown.

- Press **F2** and choose **1:Plot Setup...**

- With Plot1 highlighted, press **F1** to define.

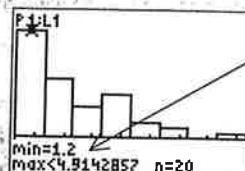


- Adjust the settings as shown.

3. Use ZoomStat (ZoomData on the TI-89) to let the calculator choose classes and make a histogram.

- Press **ZOOM** and choose **9:ZoomStat**.

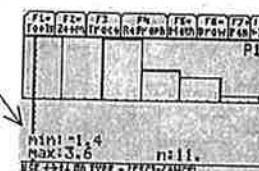
- Press **TRACE** and **◀ ▶** to examine the classes.



Note the calculator's unusual choice of classes.

- Press **F5** (ZoomData).

- Press **F3** (Trace) and **◀ ▶** to examine the classes.



4. Adjust the classes to match those in Figure 1.16, and then graph the histogram.

- Press **WINDOW** and enter the values shown.

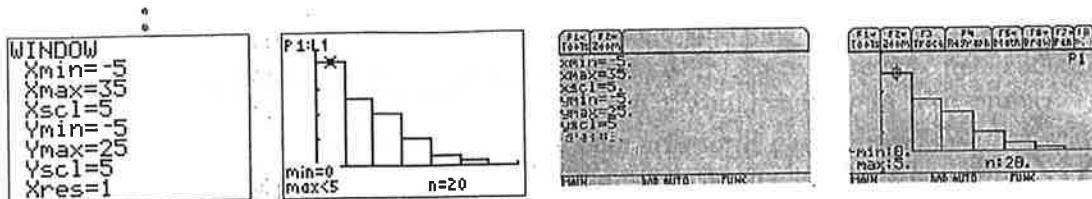
- Press **GRAPH**.

- Press **TRACE** and **◀ ▶** to examine the classes.

- Press **♦ F2** (WINDOW) and enter the values shown.

- Press **♦ F3** (GRAPH)

- Press **F3** (trace and **◀ ▶**) to examine the classes)



- 5. See if you can match the histogram in Figure 1.17.

TI-Nspire instructions in Appendix B

Here are some important things to consider when you are constructing a histogram:

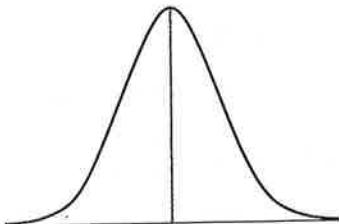
- Our eyes respond to the area of the bars in a histogram, so *be sure to choose classes that are all the same width*. Then area is determined by height, and all classes are fairly represented.
- There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, 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. Five classes is a good minimum.

**AP EXAM TIP** If you're asked to make a graph on a free-response question, be sure to label and scale your axes. Unless your calculator shows labels and scaling, don't just transfer a calculator screen shot to your paper.

### CHECK YOUR UNDERSTANDING



Many people believe that the distribution of IQ scores follows a “bell curve,” like the one shown in the margin. But is this really how such scores are distributed? The IQ scores of 60 fifth-grade students chosen at random from one school are shown below.<sup>22</sup>



145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

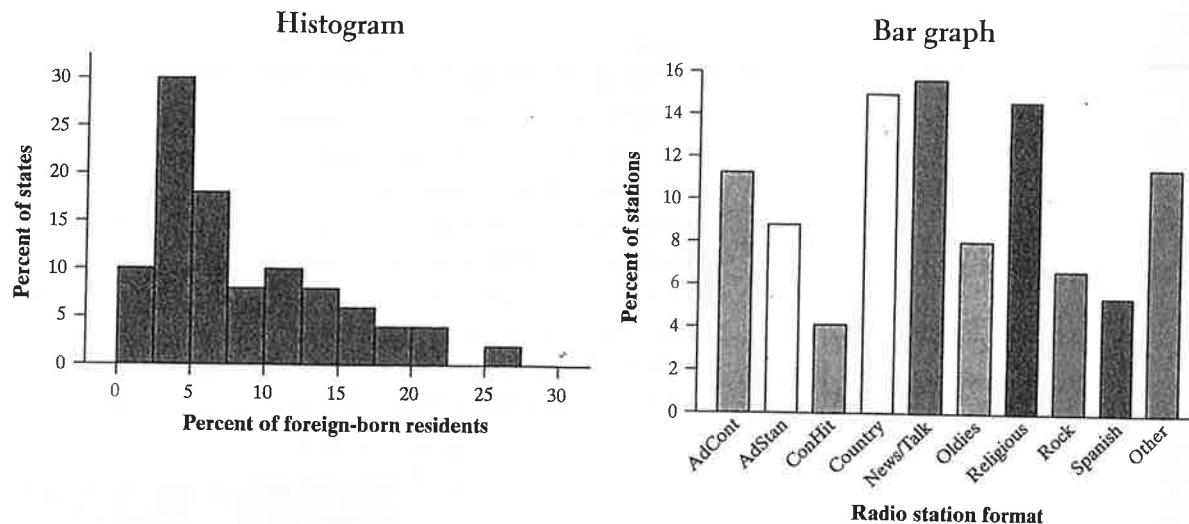
1. Construct a histogram that displays the distribution of IQ scores effectively.
2. Describe what you see. Is the distribution bell-shaped?

## Using Histograms Wisely

We offer several cautions based on common mistakes students make when using histograms.

1. *Don't confuse histograms and bar graphs.* 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 is used to display the distribution of a categorical variable or to compare the sizes of different quantities. The horizontal axis of a bar graph identifies the categories or quantities being compared. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to show the equal-width classes. For comparison, here is one of each type of graph from previous examples.

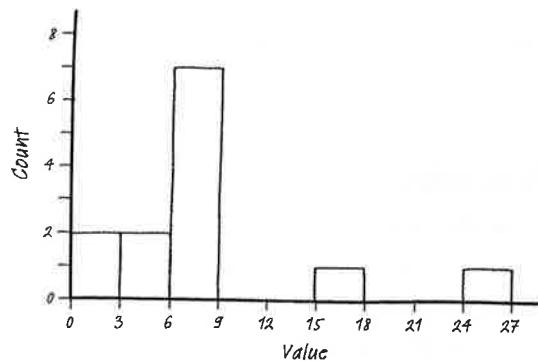




2. Don't use counts (in a frequency table) or percents (in a relative frequency table) as data. Below is a frequency table displaying the lengths (number of letters) of the first 100 words in a journal article.



Length:	1	2	3	4	5	6	7	8	9	10	11	12	13
Count:	1	15	25	7	5	7	8	7	7	6	8	3	1



Billy made the histogram shown to display these data. Can you see what Billy did wrong? (He used the counts as data when drawing the histogram—so there were two counts of 1, two counts between 3 and 5, and so on.) Question 1 in the Check Your Understanding below asks you to make a correct graph.

3. Use percents instead of counts on the vertical axis when comparing distributions with different numbers of observations. Mary was interested in comparing the reading levels of a medical journal and an airline's in-flight magazine. She counted the number of letters in the first 200 words of an article in the medical journal and of



the first 100 words of an article in the airline magazine. Mary then used Minitab statistical software to produce the histograms shown in Figure 1.18(a). This figure is misleading—it compares frequencies, but the two samples were of very different

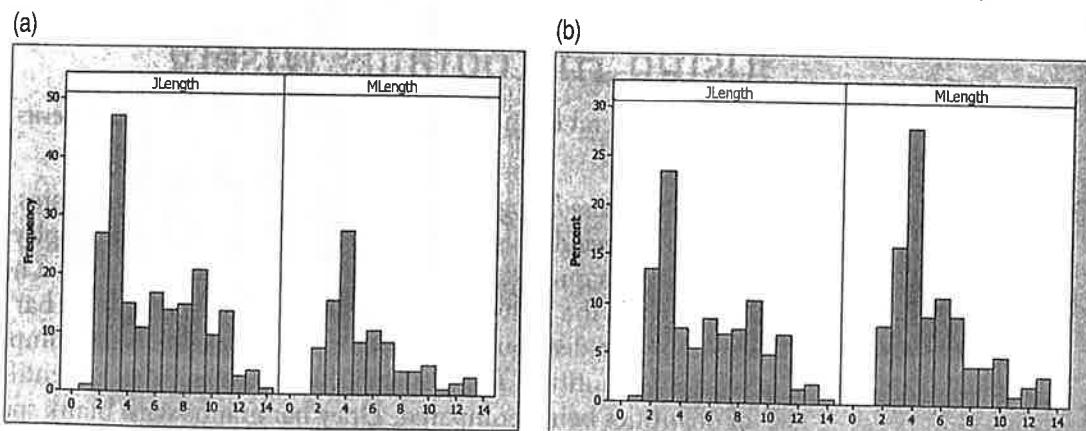
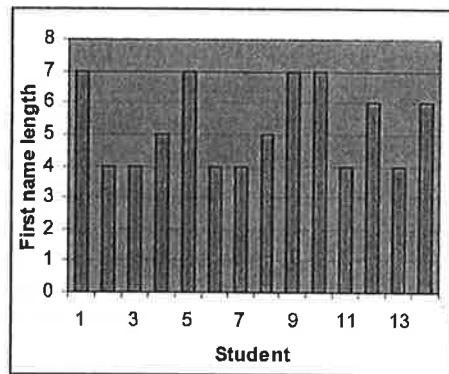


FIGURE 1.18 Two sets of histograms comparing word lengths in articles from a journal and from an airline magazine. In (a), the vertical scale uses frequencies. The graph in (b) fixes this problem by using percents on the vertical scale.

samples were of very different sizes (100 and 200). Using the same data, Mary's teacher produced the histograms in Figure 1.18(b). By using relative frequencies, this figure provides an accurate comparison of word lengths in the two samples.

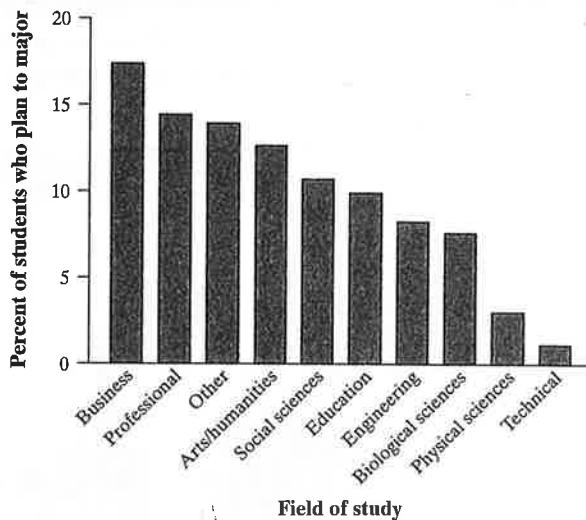
4. *Just because a graph looks nice, it's not necessarily a meaningful display of data.* The students in a small statistics class recorded the number of letters in their first names. One student entered the data into an Excel spreadsheet and then used Excel's "chart maker" to produce the graph shown. What kind of graph is this? It's neither a bar graph nor a histogram. Both of these types of graphs display the number or percent of individuals in a given category or class. This graph shows the individual data values, in the order that they were entered into the spreadsheet. It is not a very meaningful display of the data.



### CHECK YOUR UNDERSTANDING

- 1. Draw a correct histogram to replace Billy's graph of the word length data from Caution 2.
- 2. Draw a more meaningful graph of the first-name length data from Caution 4.

*Questions 3 and 4 relate to the following setting.* About 1.6 million first-year students enroll in colleges and universities each year. What do they plan to study? The graph displays data on the percents of first-year students who plan to major in several discipline areas.<sup>23</sup>



- 3. Is this a bar graph or a histogram? Explain.
- 4. Would it be correct to describe this distribution as right-skewed? Why or why not?

**SECTION 1.2**

# Summary

- You can use a **dotplot**, **stemplot**, or **histogram** to show the distribution of a quantitative variable. A dotplot displays individual values on a number line. Stemplots separate each observation into a stem and a one-digit leaf. Histograms plot the counts (frequencies) or percents (relative frequencies) of values in equal-width classes.
- When examining any graph, look for an **overall pattern** and for notable **departures** from that pattern. **Shape**, **center**, and **spread** describe the overall pattern of the distribution of a quantitative variable. **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them. Don't forget your SOCS!
- Some distributions have simple shapes, such as **symmetric** or **skewed**. The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.
- When comparing distributions of quantitative data, be sure to discuss shape, center, spread, and possible outliers.
- Remember: histograms are for quantitative data; bar graphs are for categorical data. Also, be sure to use relative frequency histograms when comparing data sets of different sizes.

**1.2 TECHNOLOGY CORNER**

Histograms on the calculator ..... page 38

TI-Nspire instructions in Appendix B

**SECTION 1.2**

# Exercises

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

9	6	8	6	8	8	6	6.5	6	7	9	4	3	4
5	6	11	6	3	6	6	10	7	8	4.5	9	7	7

(a) Make a dotplot to display the data.

(b) Describe the overall pattern of the distribution and any deviations from that pattern.

38. **Olympic gold!** The following table displays the total number of gold medals won by a sample of countries in the 2008 Summer Olympic Games in China.

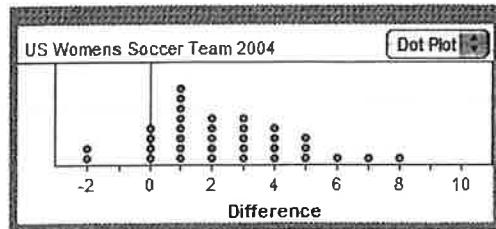
Country	Gold medals	Country	Gold medals
Sri Lanka	0	Thailand	2
China	51	Kuwait	0
Vietnam	0	Bahamas	0
Great Britain	19	Kenya	5
Norway	3	Trinidad and Tobago	0
Romania	4	Greece	0
Switzerland	2	Mozambique	0
Armenia	0	Kazakhstan	2
Netherlands	7	Denmark	2
India	0	Latvia	1
Georgia	3	Czech Republic	3
Kyrgyzstan	0	Hungary	3
Costa Rica	0	Sweden	0
Brazil	3	Uruguay	0
Uzbekistan	1	United States	36

- (a) Make a dotplot to display these data. Describe the overall pattern of the distribution and any deviations from that pattern.
- (b) Overall, 204 countries participated in the 2008 Summer Olympics, of which 55 won at least one gold medal. Do you believe that the sample of countries listed in the table is representative of this larger population? Why or why not?

39. U.S. women's soccer—2004 Earlier, we examined data on the number of goals scored by the U.S. women's soccer team in games during the 2004 season. The dotplot below displays the goal differential for those same games, computed as U.S. score minus opponent's score.

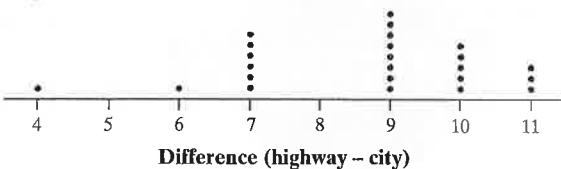
pg 28

pg 32



- (a) Explain what the two dots above  $-2$  represent.
- (b) What does the graph tell us about how well the team did in 2004? Be specific.

40. Fuel efficiency In an earlier example, we examined data on highway gas mileages of model year 2009 midsize cars. The dotplot below shows the difference (highway – city) in EPA mileage ratings for each of the 24 car models from the earlier example.

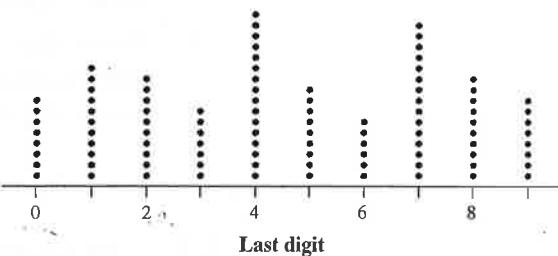


- (a) Explain what the dot above 6 represents.
- (b) What does the graph tell us about fuel economy in the city versus on the highway for these car models? Be specific.

#### 41. Dates on coins

- (a) Sketch a dotplot for a distribution that is skewed to the left.
- (b) Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.

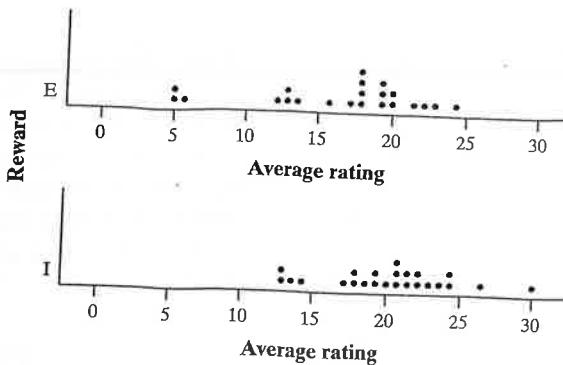
42. Phone numbers The dotplot below displays the last digit of 100 phone numbers chosen at random from a phone book. Describe the shape of the distribution. Does this shape make sense to you? Explain.



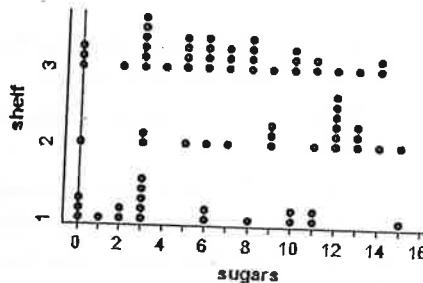
43. Creative writing The chapter-opening Case Study described research by Teresa Amabile investigating whether external rewards would promote creativity in children's artwork. Dr. Amabile conducted another study involving college students, who were divided into two groups using a chance process (like drawing names from a hat). The students in one group were given a list of statements about external reasons (E) for writing, such as public recognition, making money, or pleasing their parents. Students in the other group were given a list of statements about internal reasons (I) for writing, such as expressing yourself and enjoying playing with words. Both groups were then instructed to write a poem about laughter. Each student's poem was rated separately by 12 different poets using a creativity scale.<sup>24</sup> The 12 poets' ratings of each student's poem were averaged to obtain an overall creativity score.

A dotplot of the two groups' creativity scores is shown below. Compare the two distributions. What do you conclude about whether external rewards promote creativity?

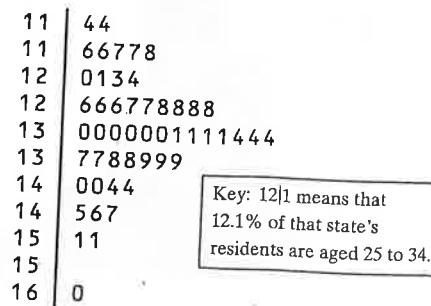
## 44 CHAPTER 1 EXPLORING DATA



44. Healthy cereal? Researchers collected data on 77 brands of cereal at a local supermarket.<sup>25</sup> For each brand, the sugar content (grams per serving) and the shelf in the store on which the cereal was located (1 = bottom, 2 = middle, 3 = top) were recorded. A dotplot of the data is shown below. Compare the three distributions. Critics claim that supermarkets tend to put sugary kids' cereals on lower shelves, where the kids can see them. Do the data from this study support this claim?



45. Where do the young live? Below is a stemplot of the percent of residents aged 25 to 34 in each of the 50 states. As in the stemplot for older residents (page 35), the stems are whole percents, and the leaves are tenths of a percent. This time, each stem has been split in two, with values having leaves 0 through 4 placed on one stem, and values ending in 5 through 9 placed on another stem.



- (a) Why did we split stems?  
 (b) Utah has the highest percent of residents aged 25 to 34. What is that percent? Why do you think Utah has an unusually high percent of residents in this age group?

(c) Describe the shape, center, and spread of the distribution, ignoring Utah.

46. Watch that caffeine! The U.S. Food and Drug Administration (USFDA) limits the amount of caffeine in a 12-ounce can of carbonated beverage to 72 milligrams. That translates to a maximum of 48 milligrams of caffeine per 8-ounce serving. Data on the caffeine content of popular soft drinks (in milligrams per 8-ounce serving) are displayed in the stemplot below.

1	556
2	033344
2	55667778888899
3	113
3	55567778
4	33
4	77

- (a) Why did we split stems?  
 (b) Give an appropriate key for this graph.  
 (c) Describe the shape, center, and spread of the distribution. Compare the caffeine content of these drinks with the USFDA's limit.

47. El Niño and the monsoon It appears that El Niño, the periodic warming of the Pacific Ocean west of South America, affects the monsoon rains that are essential for agriculture in India. Here are the monsoon rains (in millimeters) for the 23 strong El Niño years between 1871 and 2004:<sup>26</sup>

628	669	740	651	710	736	717	698	653	604	781	784
790	811	830	858	858	896	806	790	792	957	872	

- (a) To make a stemplot of these rainfall amounts, round the data to the nearest 10, so that stems are hundreds of millimeters and leaves are tens of millimeters. Make two stemplots, with and without splitting the stems. Which plot do you prefer?  
 (b) Describe the shape, center, and spread of the distribution.  
 (c) The average monsoon rainfall for all years from 1871 to 2004 is about 850 millimeters. What effect does El Niño appear to have on monsoon rains?

48. Shopping spree A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00	17.39
18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22	23.04	24.47
24.58	25.13	26.24	26.26	27.65	28.06	28.08	28.38	32.03	34.98
36.37	38.64	39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76	86.37	93.34

- (a) Round each amount to the nearest dollar. Then make a stemplot using tens of dollars as the stems and dollars as the leaves.
- (b) Make another stemplot of the data by splitting stems. Which of the plots shows the shape of the distribution better?
- (c) Write a few sentences describing the amount of money spent by shoppers at this supermarket.
49. Do women study more than men? We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

Women					Men				
180	120	180	360	240	90	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

- (a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? Are there any responses you consider suspicious?
- (b) Make a back-to-back stemplot to compare the two samples. Does it appear that women study more than men (or at least claim that they do)? Justify your answer.
50. Basketball playoffs Here are the scores of games played in the California Division I-AAA high school basketball playoffs:<sup>27</sup>

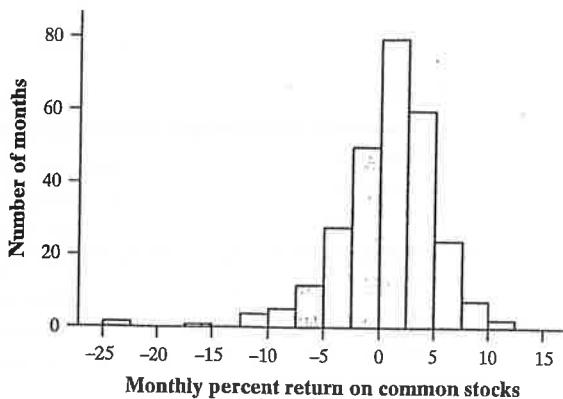
71-38	52-47	55-53	76-65	77-63	65-63	68-54	64-62
87-47	64-56	78-64	58-51	91-74	71-41	67-62	106-46

On the same day, the final scores of games in Division V-AA were

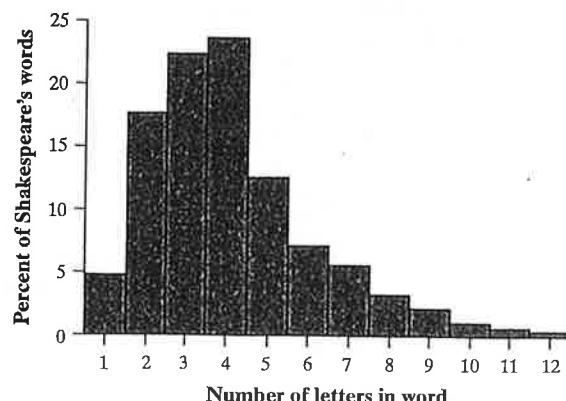
98-45	67-44	74-60	96-54	92-72	93-46
98-67	62-37	37-36	69-44	86-66	66-58

- (a) Construct a back-to-back stemplot to compare the points scored by the 32 teams in the Division I-AAA playoffs and the 24 teams in the Division V-AA playoffs.
- (b) Write a few sentences comparing the two distributions.
51. 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. The figure below shows a histogram of the distribution of the monthly returns for all common stocks listed on U.S. markets from January 1985 to September 2007 (273 months).<sup>28</sup> The extreme low outlier represents the market crash of October 1987, when stocks lost 23% of their value in one month.



- (a) Ignoring the outliers, describe the overall shape of the distribution of monthly returns.
- (b) What is the approximate center of this distribution?
- (c) Approximately what were the smallest and largest monthly returns, leaving out the outliers?
- (d) A return less than zero means that stocks lost value in that month. About what percent of all months had returns less than zero?
52. Shakespeare The histogram below shows the distribution of lengths of words used in Shakespeare's plays.<sup>29</sup> Describe the shape, center, and spread of this distribution.
53. Traveling to work How long do people travel each day to get to work? The following table gives the average travel times to work (in minutes) for workers in each state and the District of Columbia who are at least 16 years old and don't work at home.<sup>30</sup>



## 46 CHAPTER 1 EXPLORING DATA

AL	23.6	LA	25.1	OH	22.1
AK	17.7	ME	22.3	OK	20.0
AZ	25.0	MD	30.6	OR	21.8
AR	20.7	MA	26.6	PA	25.0
CA	26.8	MI	23.4	RI	22.3
CO	23.9	MN	22.0	SC	22.9
CT	24.1	MS	24.0	SD	15.9
DE	23.6	MO	22.9	TN	23.5
FL	25.9	MT	17.6	TX	24.6
GA	27.3	NE	17.7	UT	20.8
HI	25.5	NV	24.2	VT	21.2
ID	20.1	NH	24.6	VA	26.9
IL	27.9	NJ	29.1	WA	25.2
IN	22.3	NM	20.9	WV	25.6
IA	18.2	NY	30.9	WI	20.8
KS	18.5	NC	23.4	WY	17.9
KY	22.4	ND	15.5	DC	29.2

- (a) Make a histogram of the travel times using classes of width 2 minutes, starting at 14 minutes. That is, the first class is 14 to 16 minutes, the second is 16 to 18 minutes, and so on.
- (b) The shape of the distribution is a bit irregular. Is it closer to symmetric or skewed? About where is the center of the data? What is the spread in terms of the smallest and largest values? Are there any outliers?
54. **Carbon dioxide emissions** Burning fuels in power plants and motor vehicles emits carbon dioxide ( $\text{CO}_2$ ), which contributes to global warming. The table below displays  $\text{CO}_2$  emissions per person from countries with populations of at least 20 million.<sup>31</sup>
- (a) Make a histogram of the data using classes of width 2, starting at 0.
- (b) Describe the shape, center, and spread of the distribution. Which countries are outliers?

**Carbon dioxide emissions**  
(metric tons per person)

Country	$\text{CO}_2$	Country	$\text{CO}_2$
Algeria	2.6	Egypt	2.0
Argentina	3.6	Ethiopia	0.1
Australia	18.4	France	6.2
Bangladesh	0.3	Germany	9.9
Brazil	1.8	Ghana	0.3
Canada	17.0	India	1.1
China	3.9	Indonesia	1.6
Colombia	1.3	Iran	6.0
Congo	0.2	Iraq	2.9

Country	$\text{CO}_2$	Country	$\text{CO}_2$
Italy	7.8	Romania	4.2
Japan	9.5	Russia	10.8
Kenya	0.3	Saudi Arabia	13.8
Korea, North	3.3	South Africa	7.0
Korea, South	9.3	Spain	7.9
Malaysia	5.5	Sudan	0.3
Mexico	3.7	Tanzania	0.1
Morocco	1.4	Thailand	3.3
Myanmar	0.2	Turkey	3.0
Nepal	0.1	Ukraine	6.3
Nigeria	0.4	United Kingdom	8.8
Pakistan	0.8	United States	19.6
Peru	1.0	Uzbekistan	4.2
Philippines	0.9	Venezuela	5.4
Poland	7.8	Vietnam	1.0

55. **DRP test scores** There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research study on third-grade students, the DRP was administered to 44 students.<sup>32</sup> Their scores were:

40	26	39	14	42	18	25	43	46	27	19
47	19	26	35	34	15	44	40	38	31	46
52	25	35	35	33	29	34	41	49	28	52
47	35	48	22	33	41	51	27	14	54	45

Make a histogram to display the data. Write a paragraph describing the distribution of DRP scores.

56. **Drive time** Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays:

8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?

57. **The statistics of writing style** Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in *Popular Science* magazine:<sup>33</sup>

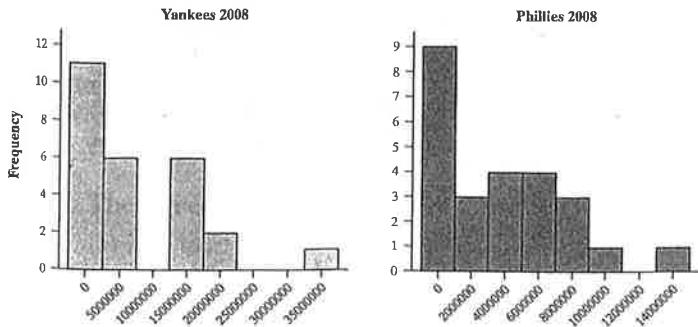
Length:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent:	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

- (a) Make a histogram of this distribution. Describe its shape, center, and spread.
- (b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution for Shakespeare's plays in Exercise 52? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).
58. **Chest out, Soldier!** In 1846, a published paper provided chest measurements (in inches) of 5738 Scottish militiamen. The table below summarizes the data.

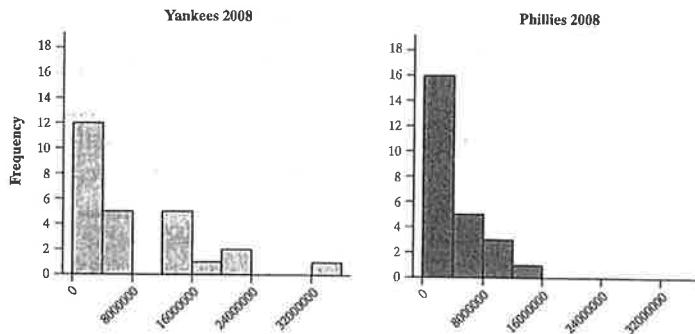
Chest size	Count	Chest size	Count
33	3	41	934
34	18	42	658
35	81	43	370
36	185	44	92
37	420	45	50
38	749	46	21
39	1073	47	4
40	1079	48	1

Source: Online Data and Story Library (DASL).

- (a) Make a histogram.
- (b) Describe the shape, center, and spread of the chest measurements distribution. Why might this information be useful?
59. **Paying for championships** Does paying high salaries lead to more victories in professional sports? The New York Yankees have long been known for having Major League Baseball's highest team payroll. And over the years, the team has won many championships. This strategy didn't pay off in 2008, when the Philadelphia Phillies won the World Series. Maybe the Yankees didn't spend enough money that year. The graph below shows histograms of the salary distributions for the two teams during the 2008 season. Why can't you use this graph to effectively compare the team payrolls?



60. **Paying for championships** Refer to Exercise 59. Here is another graph of the 2008 salary distributions for the Yankees and the Phillies. Write a few sentences comparing these two distributions.



61. **Birth months** Imagine asking a random sample of 60 students from your school about their birth months. Draw a plausible graph of the distribution of birth months. (Hint: Should you use a bar graph or a histogram?)
62. **Die rolls** Imagine rolling a fair, six-sided die 60 times. Draw a plausible graph of the distribution of die rolls. (Hint: Should you use a bar graph or a histogram?)
63. **Who makes more?** A manufacturing company is reviewing the salaries of its full-time employees below the executive level at a large plant. The clerical staff is almost entirely female, while a majority of the production workers and technical staff are male. As a result, the distributions of salaries for male and female employees may be quite different. The table below gives the frequencies and relative frequencies for women and men.

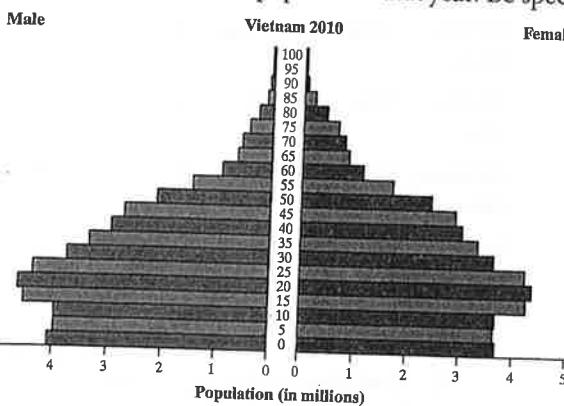
Salary (\$1000)	Women		Men	
	Number	%	Number	%
10-15	89	11.8	26	1.1
15-20	192	25.4	221	9.0
20-25	236	31.2	677	27.6
25-30	111	14.7	823	33.6
30-35	86	11.4	365	14.9
35-40	25	3.3	182	7.4
40-45	11	1.5	91	3.7
45-50	3	0.4	33	1.3
50-55	2	0.3	19	0.8
55-60	0	0.0	11	0.4
60-65	0	0.0	0	0.0
65-70	1	0.1	3	0.1
Total	756	100.1	2451	99.9

- (a) Explain why the total for women is greater than 100%.

(b) Make histograms for these data, choosing the vertical scale that is most appropriate for comparing the two distributions.

(c) Write a few sentences comparing the salary distributions for men and women.

64. **Population pyramids** A population pyramid is a helpful graph for examining the distribution of a country's population. Here is a population pyramid for Vietnam in the year 2010. Describe what the graph tells you about Vietnam's population that year. Be specific.



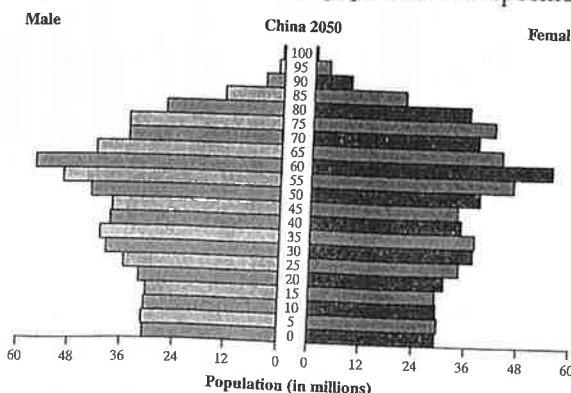
65. **Comparing AP scores** The table below gives the distribution of grades earned by students taking the AP Calculus AB and AP Statistics exams in 2009.<sup>34</sup>

No. of exams	Grade				
	5	4	3	2	1
Calculus AB	230,588	52,921	43,140	41,204	35,843
Statistics	116,876	14,353	26,050	28,276	22,283
					25,914

(a) Make an appropriate graphical display to compare the grade distributions for AP Calculus AB and AP Statistics.

(b) Write a few sentences comparing the two distributions of exam grades.

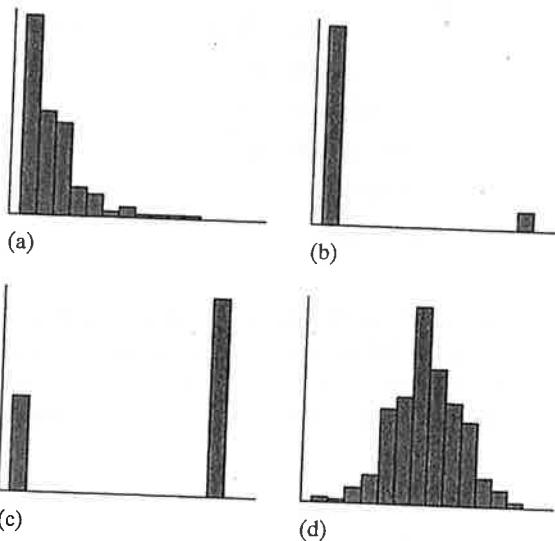
66. **Population pyramids** Refer to Exercise 64. Here is a graph of the projected population distribution for China in the year 2050. Describe what the graph suggests about China's future population. Be specific.



67. **Student survey** A survey of a large high school class asked the following questions:

- (i) Are you female or male? (In the data, male = 0, female = 1.)
- (ii) Are you right-handed or left-handed? (In the data, right = 0, left = 1.)
- (iii) What is your height in inches?
- (iv) How many minutes do you study on a typical weeknight?

The figure below shows histograms of the student responses, in scrambled order and without scale markings. Which histogram goes with each variable? Explain your reasoning.



68. **Choose a graph** What type of graph or graphs would you plan to make in a study of each of the following issues at your school? Explain your choices.

- (a) Which radio stations are most popular with students?
- (b) How many hours per week do students study?
- (c) How many calories do students consume per day?

*Multiple choice: Select the best answer for Exercises 69 to 74.*

69. Here are the amounts of money (cents) in coins carried by 10 students in a statistics class: 50, 35, 0, 97, 76, 0, 0, 87, 23, 65. To make a stemplot of these data, you would use stems

- (a) 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
- (b) 0, 2, 3, 5, 6, 7, 8, 9.
- (c) 0, 3, 5, 6, 7.
- (d) 00, 10, 20, 30, 40, 50, 60, 70, 80, 90.
- (e) None of these.

70. One of the following 12 scores was omitted from the stemplot below:

84 76 92 92 88 96 68 80 92 88 76 96

6	8
7	66
8	0488
9	2266

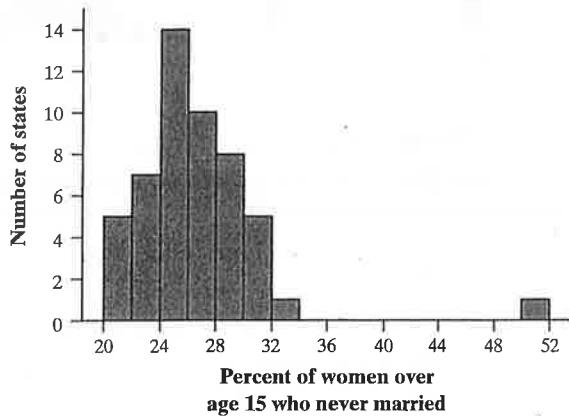
The missing number is

- (a) 76. (b) 88. (c) 90. (d) 92. (e) 96.

71. 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.  
 (d) unimodal.  
 (e) too high.

*Exercises 72 to 74 refer to the following setting.* The histogram below shows the distribution of the percents of women aged 15 and over who have never married in each of the 50 states and the District of Columbia.



72. The leftmost bar in the histogram covers percents of never-married women ranging from about  
 (a) 20% to 24%. (d) 0% to 5%.  
 (b) 20% to 22%. (e) None of these.  
 (c) 0% to 20%.

73. The center of this distribution is in the interval  
 (a) 22% to 24%. (d) 28% to 30%.  
 (b) 24% to 26%. (e) 36% to 38%.  
 (c) 26% to 28%.

74. In about what percent of states have at least 30% of women aged 15 and over never married?

- (a) 4% (b) 7% (c) 10% (d) 14% (e) 32%

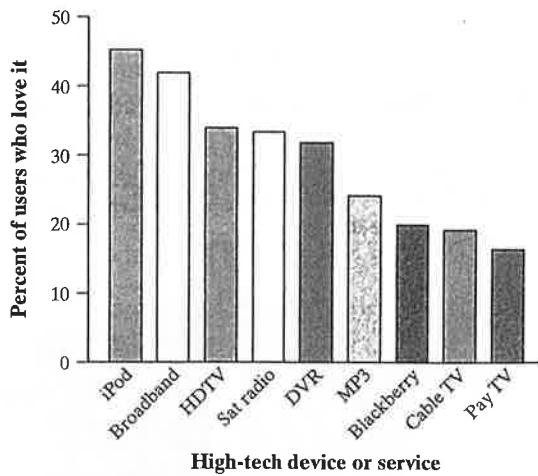
75. Baseball players (Introduction) Here is a small part of a data set that describes Major League Baseball players as of opening day of the 2009 season:

Player	Team	Position	Age	Height	Weight	Salary
Rodriguez,	Yankees	Infielder	33	6-3	230	\$33,000,000
Alex						
Ramirez,	Dodgers	Outfielder	36	6-0	200	\$23,854,494
Manny						
Santana,	Mets	Pitcher	30	6-0	210	\$18,876,139
Johan						
Zambrano,	Cubs	Pitcher	27	6-5	255	\$18,750,000
Carlos						
Suzuki,	Mariners	Outfielder	35	5-11	170	\$18,000,000
Ichiro						

- (a) What individuals does this data set describe?  
 (b) In addition to the player's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?  
 (c) What do you think are the units of measurement for each of the quantitative variables?

76. I love my iPod! (1.1) The rating service Arbitron asked adults who used several high-tech devices and services whether they "loved" using them. Below is a graph of the percents who said they did.<sup>35</sup>

- (a) Summarize what this graph tells you in a sentence or two.  
 (b) Would it be appropriate to make a pie chart of these data? Why or why not?



77. Risks of playing soccer (1.1) A study in Sweden looked at former elite soccer players, people who had played soccer but not at the elite level, and

people of the same age who did not play soccer. Here is a two-way table that classifies these individuals by whether or not they had arthritis of the hip or knee by their mid-fifties.<sup>36</sup>

	Elite	Non-Elite	Did not play
Arthritis	10	9	24
No arthritis	61	206	548

- (a) What percent of the people in this study were elite soccer players? What percent had arthritis?

- (b) What percent of the elite soccer players had arthritis? What percent of those who had arthritis were elite soccer players?

78. Risks of playing soccer (1.1) Refer to Exercise 77. We suspect that the more serious soccer players have more arthritis later in life. Do the data confirm this suspicion? Give graphical and numerical evidence to support your answer.

### 1.3

#### In Section 1.3, you'll learn about:

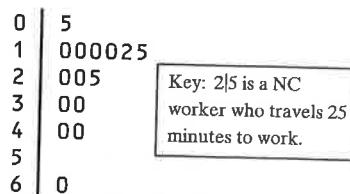
- Measuring center: The mean
- Measuring center: The median
- Comparing the mean and the median
- Measuring spread: The interquartile range (*IQR*)
- Identifying outliers
- The five-number summary and boxplots
- Measuring spread: The standard deviation
- Numerical summaries with technology
- Choosing measures of center and spread

## Describing Quantitative Data with Numbers

How long do people spend traveling to work? The answer may depend on where they live. Here are the travel times in minutes for 15 workers in North Carolina, chosen at random by the Census Bureau.<sup>37</sup>

30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

We aren't surprised that most people estimate their travel time in multiples of 5 minutes. Here is a stemplot of these data:



The distribution is single-peaked and right-skewed. The longest travel time (60 minutes) may be an outlier. Our goal in this section is to describe the center and spread of this and other distributions of quantitative data with numbers.

## Measuring Center: The Mean

The most common measure of center is the ordinary arithmetic average, or **mean**.

### DEFINITION: The mean $\bar{x}$

To find the mean  $\bar{x}$  (pronounced "x-bar") of a set of observations, add their values and divide by the number of observations. If the  $n$  observations are  $x_1, x_2, \dots, x_n$ , their mean is

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{\sum x_i}{n}$$

The  $\Sigma$  (capital Greek letter sigma) in the formula for the mean is short for “add them all up.” The subscripts on the observations  $x_i$  are just a way of keeping the  $n$  observations distinct. They do not necessarily indicate order or any other special facts about the data.

Actually, the notation  $\bar{x}$  refers to the mean of a *sample*. Most of the time, the data we’ll encounter can be thought of as a sample from some larger population. When we need to refer to a *population* mean, we’ll use the symbol  $\mu$  (Greek letter mu, pronounced “mew”). If you have the entire population of data available, then you calculate  $\mu$  in just the way you’d expect: add the values of all the observations, and divide by the number of observations.

### EXAMPLE

#### *Travel Times to Work in North Carolina*

##### Calculating the mean

Refer to the data on travel times to work for the sample of 15 North Carolinians.

0	5	Key: 2 5 is a NC worker who travels 25 minutes to work.
1	000025	
2	005	
3	00	
4	00	
5	0	
6	0	

##### PROBLEM:

- (a) Find the mean travel time for all 15 workers.
- (b) Calculate the mean again, this time excluding the person who reported a 60-minute travel time to work. What do you notice?

##### SOLUTION:

- (a) The mean travel time for the sample of 15 North Carolina workers is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{30 + 20 + \dots + 10}{15} = \frac{337}{15} = 22.5 \text{ minutes}$$

- (b) Notice that only 6 of the 15 travel times are larger than the mean. If we leave out the longest travel time, 60 minutes, the mean for the remaining 14 people is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{277}{14} = 19.8 \text{ minutes}$$

That one observation raises the mean by 2.7 minutes.



## Resistant measure

The previous example illustrates an important weakness of the mean as a measure of center: *the mean is sensitive to the influence of extreme observations.* These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure** of center.



**What does the mean mean?** A group of elementary school children was asked how many pets they have. Here are their responses, arranged from lowest to highest:<sup>38</sup>

1 3 4 4 4 5 7 8 9

What's the mean number of pets for this group of children? It's

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{1+3+4+4+4+5+7+8+9}{9} = 5 \text{ pets}$$

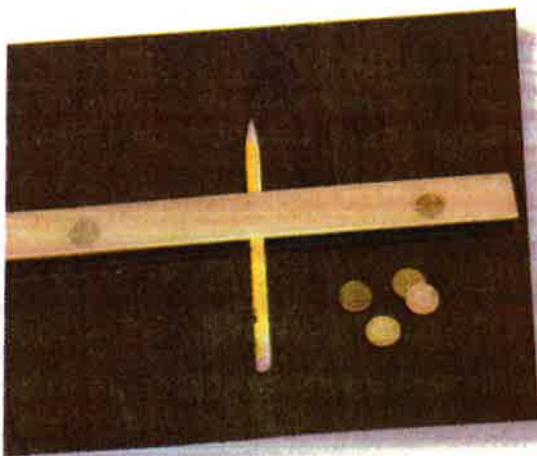
But what does that number tell us? Here's one way to look at it: if every child in the group had the same number of pets, each would have 5 pets. In other words, the mean is the "fair share" value.

The mean tells us how large each observation in the data set would be if the total were split equally among all the observations. In the language of young children, the mean is the "fair share" value. The mean of a distribution also has a physical interpretation, as the following Activity shows.

- **ACTIVITY** *Mean as a "balance point"*

- **MATERIALS:** Foot-long ruler, pencil, and 5 pennies per group of 3 to 4 students

In this Activity, you'll investigate an interesting property of the mean.



1. Stack all 5 pennies above the 6-inch mark on your ruler. Place your pencil under the ruler to make a "seesaw" on a desk or table. Move the pencil until the ruler balances. What is the relationship between the location of the pencil and the mean of the five data values: 6, 6, 6, 6, 6?
2. Move one penny off the stack to the 8-inch mark on your ruler. Now move one other penny so that the ruler balances again without moving the pencil. Where did you put the other penny? What is the mean of the five data values represented by the pennies now?
3. Move one more penny off the stack to the 2-inch mark on your ruler. Now move both remaining pennies from the 6-inch mark so that the ruler still balances with the pencil in the same location. Is the mean of the data values still 6?
4. Do you see why the mean is sometimes called the "balance point" of a distribution?

## Measuring Center: The Median

In Section 1.2, we introduced the median as an informal measure of center that describes the “midpoint” of a distribution. Now it’s time to offer an official “rule” for calculating the median.

### DEFINITION: The median $M$

The median  $M$  is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations  $n$  is odd, the median  $M$  is the center observation in the ordered list.
3. If the number of observations  $n$  is even, the median  $M$  is the average of the two center observations in the ordered list.

Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of values in order is tedious, however, so finding the median by hand for larger sets of data is unpleasant.

### EXAMPLE

#### *Travel Times to Work in North Carolina*

##### Finding the median when $n$ is odd

What is the median travel time for our 15 North Carolina workers? Here are the data arranged in order:

5 10 10 10 10 10 12 15 20 20 25 30 30 40 40 60

The count of observations  $n = 15$  is odd. The bold 20 is the center observation in the ordered list, with 7 observations to its left and 7 to its right. This is the median,  $M = 20$  minutes.

### EXAMPLE

#### *Stuck in Traffic*

##### Finding the median when $n$ is even

People say that it takes a long time to get to work in New York State due to the heavy traffic near big cities. What do the data say? Here are the travel times in minutes of 20 randomly chosen New York workers:

10 30 5 25 40 20 10 15 30 20 15 20 85 15 65 15 60 60 40 45



0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 is a New York worker who reported a 45-minute travel time to work.

A stemplot makes finding the median easy because it arranges the observations in order.

#### PROBLEM:

- (a) Make a stemplot of the data. Be sure to include a key.  
 (b) Find and interpret the median.

#### SOLUTION:

- (a) Here is a stemplot of the data. The stems are in 10 minutes and the leaves are in minutes.  
 (b) Since there is an even number of data values, there is no center observation. There is a center pair—the bold 20 and 25 in the stemplot—which have 9 observations before them and 9 after them in the ordered list. The median is the average of these two observations:

$$M = \frac{20 + 25}{2} = 22.5 \text{ minutes}$$

*Interpretation:* In the sample of New York workers, about half of the people reported traveling less than 22.5 minutes to work, and about half reported traveling more.

For Practice Try Exercise 81.

## Comparing the Mean and the Median

0	5
1	000025
2	005
3	00
4	00
5	
6	0

Key: 2|5 is a NC worker who travels 25 minutes to work.

Our discussion of travel times to work in North Carolina illustrates an important difference between the mean and the median. The median travel time (the midpoint of the distribution) is 20 minutes. The mean travel time is higher, 22.5 minutes. The mean is pulled toward the right tail of this right-skewed distribution. The median, unlike the mean, is *resistant*. If the longest travel time were 600 minutes rather than 60 minutes, the mean would increase to more than 58 minutes but the median would not change at all. The outlier just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward.

You can compare the behavior of the mean and median by using the *Mean and Median* applet at the book's Web site, [www.whfreeman.com/tps4e](http://www.whfreeman.com/tps4e).



### Comparing the Mean and Median

The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is usually farther out in the long tail than is the median.<sup>39</sup>

The mean and median measure center in different ways, and both are useful. Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.



Should we choose the mean or the median? Many economic variables have distributions that are skewed to the right. College tuitions, home prices, and personal incomes are all right-skewed. In Major League Baseball (MLB), for instance, most players earn close to the minimum salary (which was \$400,000 in 2009), while a few earn more than \$10 million. The median salary for MLB players in 2009 was about \$1.15 million—but the mean salary was over \$3.24 million. Alex

THINK  
ABOUT  
IT

Rodriguez, Manny Ramirez, Derek Jeter, and several other highly paid superstars pull the mean up but do not affect the median. Reports about incomes and other strongly skewed distributions usually give the median (“midpoint”) rather than the mean (“arithmetic average”). However, a county that is about to impose a tax of 1% on the incomes of its residents cares about the mean income, not the median. The tax revenue will be 1% of total income, and the total is the mean times the number of residents.



### CHECK YOUR UNDERSTANDING

Questions 1 through 4 refer to the following setting. Here, once again, is the stemplot of travel times to work for 20 randomly selected New Yorkers. Earlier, we found that the median was 22.5 minutes.

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 is a  
New York worker  
who reported a  
45-minute travel  
time to work.

1. Based only on the stemplot, would you expect the mean travel time to be less than, about the same as, or larger than the median? Why?
2. Use your calculator to find the mean travel time. Was your answer to Question 1 correct?
3. Interpret your result from Question 2 in context without using the words “mean” or “average.”
4. Would the mean or the median be a more appropriate summary of the center of this distribution of drive times? Justify your answer.

## Measuring Spread: The Interquartile Range (*IQR*)

A measure of center alone can be misleading. The mean annual temperature in San Francisco, California, is 57°F—the same as in Springfield, Missouri. But the wardrobe needed to live in these two cities is very different! That’s because daily temperatures vary a lot more in Springfield than in San Francisco. A useful numerical description of a distribution requires both a measure of center and a measure of spread.

The simplest measure of variability is the *range*, which we defined earlier as the difference between the largest and smallest observations. The range shows the full spread of the data. But it depends on only the maximum and minimum values, which may be outliers.

We can improve our description of spread by also looking at the spread of the middle half of the data. Here’s the idea. Count up the ordered list of observations, starting from the minimum. The first quartile  $Q_1$  lies one-quarter of the way up the list. The second quartile is the median, which is halfway up the list. The third quartile  $Q_3$  lies three-quarters of the way up the list. These quartiles mark out the middle half of the distribution. The *interquartile range* (*IQR*) measures the range of the middle 50% of the data. We need a rule to make this idea exact. The process for calculating the quartiles and the *IQR* uses the rule for finding the median.

### How to Calculate the Quartiles $Q_1$ and $Q_3$ and the Interquartile Range ( $IQR$ )

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median  $M$  in the ordered list of observations.
2. The first quartile  $Q_1$  is the median of the observations whose position in the ordered list is to the left of the median.
3. The third quartile  $Q_3$  is the median of the observations whose position in the ordered list is to the right of the median.

The interquartile range ( $IQR$ ) is defined as

$$IQR = Q_3 - Q_1$$

Be careful in locating the quartiles when several observations take the same numerical value. Write down all the observations, arrange them in order, and apply the rules just as if they all had distinct values.

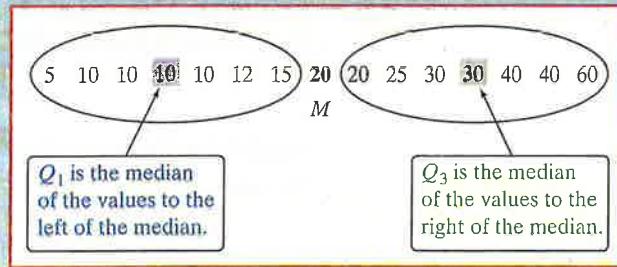
Let's look at how this process works using a familiar set of data.

#### EXAMPLE

#### Travel Times to Work in North Carolina

##### Calculating quartiles

Our North Carolina sample of 15 workers' travel times, arranged in increasing order, is



There is an odd number of observations, so the median is the middle one, the bold 20 in the list. The first quartile is the median of the 7 observations to the left of the median. This is the 4th of these 7 observations, so  $Q_1 = 10$  minutes (shown in blue). The third quartile is the median of the 7 observations to the right of the median,  $Q_3 = 30$  minutes (shown in green). So the spread of the middle 50% of the travel times is  $IQR = Q_3 - Q_1 = 30 - 10 = 20$  minutes. Be sure to leave out the overall median  $M$  when you locate the quartiles.



The quartiles and the interquartile range are *resistant* because they are not affected by a few extreme observations. For example,  $Q_3$  would still be 30 and the  $IQR$  would still be 20 if the maximum were 600 rather than 60.

**EXAMPLE*****Stuck in Traffic Again*****Finding and interpreting the IQR**

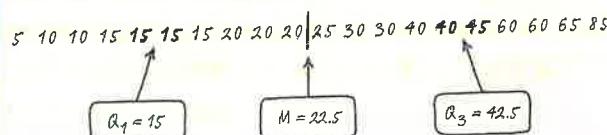
In an earlier example, we looked at data on travel times to work for 20 randomly selected New Yorkers. Here is the stemplot once again:

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 is a  
New York worker  
who reported a  
45-minute travel  
time to work.

**PROBLEM:** Find and interpret the interquartile range (IQR).

**SOLUTION:** We begin by writing the travel times arranged in increasing order:



There is an even number of observations, so the median lies halfway between the middle pair. Its value is  $M = 22.5$  minutes. (We marked the location of the median by |.) The first quartile is the median of the 10 observations to the left of  $M = 22.5$ . So it's the average of the two bold 15s:  $Q_1 = 15$  minutes. The third quartile is the median of the 10 observations to the right of  $M = 22.5$ . It's the average of the bold numbers 40 and 45:  $Q_3 = 42.5$  minutes. The interquartile range is

$$IQR = Q_3 - Q_1 = 42.5 - 15 = 27.5 \text{ minutes}$$

*Interpretation:* The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.

**For Practice Try Exercise 89(a)**

**Identifying Outliers**

In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

**DEFINITION: The  $1.5 \times IQR$  rule for outliers**

Call an observation an outlier if it falls more than  $1.5 \times IQR$  above the third quartile or below the first quartile.

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 is a  
New York worker  
who reported a  
45-minute travel  
time to work.

Does the  $1.5 \times IQR$  rule identify any outliers for the New York travel time data? In the previous example, we found that  $Q_1 = 15$  minutes,  $Q_3 = 42.5$  minutes, and  $IQR = 27.5$  minutes. For these data,

$$1.5 \times IQR = 1.5(27.5) = 41.25$$

Any values not falling between

$$Q_1 - 1.5 \times IQR = 15 - 41.25 = -26.25 \quad \text{and}$$

$$Q_3 + 1.5 \times IQR = 42.5 + 41.25 = 83.75$$

are flagged as outliers. Look again at the stemplot: the only outlier is the longest travel time, 85 minutes. The  $1.5 \times IQR$  rule suggests that the three next-longest travel times (60 and 65 minutes) are just part of the long right tail of this skewed distribution.

### EXAMPLE

#### Travel Times to Work in North Carolina

##### Identifying outliers

0	5
1	000025
2	005
3	00
4	00
5	0
6	0

Key: 25 is a NC worker who travels 25 minutes to work.

Earlier, we noted the influence of one long travel time of 60 minutes in our sample of 15 North Carolina workers.

**PROBLEM:** Determine whether this value is an outlier.

**SOLUTION:** Earlier, we found that  $Q_1 = 10$  minutes,  $Q_3 = 30$  minutes, and  $IQR = 20$  minutes. To check for outliers, we first calculate

$$1.5 \times IQR = 1.5(20) = 30$$

By the  $1.5 \times IQR$  rule, any value greater than

$$Q_3 + 1.5 \times IQR = 30 + 30 = 60$$

or less than

$$Q_1 - 1.5 \times IQR = 10 - 30 = -20$$

would be classified as an outlier. The maximum value of 60 minutes is not quite large enough to be flagged as an outlier.

For Practice Try Exercise 89(b)

**AP EXAM TIP** You may be asked to determine whether a quantitative data set has any outliers. Be prepared to state and use the rule for identifying outliers.

Whenever you find outliers in your data, try to find an explanation for them. Sometimes the explanation is as simple as a typing error, like typing 10.1 as 101. Sometimes a measuring device broke down or someone gave a silly response, like the student in a class survey who claimed to study 30,000 minutes per night. (Yes, that really happened.) In all these cases, you can simply remove the outlier from your data. When outliers are “real data,” like the long travel times of some New York workers, you should choose statistical methods that are not greatly affected by the outliers.

## The Five-Number Summary and Boxplots

The smallest and largest observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only the median and the quartiles. To get a quick summary of both center and spread, combine all five numbers.

**DEFINITION: The five-number summary**

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum  $Q_1$   $M$   $Q_3$  Maximum

These five numbers divide each distribution roughly into quarters. About 25% of the data values fall between the minimum and  $Q_1$ , about 25% are between  $Q_1$  and the median, about 25% are between the median and  $Q_3$ , and about 25% are between  $Q_3$  and the maximum.

**Boxplot**

The five-number summary of a distribution leads to a new graph, the boxplot (sometimes called a box and whisker plot).

**How to Make a Boxplot**

- A central box is drawn from the first quartile ( $Q_1$ ) to the third quartile ( $Q_3$ ).
- A line in the box marks the median.
- Lines (called whiskers) extend from the box out to the smallest and largest observations that are not outliers.

Here's an example that shows how to make a boxplot.

**EXAMPLE****Home Run King****Making a boxplot**

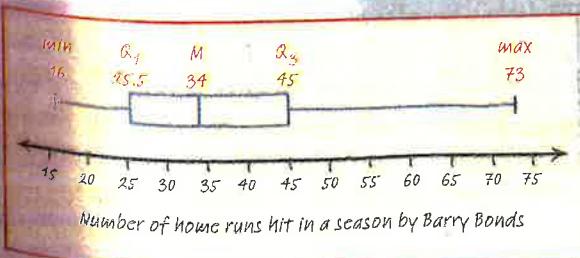
Barry Bonds set the major league record by hitting 73 home runs in a single season in 2001. On August 7, 2007, Bonds hit his 756th career home run, which broke Hank Aaron's longstanding record of 755. By the end of the 2007 season when Bonds retired, he had increased the total to 762. Here are data on the number of home runs that Bonds hit in each of his 21 complete seasons:

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

**PROBLEM:** We want to make a boxplot for these data.

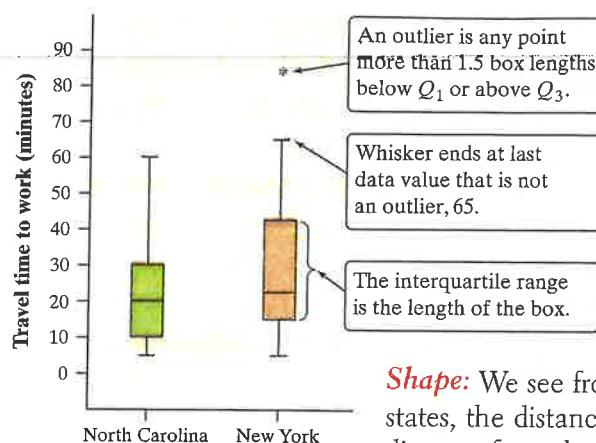
**SOLUTION:** Let's start by sorting the data values so that we can find the five-number summary.

16 19 24 25	25 26 28 33 33 34 34 37 37 40 42	45 45 46 46 49 73
Min	$Q_1 = 25.5$	$M = 42$
	$Q_3 = 45$	Max



Now we check for outliers. Since  $IQR = 45 - 25.5 = 19.5$ , by the  $1.5 \times IQR$  rule, any value greater than  $Q_3 + 1.5 \times IQR = 45 + 1.5 \times 19.5 = 74.25$  or less than  $Q_1 - 1.5 \times IQR = 25.5 - 1.5 \times 19.5 = -3.75$  would be classified as an outlier. So there are no outliers in this data set. Now we are ready to draw the boxplot. See the finished graph in the margin.

**For Practice Try Exercise 91**



**FIGURE 1.19** Boxplots comparing the travel times to work of samples of workers in North Carolina and New York.

Figure 1.19 shows boxplots (this time, they are oriented vertically) comparing travel times to work for the samples of workers from North Carolina and New York. We will identify outliers as isolated points in the boxplot (like the \* for the maximum value in the New York data set).

Boxplots show less detail than histograms or stemplots, so they are best used for side-by-side comparison of more than one distribution, as in Figure 1.19. As always, be sure to discuss shape, center, spread, and outliers as part of your comparison. For the travel time to work data:

**Shape:** We see from the graph that both distributions are right-skewed. For both states, the distance from the minimum to the median is much smaller than the distance from the median to the maximum.

**Center:** It appears that travel times to work are generally a bit longer in New York than in North Carolina. The median, both quartiles, and the maximum are all larger in New York.

**Spread:** Travel times are also more variable in New York, as shown by the heights of the boxes (the *IQR*) and the spread from smallest to largest time.

**Outliers:** Earlier, we showed that the maximum travel time of 85 minutes is an outlier for the New York data. There are no outliers in the North Carolina sample.

The following Activity reinforces the important ideas of working with boxplots.<sup>40</sup>

### ACTIVITY *Do you know your geography?*

How well do you know the U.S. state capitals? The locations of major rivers and mountain chains? The populations of states and cities? In this Activity, you'll have a chance to demonstrate your geographic knowledge by estimating the population of a capital city.

1. Your teacher will give each student a slip of paper with two questions about the population of a U.S. capital city. Look only at your paper. Answer the questions to the best of your ability.
2. Now your teacher will share the details of what was on the slips of paper. About half the students in the class received one version of the first question, and the other half received a different version. The second question was the same on all the slips.
3. Students in the class should divide into two groups based on which version of the first question they received. Within each group, line up in order of your population estimates, from smallest to largest. Then find the five-number summary for your estimates.
4. With the other members of your group, determine whether there are any outliers.
5. Have a representative from each group draw a boxplot of the group's population estimates on the board or overhead-projector sheet using the scale provided by your teacher. What do the side-by-side boxplots reveal about the two groups' estimates?
6. Your teacher will provide the capital city's actual population. How close was your estimate? Which group did better at estimating the city's population? How can you tell?





### CHECK YOUR UNDERSTANDING

The 2009 roster of the Dallas Cowboys professional football team included 10 offensive linemen. Their weights (in pounds) were

338 318 353 313 318 326 307 317 311 311

- Find the five-number summary for these data by hand. Show your work.
- Calculate the  $IQR$ . Interpret this value in context.
- Determine whether there are any outliers using the  $1.5 \times IQR$  rule.
- Draw a boxplot of the data.

### TECHNOLOGY CORNER Making calculator boxplots

The TI-83/84 and TI-89 can plot up to three boxplots in the same viewing window. Let's use the calculator to make side-by-side boxplots of the travel time to work data for the samples from North Carolina and New York.

- Enter the travel time data for North Carolina in L1/list1 and for New York in L2/list2.
- Set up two statistics plots: Plot1 to show a boxplot of the North Carolina data and Plot2 to show a boxplot of the New York data.

TI-83/84



TI-89

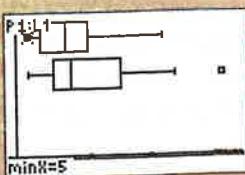


**Note:** The calculator offers two types of boxplots: a “modified” boxplot that shows outliers and a standard boxplot that doesn’t. We’ll always use the modified boxplot.

- Use the calculator’s Zoom feature to display the side-by-side boxplots. Then Trace to view the five-number summary.

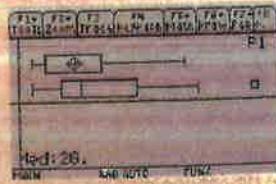
TI-83/84

- Press **ZOOM** and select 9 : ZoomStat.
- Press **TRACE**.

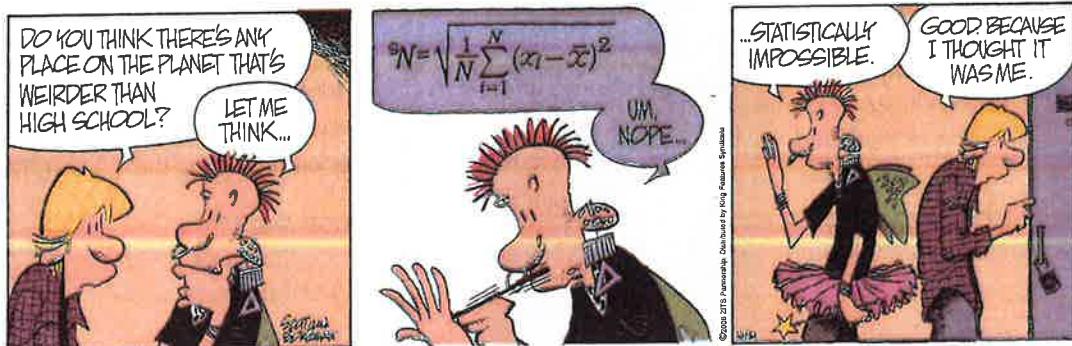


TI-89

- Press **F5** (ZoomData).
- Press **F3** (Trace).



## Measuring Spread: The Standard Deviation



The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread. The standard deviation and its close relative, the *variance*, measure spread by looking at how far the observations are from their mean. Let's explore this idea using a simple set of data.

### EXAMPLE

#### How Many Pets?

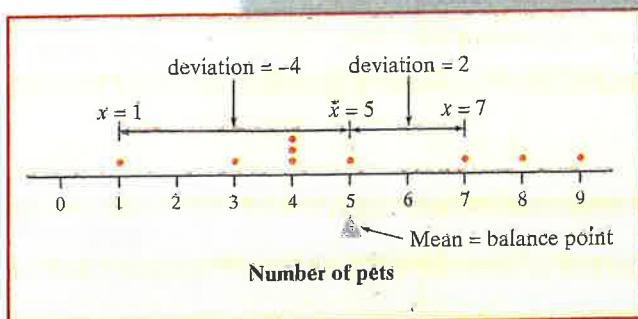
##### Investigating spread around the mean

In the Think About It on page 52, we examined data on the number of pets owned by a group of 9 children. Here are the data again, arranged from lowest to highest:

1 3 4 4 4 5 7 8 9

Earlier, we found the mean number of pets to be  $\bar{x} = 5$ . Let's look at where the observations in the data set are relative to the mean.

Figure 1.20 displays the data in a dotplot, with the mean clearly marked. The data value 1 is 4 units below the mean. We say that its *deviation* from the mean is  $-4$ . What about the data value 7? Its deviation is  $7 - 5 = 2$  (it is 2 units above the mean). The arrows in the figure mark these two deviations from the mean. The deviations show how much the data vary about their mean. They are the starting point for calculating the variance and standard deviation.



**FIGURE 1.20** Dotplot of the pet data with the mean and two of the deviations marked.

Since the sum of the deviations from the mean will be 0 for any set of data, we need another way to calculate spread around the mean. How can we fix the problem of the positive and negative deviations canceling out? We could take the absolute value of each deviation. Or we could square the deviations. For math-

ematical reasons beyond the scope of this book, statisticians choose to square rather than to use absolute values.

Observations $x_i$	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1	$1 - 5 = -4$	$(-4)^2 = 16$
3	$3 - 5 = -2$	$(-2)^2 = 4$
4	$4 - 5 = -1$	$(-1)^2 = 1$
4	$4 - 5 = -1$	$(-1)^2 = 1$
4	$4 - 5 = -1$	$(-1)^2 = 1$
5	$5 - 5 = 0$	$0^2 = 0$
7	$7 - 5 = 2$	$2^2 = 4$
8	$8 - 5 = 3$	$3^2 = 9$
9	$9 - 5 = 4$	$4^2 = 16$
	sum = ??	sum = ??

We have added a column to the table that shows the square of each deviation  $(x_i - \bar{x})^2$ . Add up the squared deviations. Did you get 52? Now we compute the average squared deviation—sort of. Instead of dividing by the number of observations  $n$ , we divide by  $n - 1$ :

$$\text{"average" squared deviation} = \frac{16 + 4 + 1 + 1 + 1 + 0 + 4 + 9 + 16}{9 - 1} = \frac{52}{8} = 6.5$$

This value, 6.5, is called the **variance**.

Because we squared all the deviations, our units are in “squared pets.” That’s no good. We’ll take the square root to get back to the correct units—pets. The resulting value is the **standard deviation**:

$$\text{standard deviation} = \sqrt{\text{"average" squared deviation}} = \sqrt{6.5} = 2.55$$

This 2.55 is roughly the average distance of the values in the data set from the mean.



As you can see, the average distance in the standard deviation is found in a rather unexpected way. Why do we divide by  $n - 1$  instead of  $n$  when calculating the variance and standard deviation? The answer is complicated but will be revealed in Chapter 7.

### DEFINITION: The standard deviation $s_x$ and variance $s_x^2$

The **standard deviation**  $s_x$  measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances and then taking the square root. This average squared distance is called the **variance**. In symbols, the variance  $s_x^2$  is given by

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

Here's a brief summary of the process for calculating the standard deviation.

### How to Find the Standard Deviation

To find the standard deviation of  $n$  observations:

1. Find the distance of each observation from the mean and square each of these distances.
2. Average the distances by dividing their sum by  $n - 1$ .
3. The standard deviation  $s_x$  is the square root of this average squared distance:

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

Many calculators report two standard deviations, giving you a choice of dividing by  $n$  or by  $n - 1$ . The former is usually labeled  $\sigma_x$ , the symbol for the standard deviation of a population. If your data set consists of the entire population, then it's appropriate to use  $\sigma_x$ . More often, the data we're examining come from a sample. In that case, we should use  $s_x$ .

More important than the details of calculating  $s_x$  are the properties that determine the usefulness of the standard deviation:

- $s_x$  measures *spread about the mean* and should be used only when the mean is chosen as the measure of center.
- $s_x$  is *always greater than or equal to 0*.  $s_x = 0$  only when there is no variability. This happens only when all observations have the same value. Otherwise,  $s_x > 0$ . As the observations become more spread out about their mean,  $s_x$  gets larger.
- $s_x$  has the *same units of measurement as the original observations*. For example, if you measure metabolic rates in calories, both the mean  $\bar{x}$  and the standard deviation  $s_x$  are also in calories. This is one reason to prefer  $s_x$  to the variance  $s_x^2$ , which is in squared calories.
- Like the mean  $\bar{x}$ ,  $s_x$  is *not resistant*. A few outliers can make  $s_x$  very large.

*The use of squared deviations makes  $s_x$  even more sensitive than  $\bar{x}$  to a few extreme observations.* For example, the standard deviation of the travel times for the 15 North Carolina workers is 15.23 minutes. If we omit the maximum value of 60 minutes, the standard deviation drops to 11.56 minutes.



### CHECK YOUR UNDERSTANDING



The heights (in inches) of the five starters on a basketball team are 67, 72, 76, 76, and 84.

1. Find and interpret the mean.
2. Make a table that shows, for each value, its deviation from the mean and its squared deviation from the mean.
3. Show how to calculate the variance and standard deviation from the values in your table.
4. Interpret the meaning of the standard deviation in this setting.

## Numerical Summaries with Technology

Graphing calculators and computer software will calculate numerical summaries for you. That will free you up to concentrate on choosing the right methods and interpreting your results.

### TECHNOLOGY CORNER Computing numerical summaries with technology

Let's find numerical summaries for the travel times of North Carolina and New York workers from the previous Technology Corner (page 61). We'll start by showing you the necessary calculator techniques and then look at output from computer software.

**I. One-variable statistics on the calculator** If you haven't done so already, enter the North Carolina data in L1/list1 and the New York data in L2/list2.

- Find the summary statistics for the North Carolina travel times.

TI-83/84

- Press **STAT** **►** (CALC); choose **1:1-VarStats**.
- Press **[ENTER]**. Now press **2nd** **1** (L1) and **[ENTER]**.

TI-89

- Press **F4** (Calc); choose **1:1-Var Stats**.
- Type **list1** in the list box. Press **[ENTER]**

Press **▼** to see the rest of the one-variable statistics for North Carolina.

**1-Var Stats**  
 $\bar{x}=22.46666667$   
 $\sum x=337$   
 $\sum x^2=10819$   
 $Sx=15.23092093$   
 $s_x=14.71445756$   
 $\downarrow n=15$

**1-Var Stats**  
 $n=15$   
 $\min x=5$   
 $Q_1=10$   
 $Med=20$   
 $Q_3=30$   
 $\max x=60$

**1-Var Stats**  
 $n=22$   
 $\bar{x}=22.46666667$   
 $\sum x=337$   
 $\sum x^2=10819$   
 $Sx=15.23092093$   
 $s_x=14.71445756$   
 $\min x=5$   
 $Q_1=10$   
 $Med=20$   
 $Q_3=30$   
 $\max x=60$

**1-Var Stats**  
 $n=22$   
 $\bar{x}=22.46666667$   
 $\sum x=337$   
 $\sum x^2=10819$   
 $Sx=15.23092093$   
 $s_x=14.71445756$   
 $\min x=5$   
 $Q_1=10$   
 $Med=20$   
 $Q_3=30$   
 $\max x=60$

- Repeat Step 1 using L2/list2 to find the summary statistics for the New York travel times.

**1-Var Stats**  
 $\bar{x}=31.25$   
 $\sum x=625$   
 $\sum x^2=28625$   
 $Sx=21.8773495$   
 $s_x=21.32340264$   
 $\downarrow n=20$

**1-Var Stats**  
 $n=20$   
 $\min x=5$   
 $Q_1=15$   
 $Med=22.5$   
 $Q_3=42.5$   
 $\max x=85$

**1-Var Stats**  
 $n=22$   
 $\bar{x}=31.25$   
 $\sum x=625$   
 $\sum x^2=28625$   
 $Sx=21.8773495$   
 $s_x=21.32340264$   
 $\min x=5$   
 $Q_1=15$   
 $Med=22.5$   
 $Q_3=42.5$   
 $\max x=85$

**1-Var Stats**  
 $n=22$   
 $\bar{x}=31.25$   
 $\sum x=625$   
 $\sum x^2=28625$   
 $Sx=21.8773495$   
 $s_x=21.32340264$   
 $\min x=5$   
 $Q_1=15$   
 $Med=22.5$   
 $Q_3=42.5$   
 $\max x=85$

**II. Output from statistical software** We used Minitab statistical software to produce descriptive statistics for the New York and North Carolina travel time data. Minitab allows you to choose which numerical summaries are included in the output.

#### Descriptive Statistics: Travel time to work

Variable	N	Mean	StDev	Minimum	Q <sub>1</sub>	Median	Q <sub>3</sub>	Maximum
NY Time	20	31.25	21.88	5.00	15.00	22.50	43.75	85.00
NC Time	15	22.47	15.23	5.00	10.00	20.00	30.00	60.00

TI-Nspire instructions in Appendix B

THINK  
ABOUT  
IT

What's with that third quartile? Earlier, we saw that the quartiles of the New York travel times are  $Q_1 = 15$  and  $Q_3 = 42.5$ . Look at the Minitab output in the Technology Corner. Minitab says that  $Q_3 = 43.75$ . What happened? Minitab and some other software use different rules for locating quartiles. Results from the

various rules are always close to each other, so the differences are rarely important in practice. But because of the slight difference, Minitab wouldn't identify the maximum value of 85 as an outlier by the  $1.5 \times IQR$  rule.

## Choosing Measures of Center and Spread

We now have a choice between two descriptions of the center and spread of a distribution: the median and  $IQR$ , or  $\bar{x}$  and  $s_x$ . Because  $\bar{x}$  and  $s_x$  are sensitive to extreme observations, they can be misleading when a distribution is strongly skewed or has outliers. In these cases, the median and  $IQR$ , which are both resistant to extreme values, provide a better summary. We'll see in the next chapter that the mean and standard deviation are the natural measures of center and spread for a very important class of symmetric distributions, the Normal distributions.

### Choosing Measures of Center and Spread

The median and  $IQR$  are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use  $\bar{x}$  and  $s_x$  only for reasonably symmetric distributions that don't have outliers.

*Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not highlight the presence of multiple peaks or clusters, for example. Always plot your data.*



Here's a final example that shows the thinking involved in choosing measures of center and spread when comparing two sets of quantitative data.

### EXAMPLE

#### Who Texts More—Males or Females?

Pulling it all together



For their final project, a group of AP Statistics students investigated their belief that females text more than males. They asked a random sample of students from their school to record the number of text messages sent and received over a two-day period. Here are their data:

Males:	127	44	28	83	0	6	78	6	5	213	73	20	214	28	11
Females:	112	203	102	54	379	305	179	24	127	65	41	27	298	6	130

What conclusion should the students draw? Give appropriate evidence to support your answer.

**STATE:** Do the data give convincing evidence that females text more than males?



**PLAN:** We'll begin by making side-by-side boxplots. Then we'll calculate one-variable statistics. Finally, we'll compare shape, center, spread, and outliers for the two distributions.

**DO:** Figure 1.21 is a sketch of the boxplots we obtained from our calculator. The table below shows numerical summaries for males and females.

	$\bar{x}$	$s_x$	Min	$Q_1$	$M$	$Q_3$	Max	$IQR$
Male	62.4	71.4	0	6	28	83	214	77
Female	128.3	116.0	0	34	107	191	379	157

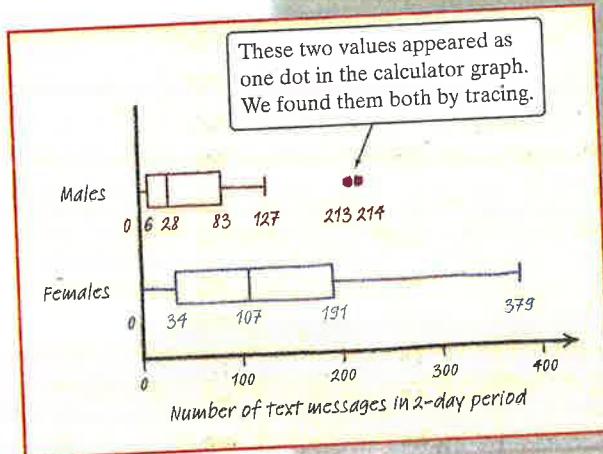


FIGURE 1.21 Side-by-side boxplots of the texting data.

evidence to support the students' belief that females text more than males. Females sent and received a median of 107 texts over the two-day period, which exceeded the number of texts reported by over 75% of the males.

**Due to the strong skewness and outliers, we'll use the median and  $IQR$  instead of the mean and standard deviation when discussing center and spread. Shape:** Both distributions are heavily right-skewed. **Center:** On average, females text more than males. The median number of texts for females (107) is about four times as high as for males (28). In fact, the median for the females is above the third quartile for the males. This indicates that over 75% of the males texted less than the "typical" (median) female. **Spread:** There is much more variability in texting among the females than the males. The  $IQR$  for females (157) is about twice the  $IQR$  for males (77). **Outliers:** There are two outliers in the male distribution: students who reported 213 and 214 texts in two days. The female distribution has no outliers.

**CONCLUDE:** The data from this survey project give very strong

### For Practice Try Exercise 105

**AP EXAM TIP** Use statistical terms carefully and correctly on the AP exam. Don't say "mean" if you really mean "median." Range is a single number; so are  $Q_1$ ,  $Q_3$ , and  $IQR$ . Avoid colloquial use of language, like "the outlier skews the mean." Skewed is a shape. If you misuse a term, expect to lose some credit.

Now it's time for you to put what you have learned into practice in the following Data Exploration.

### DATA EXPLORATION Did Mr. Starnes stack his class?

Mr. Starnes teaches AP Statistics, but he also does the class scheduling for the high school. There are two AP Statistics classes—one taught by Mr. Starnes and one taught by Ms. McGrail. The two teachers give the same first test to their classes and grade the test together. Mr. Starnes's students earned an average score that was 8 points higher than the average for Ms. McGrail's class. Ms. McGrail wonders whether Mr. Starnes might have "adjusted" the class rosters from the computer scheduling program. In other words, she thinks he might have "stacked" his class. He denies this, of course.

To help resolve the dispute, the teachers collect data on the cumulative grade point averages and SAT Math scores of their students. Mr. Starnes provides the GPA data from his computer. The students report their SAT Math scores. The tables below show the data for each student in the two classes. Note that the two data values in each row come from a single student.

Starnes GPA	Starnes SAT-M	McGrail GPA	McGrail SAT-M
2.9	670	2.9	620
2.86	520	3.3	590
2.6	570	3.98	650
3.6	710	2.9	600
3.2	600	3.2	620
2.7	590	3.5	680
3.1	640	2.8	500
3.085	570	2.9	502.5
3.75	710	3.95	640
3.4	630	3.1	630
3.338	630	2.85	580
3.56	670	2.9	590
3.8	650	3.245	600
3.2	660	3.0	600
3.1	510	3.0	620
		2.8	580
		2.9	600
		3.2	600

Did Mr. Starnes stack his class? Give appropriate graphical and numerical evidence to support your conclusion.

At the end of each chapter, we'll return to the chapter-opening Case Study to let you know what happened. We call this feature "Case Closed!"



case closed

## Do Rewards Promote Creativity?

Refer to the chapter-opening Case Study (page 1). In Dr. Amabile's experiment, the children who believed they were competing for a prize produced *less creative* collages than the children who were told that they would share their finished collages at an art party. As a result of this study, Dr. Amabile argued that external rewards—things like money, grades, and fame—actually reduce creativity. She went on to conduct several more experiments, like the one investigating the effects of internal and external motivation on college students' creative writing (Exercise 43, page 43). Her findings were remarkably consistent: for most people, creativity is its own reward.

## SECTION 1.3

## Summary

- A numerical summary of a distribution should report at least its **center** and its **spread**, or **variability**.
- The **mean**  $\bar{x}$  and the **median**  $M$  describe the center of a distribution in different ways. The mean is the average of the observations, and the median is the midpoint of the values.
- When you use the median to indicate the center of a distribution, describe its spread using the **quartiles**. The first quartile  $Q_1$  has about one-fourth of the observations below it, and the third quartile  $Q_3$  has about three-fourths of the observations below it. The **interquartile range (IQR)** is the range of the middle 50% of the observations and is found by  $IQR = Q_3 - Q_1$ . An extreme observation is an **outlier** if it is smaller than  $Q_1 - (1.5 \times IQR)$  or larger than  $Q_3 + (1.5 \times IQR)$ .
- The **five-number summary** consisting of the median, the quartiles, and the maximum and minimum values provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.
- **Boxplots** based on the five-number summary are useful for comparing distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the smallest and the largest observations that are not outliers. Outliers are plotted as isolated points.
- The **variance**  $s_x^2$  and especially its square root, the **standard deviation**  $s_x$ , are **common** measures of spread about the mean as center. The standard deviation  $s_x$  is zero when there is no variability and gets larger as the spread increases.
- The median is a **resistant** measure of center because it is relatively unaffected by extreme observations. The mean is **nonresistant**. Among measures of spread, the **IQR** is resistant, but the standard deviation is not.
- The mean and standard deviation are **good** descriptions for **symmetric** distributions without outliers. They are **most useful** for the **Normal** distributions introduced in the next chapter. The median and **IQR** are a better description for skewed distributions.
- Numerical summaries do not fully describe the shape of a distribution.  
*Always plot your data.*

## 1.3 TECHNOLOGY CORNERS

Making calculator boxplots ..... page 61  
 Computing numerical summaries with technology ..... page 65

TI-Nspire instructions in Appendix B

## SECTION 1.3

## Exercises

- 79. Quiz grades** Joey's first 14 quiz grades in a marking period were

pg 51

86	84	91	75	78	80	74
87	76	96	82	90	98	93

Calculate the mean. Show your work. Interpret your result in context.

- 80. Cowboys** The 2009 roster of the Dallas Cowboys professional football team included 7 defensive linemen. Their weights (in pounds) were 306, 305, 315, 303, 318, 309, and 285. Calculate the mean. Show your work. Interpret your result in context.

pg 53

- 81. Quiz grades** Refer to Exercise 79.

- (a) Find the median by hand. Show your work. Interpret your result in context.  
 (b) Suppose Joey has an unexcused absence for the 15th quiz, and he receives a score of zero. Recalculate the mean and the median. What property of measures of center does this illustrate?

- 82. Cowboys** Refer to Exercise 80.

- (a) Find the median by hand. Show your work. Interpret your result in context.  
 (b) Suppose the lightest lineman had weighed 265 pounds instead of 285 pounds. How would this change affect the mean and the median? What property of measures of center does this illustrate?

- 83. Incomes of college grads** According to the Census Bureau, the mean and median 2008 income of people at least 25 years old who had a bachelor's degree but no higher degree were \$48,097 and \$60,954. Which of these numbers is the mean and which is the median? Explain your reasoning.

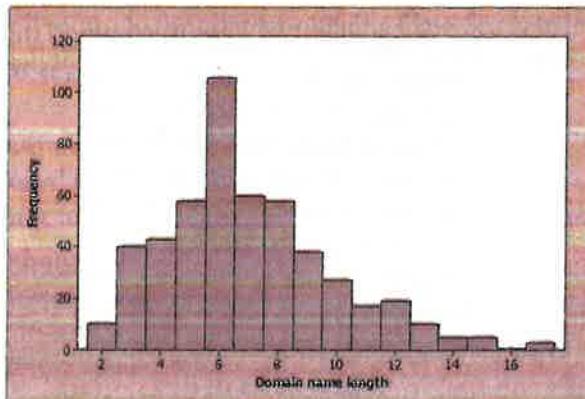
- 84. House prices** The mean and median selling prices of existing single-family homes sold in November 2009 were \$216,400 and \$172,600.<sup>41</sup> Which of these numbers is the mean and which is the median? Explain how you know.

- 85. Baseball salaries** Suppose that a Major League Baseball team's mean yearly salary for its players is \$1.2 million and that the team has 25 players on its active roster. What is the team's total annual payroll?

If you knew only the median salary, would you be able to answer this question? Why or why not?

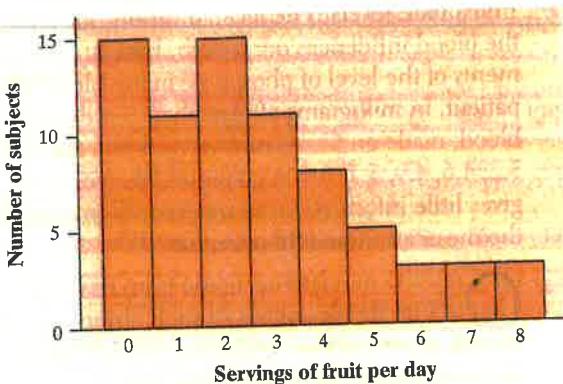
- 86. Mean salary?** Last year a small accounting firm paid each of its five clerks \$22,000, two junior accountants \$50,000 each, and the firm's owner \$270,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary? Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

- 87. Domain names** When it comes to Internet domain names, is shorter better? According to one ranking of Web sites in 2008, the top 8 sites (by number of "hits") were yahoo.com, google.com, youtube.com, live.com, msn.com, myspace.com, wikipedia.org, and facebook.com. These familiar sites certainly have short domain names. The histogram below shows the domain name lengths (in number of letters in the name, not including the extensions .com and .org) for the 500 most popular Web sites.



- (a) Estimate the mean and median of the distribution. Explain your method clearly.  
 (b) If you wanted to argue that shorter domain names were more popular, which measure of center would you choose—the mean or the median? Justify your answer.

- 88. Do adolescent girls eat fruit?** We all know that fruit is good for us. Below is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.<sup>42</sup>



- (a) With a little care, you can find the median and the quartiles from the histogram. What are these numbers? How did you find them?

- (b) Estimate the mean of the distribution. Explain your method clearly.

89. Quiz grades Refer to Exercise 79.

pg 57, 58

- (a) Find and interpret the interquartile range (IQR).  
(b) Determine whether there are any outliers. Show your work.

90. Cowboys Refer to Exercise 80.

- (a) Find and interpret the interquartile range (IQR).  
(b) Determine whether there are any outliers. Show your work.

91. Don't call me In a September 28, 2008, article titled "Letting Our Fingers Do the Talking," the *New York Times* reported that Americans now send more text messages than they make phone calls. According to a study by Nielsen Mobile, "Teenagers ages 13 to 17 are by far the most prolific texters, sending or receiving 1,742 messages a month." Mr. Williams, a high school statistics teacher, was skeptical about the claims in the article. So he collected data from his first-period statistics class on the number of text messages and calls they had sent or received in the past 24 hours. Here are the texting data:

0	7	1	29	25	8	5	1	25	98	9	0	26
8	118	72	0	92	52	14	3	3	44	5	42	

- (a) Make a boxplot of these data by hand. Be sure to check for outliers.  
(b) Do these data support the claim in the article about the number of texts sent by teens? Justify your answer with appropriate evidence.

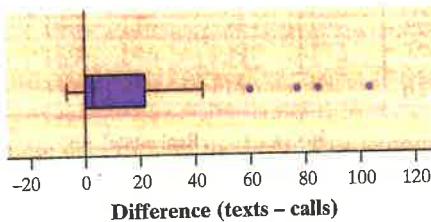
92. Acing the first test Here are the scores of Mrs. Liao's students on their first statistics test:

93	93	87.5	91	94.5	72	96	95	93.5	93.5	73
82	45	88	80	86	85.5	87.5	81	78	86	89
92	91	98	85	82.5	88	94.5	43			

- (a) Make a boxplot of the test score data by hand. Be sure to check for outliers.

- (b) How did the students do on Mrs. Liao's first test? Justify your answer.

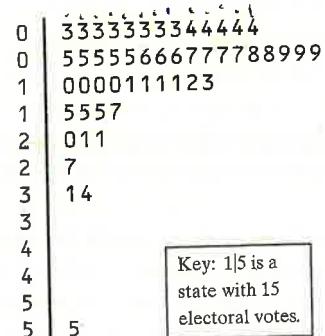
93. Texts or calls? Refer to Exercise 91. A boxplot of the difference (texts – calls) in the number of texts and calls for each student is shown below.



- (a) Do these data support the claim in the article about texting versus calling? Justify your answer with appropriate evidence.

- (b) Can we draw any conclusion about the preferences of all students in the school based on the data from Mr. Williams's statistics class? Why or why not?

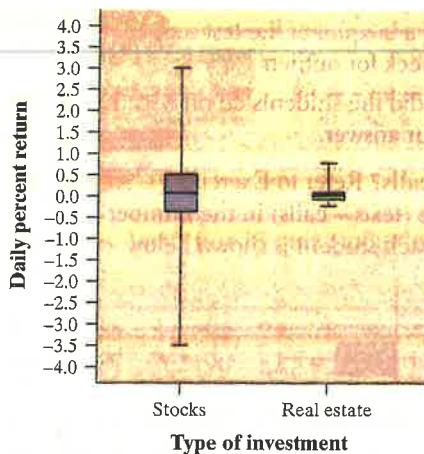
94. Electoral votes To become president of the United States, a candidate does not have to receive a majority of the popular vote. The candidate does have to win a majority of the 538 electoral votes that are cast in the Electoral College. Here is a stemplot of the number of electoral votes for each of the 50 states and the District of Columbia.



- (a) Make a boxplot of these data by hand. Be sure to check for outliers.

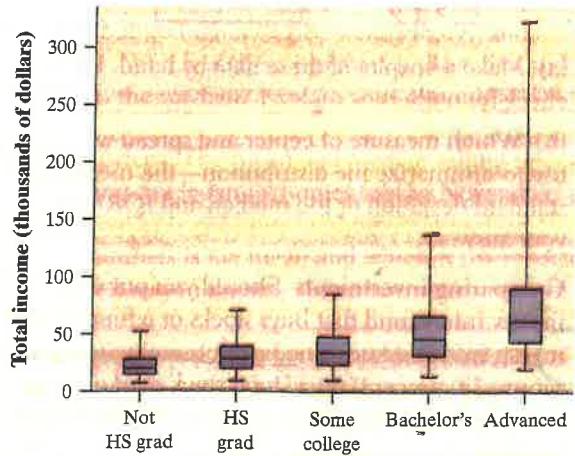
- (b) Which measure of center and spread would you use to summarize the distribution—the mean and standard deviation or the median and IQR? Justify your answer.

95. Comparing investments Should you put your money into a fund that buys stocks or a fund that invests in real estate? The boxplots compare the daily returns (in percent) on a "total stock market" fund and a real estate fund over a year ending in November 2007.<sup>43</sup>



- (a) Read the graph: about what were the highest and lowest daily returns on the stock fund?
- (b) Read the graph: the median return was about the same on both investments. About what was the median return?
- (c) What is the most important difference between the two distributions?

**96. Income and education level** Each March, the Bureau of Labor Statistics compiles an Annual Demographic Supplement to its monthly Current Population Survey.<sup>44</sup> Data on about 71,067 individuals between the ages of 25 and 64 who were employed full-time were collected in one of these surveys. The boxplots below compare the distributions of income for people with five levels of education. This figure is a variation of the boxplot idea: because large data sets often contain very extreme observations, we omitted the individuals in each category with the top 5% and bottom 5% of incomes. Write a brief description of how the distribution of income changes with the highest level of education reached. Give specifics from the graphs to support your statements.



**97. Phosphate levels** The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic: 5.6, 5.2, 4.6, 4.9, 5.7, 6.4. A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

(a) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation.

(b) Interpret the value of  $s_x$  you obtained in (a).

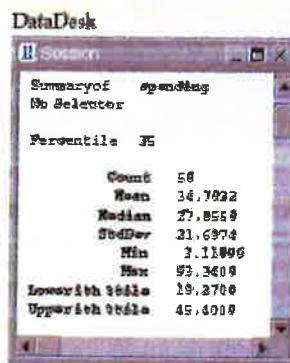
**98. Feeling sleepy?** The first four students to arrive for a first-period statistics class were asked how much sleep (to the nearest hour) they got last night. Their responses were 7, 7, 9, and 9.

(a) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation.

(b) Interpret the value of  $s_x$  you obtained in (a).

(c) Do you think it's safe to conclude that the mean amount of sleep for all 30 students in this class is close to 8 hours? Why or why not?

**99. Shopping spree** The figure displays computer output from Data Desk for data on the amount spent by 50 grocery shoppers.



(a) What would you guess is the shape of the distribution based only on the computer output? Explain.

(b) Interpret the value of the standard deviation.

(c) Are there any outliers? Justify your answer.

**100. C-sections** Do male doctors perform more cesarean sections (C-sections) than female doctors? A study in Switzerland examined the number of cesarean sections (surgical deliveries of babies) performed in a

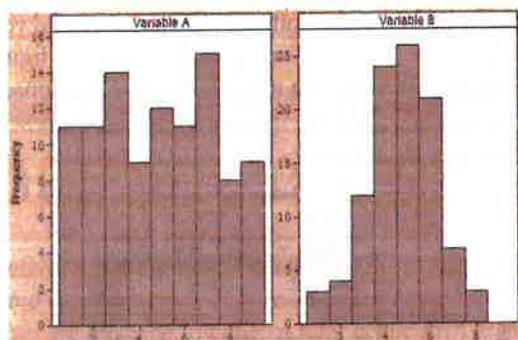
year by samples of male and female doctors. Here are summary statistics for the two distributions:

	$\bar{x}$	$s_x$	Min.	$Q_1$	$M$	$Q_3$	Max.	IQR
Male doctors	41.333	20.607	20	27	34	50	86	23
Female doctors	19.1	10.126	5	10	18.5	29	33	19

- (a) Based on the computer output, which distribution would you guess has a more symmetrical shape? Explain.  
 (b) Explain how the IQRs of these two distributions can be so similar even though the standard deviations are quite different.  
 (c) Does it appear that males perform more C-sections? Justify your answer.

101. The IQR Is the interquartile range a resistant measure of spread? Give an example of a small data set that supports your answer.

102. Measuring spread Which of the distributions shown has a larger standard deviation? Justify your answer.



103. SD contest This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

- (a) Choose four numbers that have the smallest possible standard deviation.  
 (b) Choose four numbers that have the largest possible standard deviation.  
 (c) Is more than one choice possible in either (a) or (b)? Explain.

104. What do they measure? For each of the following summary statistics, decide (i) whether it could be used to measure center or spread and (ii) whether it is resistant.

(a)  $\frac{Q_1 + Q_3}{2}$       (b)  $\frac{\text{Max} - \text{Min}}{2}$

105. SSHA scores Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

pg 66



154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

and for 20 first-year college men:

108	140	114	91	180	115	126
92	169	146	109	132	75	88
113	151	70	115	187	104	

Do these data support the belief that women have better study habits and attitudes toward learning than men? (Note that high scores indicate good study habits and attitudes toward learning.) Follow the four-step process.



106. Hummingbirds and tropical flowers Researchers from Amherst College studied the relationship between varieties of the tropical flower *Heliconia* on the island of Dominica and the different species of hummingbirds that fertilize the flowers.<sup>45</sup> Over time, the researchers believe, the lengths of the flowers and the forms of the hummingbirds' beaks have evolved to match each other. If that is true, flower varieties fertilized by different hummingbird species should have distinct distributions of length.

The table below gives length measurements (in millimeters) for samples of three varieties of *Heliconia*, each fertilized by a different species of hummingbird. Do these data support the researchers' belief? Follow the four-step process.

#### *H. bihai*

47.12	46.75	46.80	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36

#### *H. caribaea red*

41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	

#### *H. caribaea yellow*

36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.10
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

Multiple choice: Select the best answer for Exercises 107 to 110.

107. If a distribution is skewed to the right with no outliers,
- (a) mean < median.
  - (b) mean ≈ median.
  - (c) mean = median.
  - (d) mean > median.
  - (e) We can't tell without examining the data.

- 108** You have data on the weights in grams of 5 baby pythons. The mean weight is 31.8 and the standard deviation of the weights is 2.39. The correct units for the standard deviation are

- (a) no units—it's just a number.
- (b) grams.
- (c) grams squared.
- (d) pythons.
- (e) pythons squared.

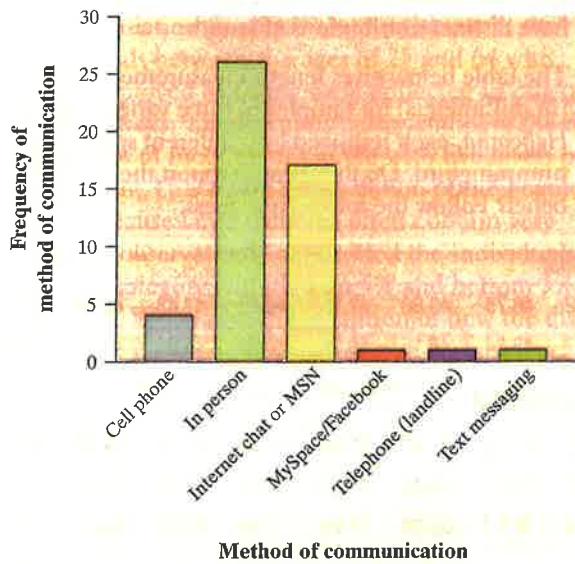
- 109** Which of the following is least affected if an extreme high outlier is added to your data?

- (a) Median
- (b) Mean
- (c) Standard deviation
- (d) Range
- (e) Maximum

- 110** What are all the values that a standard deviation  $s_x$  can possibly take?

- (a)  $s_x \geq 0$
- (b)  $s_x > 0$
- (c)  $0 \leq s_x \leq 1$
- (d)  $-1 \leq s_x \leq 1$
- (e) Any number

Exercises 111 and 112 refer to the following setting. We used CensusAtSchool's "Random Data Selector" to choose a sample of 50 Canadian students who completed a survey in 2007–2008.



- 111** Let's chat (1.1) The bar graph displays data on students' responses to the question "Which of these methods do you most often use to communicate with your friends?"

- (a) Would it be appropriate to make a pie chart for these data? Why or why not?

- (b) Jerry says that he would describe this bar graph as skewed to the right. Explain why Jerry is wrong.

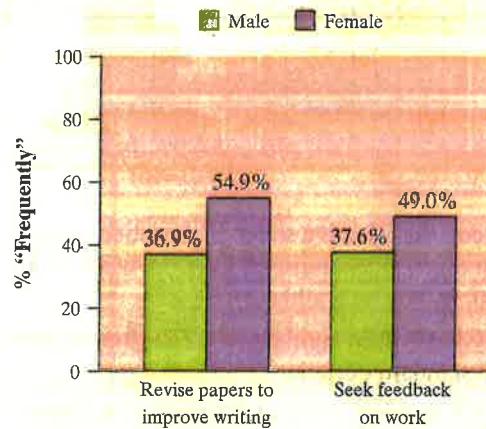
- 112** How tall are you? (1.2) Here are the heights (in centimeters) of 50 randomly selected Canadian students who participated in CensusAtSchool in 2007–2008:

166.5	170	178	163	150.5	169	173	169	171	166
190	183	178	161	171	170	191	168.5	178.5	173
175	160.5	166	164	163	174	160	174	182	167
166	170	170	181	171.5	160	178	157	165	187
168	157.5	145.5	156	182	168.5	177	162.5	160.5	185.5

Make an appropriate graph to display these data.

Describe the shape, center, and spread of the distribution. Are there any outliers?

- 113** Success in college (1.1) The 2007 Freshman Survey asked first-year college students about their "habits of mind"—specific behaviors that college faculty have identified as being important for student success. One question asked students, "How often in the past year did you revise your papers to improve your writing?" Another asked, "How often in the past year did you seek feedback on your academic work?" The figure is a bar graph comparing male and female responses to these two questions.<sup>46</sup>



What does the graph tell us about the habits of mind of male and female college freshmen?