AP Statistics Summer Assignment

Welcome to AP Statistics. I am looking forward to teaching you and learning from you!

We will speak at length about Statistics and its relevance when we meet in September.

Meanwhile, I would like you to get you started with some reading and exercises. As you read the assigned text, get into the habit of taking your own notes. These notes will be valuable to you when doing the assigned exercises and when studying for tests.

Do a little bit at a time. It becomes overwhelming if you leave it to the last minute.

Below is the assignment. You will also be getting first chapter of our AMAZING textbook: The Practice of Statistics. And guess what else? I am attaching solutions to the assigned exercises.

After you are done doing the exercises, you should check your work against my solutions. Write down all the questions you still have or anything that you don't understand. I will be happy to go over and answer those questions and clear any misunderstandings when we meet in September.

Have an amazing summer,

Dina Klapper

AP Statistics Summer Assignment

Read Pages 2-7
To Do: Page 7 # 1,3,5,7,9,10

Read Section 1.1: Pages 9-22
To Do: Page 24 # 13,15,17

Read Section 1.2: Pages 30-46
To Do: Page 47 # 45, 49, 51, 55, 59, 63, 65, 69, 77, 80-85

Read Section 1.3: Pages 54-74
To Do: Page 47 # 87,89,91,95,97,101,109,11,113,115, 123-126

Chapter 1

Data Analysis

AfriPics.com/Alamy Stock Photo

INTRODUCTION Statistics: The Science and Art of Data

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

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

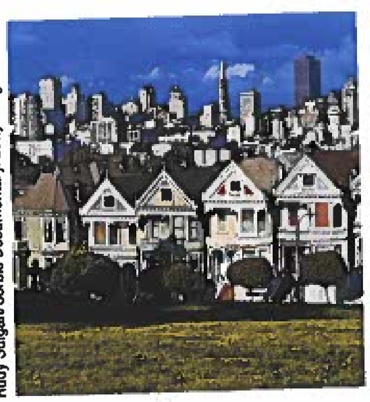
We live in a world of *data*. Every day, the media report poll results, outcomes of medical studies, and analyses of data on everything from stock prices to standardized test scores to global warming. The data are trying to tell us a story. To understand what the data are saying, you need to learn more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

DEFINITION Statistics

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

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

Organizing Data


Rudy Sulgan/Corbis Documentary/Getty Images

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

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

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

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

DEFINITION Individual, Variable

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

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

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

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

DEFINITION Categorical variable, Quantitative variable

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

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

Caution

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

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

EXAMPLE

Census At School
Individuals and Variables

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

Garry Black/Alamy

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

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

SOLUTION:

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

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

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

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

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

FOR PRACTICE, TRY EXERCISE 1

AP® EXAM TIP

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

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

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

DEFINITION Distribution

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

Let's return to the data for the sample of 10 Canadian students from the preceding example. Figure 1.1(a) shows the distribution of preferred communication method for these students in a *bar graph*. We can see how many students chose each method from the heights of the bars: cell phone (2), Facebook (1), in person (3), text messaging (4). Figure 1.1(b) shows the distribution of number of languages spoken in a *dotplot*. We can see that 6 students speak one language, 3 students speak two languages, and 1 student speaks four languages.

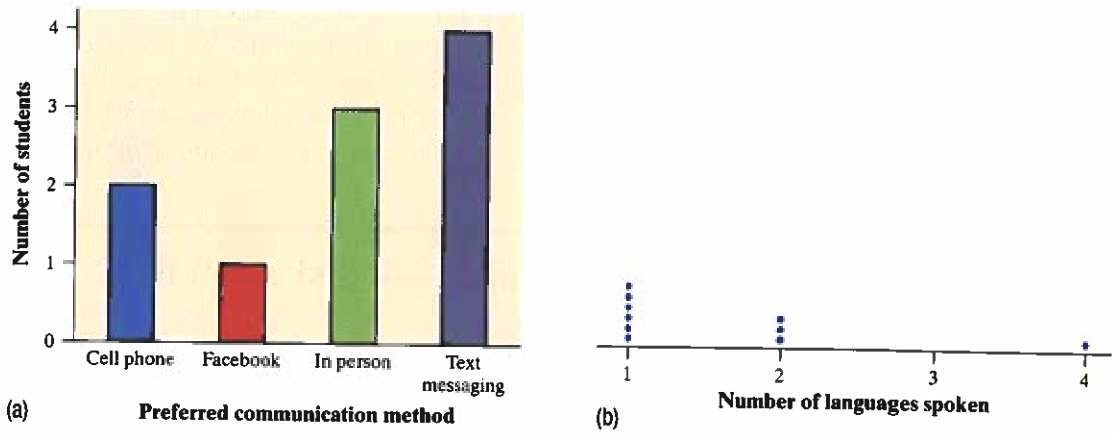


FIGURE 1.1 (a) Bar graph showing the distribution of preferred communication method for the sample of 10 Canadian students. (b) Dotplot showing the distribution of number of languages spoken by these students.

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

HOW TO ANALYZE DATA

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

CHECK YOUR UNDERSTANDING

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

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

From Data Analysis to Inference

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

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

ACTIVITY Hiring discrimination—it just won't fly!

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


Choja/Getty Images

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

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

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

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

Introduction Summary

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

Introduction Exercises

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

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

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

(a) Identify the individuals in this data set.

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

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

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

(a) Identify the individuals in this data set.

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

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

Movie	Year	Rating	Time (min)	Genre	Box office ($)
Avatar	2009	PG-13	162	Action	2,783,918,982
Titanic	1997	PG-13	194	Drama	2,207,615,668
Star Wars: The Force Awakens	2015	PG-13	136	Adventure	2,040,375,795
Jurassic World	2015	PG-13	124	Action	1,669,164,161
Marvel's The Avengers	2012	PG-13	142	Action	1,519,479,547
Furious 7	2015	PG-13	137	Action	1,516,246,709
The Avengers: Age of Ultron	2015	PG-13	141	Action	1,404,705,868
Harry Potter and the Deathly Hallows: Part 2	2011	PG-13	130	Fantasy	1,328,111,219
Frozen	2013	PG	108	Animation	1,254,512,386
Iron Man 3	2013	PG-13	129	Action	1,172,805,920

4. **Skyscrapers** Here is some information about the tallest buildings in the world as of February 2017. Identify the individuals and variables in this data set. Classify each variable as categorical or quantitative.

Building	Country	Height (m)	Floors	Use	Year completed
Burj Khalifa	United Arab Emirates	828.0	163	Mixed	2010
Shanghai Tower	China	632.0	121	Mixed	2014
Makkah Royal Clock Tower Hotel	Saudi Arabia	601.0	120	Hotel	2012
Ping An Finance Center	China	599.0	115	Mixed	2016
Lotte World Tower	South Korea	554.5	123	Mixed	2016
One World Trade Center	United States	541.0	104	Office	2013
Taipei 101	Taiwan	509.0	101	Office	2004
Shanghai World Financial Center	China	492.0	101	Mixed	2008
International Commerce Center	China	484.0	118	Mixed	2010
Petronas Tower 1	Malaysia	452.0	88	Office	1998

5. **Protecting wood** What measures can be taken, especially when restoring historic wooden buildings, to help wood surfaces resist weathering? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Some of the variables recorded were 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). Classify each variable as categorical or quantitative.

6. **Medical study variables** Data from a medical study contain values of many variables for each subject in the study. Some of the variables recorded were 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). Classify each variable as categorical or quantitative.

7. **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.

8. **Social media** You are preparing to study the social media habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student.

Multiple Choice: *Select the best answer.*

Exercises 9 and 10 refer to the following setting. At the Census Bureau website www.census.gov, you can view detailed data collected by the American Community Survey. The following table includes data for 10 people chosen at random from the more than 1 million people in households contacted by the survey. "School" gives the highest level of education completed.

Weight (lb)	Age (years)	Travel to work (min)	School	Gender	Income last year ($)
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	6000
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	2500
193	11	n/a	Fifth grade	1	0

9. The individuals in this data set are

(a) households. (b) people. (c) adults.

(d) 120 variables. (e) columns.

10. This data set contains

(a) 7 variables, 2 of which are categorical.

(b) 7 variables, 1 of which is categorical.

(c) 6 variables, 2 of which are categorical.

(d) 6 variables, 1 of which is categorical.

(e) None of these.

SECTION 1.1 Analyzing Categorical Data

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

- Make and interpret bar graphs for categorical data.
- Identify what makes some graphs of categorical data misleading.
- Calculate marginal and joint relative frequencies from a two-way table.
- Calculate conditional relative frequencies from a two-way table.
- Use bar graphs to compare distributions of categorical data.
- Describe the nature of the association between two categorical variables.

Here are the data on preferred communication method for the 10 randomly selected Canadian students from the example on page 3:

In person	In person	Facebook	Cell phone	In person
Text messaging	Cell phone	Text messaging	Text messaging	Text messaging

We can summarize the distribution of this categorical variable with a **frequency table** or a **relative frequency table.**

Some people use the terms *frequency distribution* and *relative frequency distribution* instead.

DEFINITION Frequency table, Relative frequency table

A **frequency table** shows the number of individuals having each value.

A **relative frequency table** shows the proportion or percent of individuals having each value.

To make either kind of table, start by tallying the number of times that the variable takes each value. **Note that the frequencies and relative frequencies listed in these tables are not data.** The tables summarize the data by telling us how many (or what proportion or percent of) students in the sample said "Cell phone," "Facebook," "In person," and "Text messaging."

Preferred method	Tally
Cell phone	II
Facebook	I
In person	III
Text messaging	IIII

Frequency table

Preferred method	Frequency
Cell phone	2
Facebook	1
In person	3
Text messaging	4

Relative frequency table

Preferred method	Relative frequency
Cell phone	2/10 = 0.20 or 20%
Facebook	1/10 = 0.10 or 10%
In person	3/10 = 0.30 or 30%
Text messaging	4/10 = 0.40 or 40%

The same process can be used to summarize the distribution of a quantitative variable. Of course, it would be hard to make a frequency table or a relative frequency table for quantitative data that take many different values, like the ages of people attending a Major League Baseball game. We'll look at a better option for quantitative variables with many possible values in Section 1.2.

Displaying Categorical Data: Bar Graphs and Pie Charts

A frequency table or relative frequency table summarizes a variable's distribution with numbers. To display the distribution more clearly, use a graph. You can make a **bar graph** or a **pie chart** for categorical data.

Bar graphs are sometimes called *bar charts*. Pie charts are sometimes called *circle graphs*.

DEFINITION Bar graph, Pie chart

A **bar graph** shows each category as a bar. The heights of the bars show the category frequencies or relative frequencies.

A **pie chart** shows each category as a slice of the "pie." The areas of the slices are proportional to the category frequencies or relative frequencies.

Figure 1.2 shows a bar graph and a pie chart of the data on preferred communication method for the random sample of Canadian students. Note that the percents for each category come from the relative frequency table.

Relative frequency table

Preferred method	Relative frequency
Cell phone	2/10 = 0.20 or 20%
Facebook	1/10 = 0.10 or 10%
In person	3/10 = 0.30 or 30%
Text messaging	4/10 = 0.40 or 40%

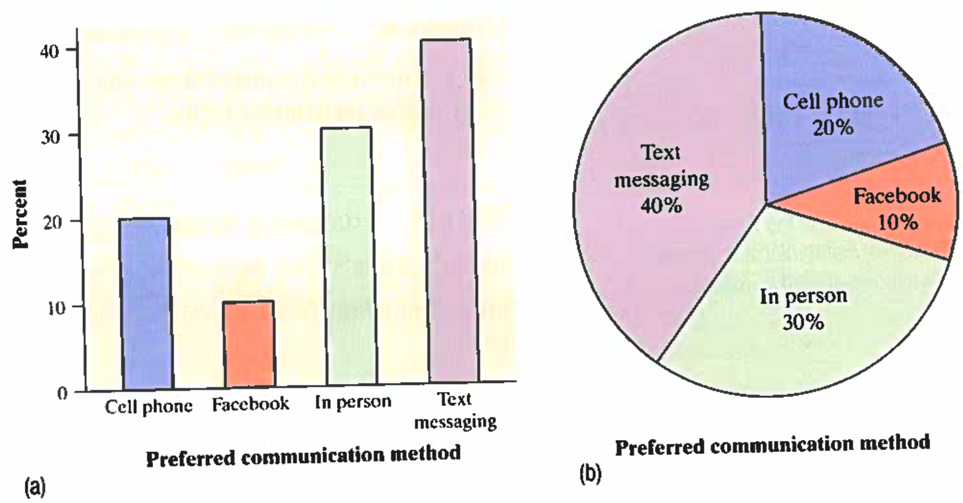


FIGURE 1.2 (a) Bar graph and (b) pie chart of the distribution of preferred communication method for a random sample of 10 Canadian students.

It is fairly easy to make a bar graph by hand. Here's how you do it.

HOW TO MAKE A BAR GRAPH

- **Draw and label the axes.** Put the name of the categorical variable under the horizontal axis. To the left of the vertical axis, indicate whether the graph shows the frequency (count) or relative frequency (percent or proportion) of individuals in each category.
- **"Scale" the axes.** Write the names of the categories at equally spaced intervals under the horizontal axis. On the vertical axis, start at 0 and place tick marks at equal intervals until you exceed the largest frequency or relative frequency in any category.
- **Draw bars above the category names.** Make the bars equal in width and leave gaps between them. Be sure that the height of each bar corresponds to the frequency or relative frequency of individuals in that category.

Making a graph is not an end in itself. The purpose of a graph is to help us understand the data. When looking at a graph, always ask, "What do I see?" We can see from both graphs in Figure 1.2 that the most preferred communication method for these students is text messaging.

EXAMPLE What's on the radio?

Making and interpreting bar graphs

PROBLEM: Arbitron, the rating service for radio audiences, categorizes U.S. radio stations in terms of the kinds of programs they broadcast. The frequency table summarizes the distribution of station formats in a recent year.[2]

Format	Number of stations	Format	Number of stations
Adult contemporary	2536	Religious	3884
All sports	1274	Rock	1636
Contemporary hits	1012	Spanish language	878
Country	2893	Variety	1579
News/Talk/Information	4077	Other formats	4852
Oldies	831	**Total**	**25,452**

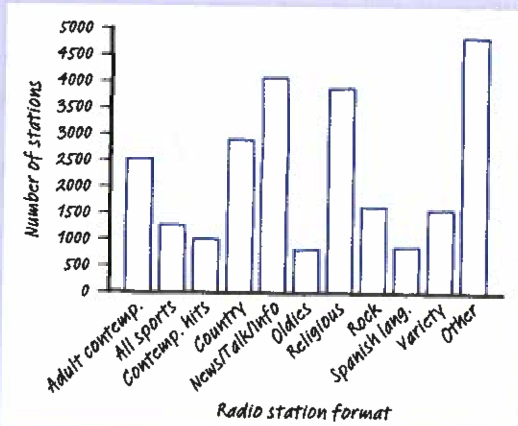
(a) Identify the individuals in this data set.

(b) Make a frequency bar graph of the data. Describe what you see.

SOLUTION:

(a) U.S. radio stations

(b)

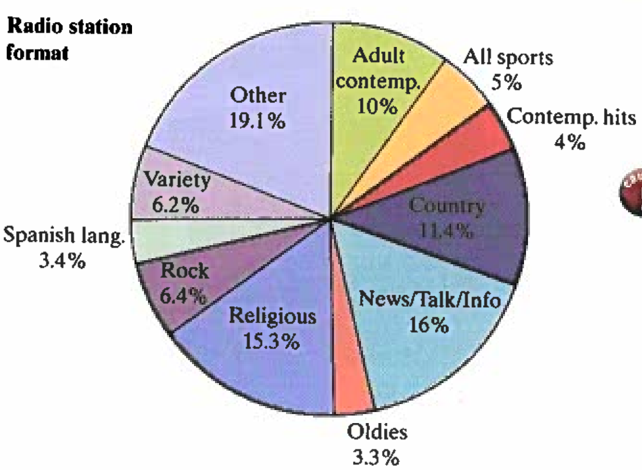


To make the bar graph:

- **Draw and label the axes.**
- **"Scale" the axes.** The largest frequency is 4852. So we choose a vertical scale from 0 to 5000, with tick marks 500 units apart.
- **Draw bars above the category names.**

On U.S. radio stations, the most frequent formats are Other (4852), News/talk/information (4077), and Religious (3884), while the least frequent are Oldies (831), Spanish language (878), and Contemporary hits (1012).

FOR PRACTICE, TRY EXERCISE 11



Here is a pie chart of the radio station format data from the preceding example. You can use a pie chart when you want to emphasize each category's relation to the whole. Pie charts are challenging to make by hand, but technology will do the job for you. Note that **a pie chart must include all categories that make up a whole,** which might mean adding an "other" category, as in the radio station example.

CHECK YOUR UNDERSTANDING

The American Statistical Association sponsors a web-based project that collects data about primary and secondary school students using surveys. We used the site's "Random Sampler" to choose 40 U.S. high school students who completed the survey in a recent year.[3] One of the questions asked:

Which would you prefer to be? Select one.

_____ Rich _____ Happy _____ Famous _____ Healthy

Here are the responses from the 40 randomly selected students:

Famous	Healthy	Healthy	Famous	Happy	Famous	Happy	Happy	Famous
Rich	Happy	Happy	Rich	Happy	Happy	Happy	Rich	Happy
Famous	Healthy	Rich	Happy	Happy	Rich	Happy	Happy	Rich
Healthy	Happy	Happy	Rich	Happy	Happy	Rich	Happy	Famous
Famous	Happy	Happy	Happy					

Make a relative frequency bar graph of the data. Describe what you see.

Graphs: Good and Bad

Bar graphs are a bit dull to look at. It is tempting to replace the bars with pictures or to use special 3-D effects to make the graphs seem more interesting. Don't do it! Our eyes react to the area of the bars as well as to their height. When all bars have the same width, the area (width × height) varies in proportion to the height, and our eyes receive the right impression about the quantities being compared.

EXAMPLE Who buys iMacs? Beware the pictograph!

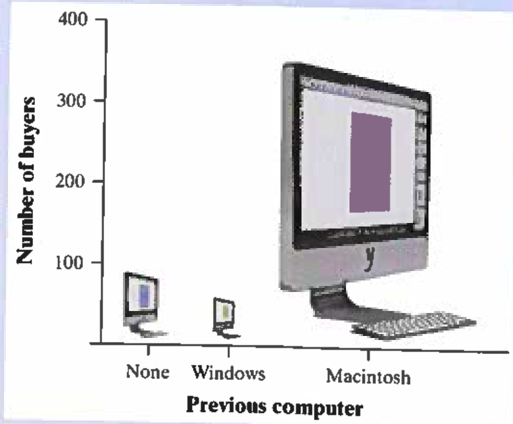

Justin Sullivan/Getty Images

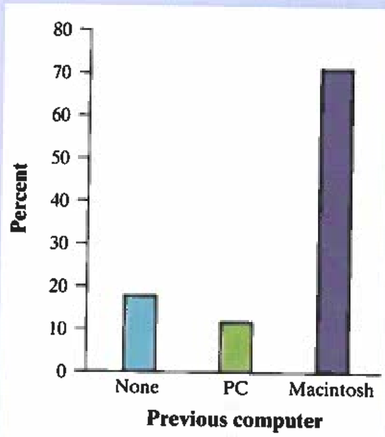
PROBLEM: When Apple, Inc., introduced the iMac, the company wanted to know whether this new computer was expanding Apple's market share. Was the iMac mainly being bought by previous Macintosh owners, or was it being purchased by first-time computer buyers and by previous PC users who were switching over? 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.[4]

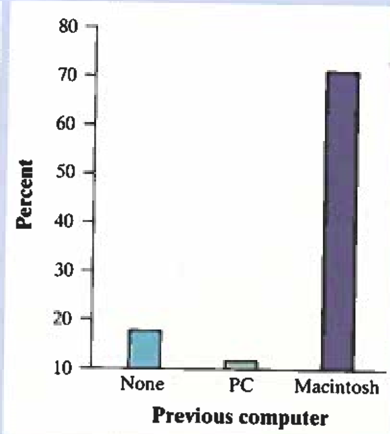
Previous ownership	Count	Percent (%)
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	500	100.0

(a) To the right is a clever graph of the data that uses pictures instead of the more traditional bars. How is this pictograph misleading?

(b) Two possible bar graphs of the data are shown below. Which one could be considered deceptive? Why?







SOLUTION:

(a) The pictograph makes it look like the percentage of iMac buyers who are former Mac owners is at least 10 times larger than either of the other two categories, which isn't true.

(b) The bar graph on the right is misleading. By starting the vertical scale at 10 instead of 0, it looks like the percentage of iMac buyers who previously owned a PC is less than half the percentage who are first-time computer buyers, which isn't true.

In part (a), the *heights* of the images are correct. But the *areas* of the images are misleading. The Macintosh image is about 6 times as tall as the PC image, but its area is about 36 times as large!

FOR PRACTICE, TRY EXERCISE 19

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

Analyzing Data on Two Categorical Variables

You have learned some techniques for analyzing the distribution of a single categorical variable. What should you do when a data set involves two categorical variables? For example, Yellowstone National Park staff surveyed a random sample of 1526 winter visitors to the park. They asked each person whether he or she belonged to an environmental club (like the Sierra Club). Respondents were also

franz12/Shutterstock

asked whether they owned, rented, or had never used a snowmobile. The data set looks something like the following:

Respondent	Environmental club?	Snowmobile use
1	No	Own
2	No	Rent
3	Yes	Never
4	Yes	Rent
5	No	Never
⋮	⋮	⋮

The **two-way table** summarizes the survey responses.

		Environmental club member?	
		No	Yes
Snowmobile use	Never	445	212
	Rent	497	77
	Own	279	16

A two-way table is sometimes called a *contingency table*.

DEFINITION Two-way table

A **two-way table** is a table of counts that summarizes data on the relationship between two categorical variables for some group of individuals.

It's easier to grasp the information in a two-way table if row and column totals are included, like the one shown here.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Now we can quickly answer questions like:

- What percent of people in the sample are environmental club members?

$$\frac{305}{1526} = 0.200 = 20.0\%$$

- What proportion of people in the sample never used a snowmobile?

$$\frac{657}{1526} = 0.431$$

These percents or proportions are known as **marginal relative frequencies** because they are calculated using values in the margins of the two-way table.

DEFINITION Marginal relative frequency

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

We could call this distribution the *marginal distribution* of environmental club membership.

We can compute marginal relative frequencies for the *column* totals to give the distribution of environmental club membership in the entire sample of 1526 park visitors:

$$\text{No: } \frac{1221}{1526} = 0.800 \text{ or } 80.0\% \qquad \text{Yes: } \frac{305}{1526} = 0.200 \text{ or } 20.0\%$$

We can compute marginal relative frequencies for the *row* totals to give the distribution of snowmobile use for all the individuals in the sample:

We could call this distribution the *marginal distribution* of snowmobile use.

$$\text{Never: } \frac{657}{1526} = 0.431 \text{ or } 43.1\%$$

$$\text{Rent: } \frac{574}{1526} = 0.376 \text{ or } 37.6\%$$

$$\text{Own: } \frac{295}{1526} = 0.193 \text{ or } 19.3\%$$

Note that we could use a bar graph or a pie chart to display either of these distributions.

A marginal relative frequency tells you about only *one* of the variables in a two-way table. It won't help you answer questions like these, which involve values of *both* variables:

- What percent of people in the sample are environmental club members and own snowmobiles?

$$\frac{16}{1526} = 0.010 = 1.0\%$$

- What proportion of people in the sample are not environmental club members and never use snowmobiles?

$$\frac{445}{1526} = 0.292$$

These percents or proportions are known as **joint relative frequencies**.

DEFINITION Joint relative frequency

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

EXAMPLE

A *Titanic* disaster

Calculating marginal and joint relative frequencies

PROBLEM: 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 gives information about adult passengers who survived and who died, by class of travel.


The New York Times.
TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSI

Blank Archives/Getty Images

(a) What proportion of adult passengers on the *Titanic* survived?

(b) Find the distribution of class of travel for adult passengers on the *Titanic* using relative frequencies.

(c) What percent of adult *Titanic* passengers traveled in third class and survived?

		Class of travel		
		First	Second	Third
Survival status	Survived	197	94	151
	Died	122	167	476

SOLUTION:

Start by finding the marginal totals.

		Class of travel			
		First	Second	Third	Total
Survival status	Survived	197	94	151	442
	Died	122	167	476	765
	Total	319	261	627	1207

(a) $\frac{442}{1207} = 0.366$

(b) First: $\frac{319}{1207} = 0.264 = 26.4\%$

Second: $\frac{261}{1207} = 0.216 = 21.6\%$

Third: $\frac{627}{1207} = 0.519 = 51.9\%$

Remember that a distribution lists the possible values of a variable and how often those values occur.

(c) $\frac{151}{1207} = 0.125 = 12.5\%$

Note that the three percentages for class of travel in part (b) do not add to exactly 100% due to roundoff error.

FOR PRACTICE, TRY EXERCISE 23

CHECK YOUR UNDERSTANDING

An article in the *Journal of the American Medical Association* reports the results of a study designed to see if the herb St. John's wort is effective in treating moderately severe cases of depression. The study involved 338 patients who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort, Zoloft (a prescription drug), or placebo (an inactive treatment) for an 8-week period. The two-way table summarizes the data from the experiment.[5]

		Treatment		
		St. John's wort	Zoloft	Placebo
Change in depression	Full response	27	27	37
	Partial response	16	26	13
	No response	70	56	66

1. What proportion of subjects in the study were randomly assigned to take St. John's wort? Explain why this value makes sense.
2. Find the distribution of change in depression for the subjects in this study using relative frequencies.
3. What percent of subjects took Zoloft and showed a full response?

Relationships Between Two Categorical Variables

Let's return to the data from the Yellowstone National Park survey of 1526 randomly selected winter visitors. Earlier, we calculated marginal and joint relative frequencies from the two-way table. These values do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

We can also use the two-way table to answer questions like:

- What percent of environmental club members in the sample are snowmobile owners?

$$\frac{16}{305} = 0.052 = 5.2\%$$

- What proportion of snowmobile renters in the sample are not environmental club members?

$$\frac{497}{574} = 0.866$$

These percents or proportions are known as **conditional relative frequencies.**

DEFINITION Conditional relative frequency

A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition).

EXAMPLE

A *Titanic* disaster
Conditional relative frequencies

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

		Class of travel			
		First	Second	Third	Total
Survival status	Survived	197	94	151	442
	Died	122	167	476	765
	Total	319	261	627	1207

(a) What proportion of survivors were third-class passengers?

(b) What percent of first-class passengers survived?

Note that a proportion is always a number between 0 and 1, whereas a percent is a number between 0 and 100. To get a percent, multiply the proportion by 100.

SOLUTION:

(a) $\frac{151}{442} = 0.342$ (b) $\frac{197}{319} = 0.618 = 61.8\%$

FOR PRACTICE, TRY EXERCISE 27

We can study the snowmobile use habits of environmental club members by looking only at the "Yes" column in the two-way table.

Snowmobile use	Environmental club: No	Yes	Total
Never used	445	212	657
Snowmobile renter	497	77	574
Snowmobile owner	279	16	295
Total	1221	305	1526

It is easy to calculate the proportions or percents of environmental club members who never use, rent, and own snowmobiles:

Never: $\frac{212}{305} = 0.695$ or 69.5% Rent: $\frac{77}{305} = 0.252$ or 25.2%

Own: $\frac{16}{305} = 0.052$ or 5.2%

This is the distribution of snowmobile use among environmental club members.

We could also refer to this distribution as the *conditional distribution* of snowmobile use among environmental club members.

We can find the distribution of snowmobile use among the survey respondents who are not environmental club members in a similar way. The table summarizes the conditional relative frequencies for both groups.

Snowmobile use	Not environmental club members	Environmental club members
Never	$\frac{445}{1221} = 0.364$ or 36.4%	$\frac{212}{305} = 0.695$ or 69.5%
Rent	$\frac{497}{1221} = 0.407$ or 40.7%	$\frac{77}{305} = 0.252$ or 25.2%
Own	$\frac{279}{1221} = 0.229$ or 22.9%	$\frac{16}{305} = 0.052$ or 5.2%

AP® EXAM TIP

When comparing groups of different sizes, be sure to use relative frequencies (percents or proportions) instead of frequencies (counts) when analyzing categorical data. Comparing only the frequencies can be misleading, as in this setting. There are many more people who never use snowmobiles among the non-environmental club members in the sample (445) than among the environmental club members (212). However, the *percentage* of environmental club members who never use snowmobiles is much higher (69.5% to 36.4%). Finally, make sure to avoid statements like "More club members never use snowmobiles" when you mean "A greater percentage of club members never use snowmobiles."

Figure 1.3 compares the distributions of snowmobile use for Yellowstone National Park visitors who are environmental club members and those who are not environmental club members with (a) a **side-by-side bar graph** and (b) a **segmented bar graph.** Notice that the segmented bar graph can be obtained by stacking the bars in the side-by-side bar graph for each of the two environmental club membership categories (no and yes).

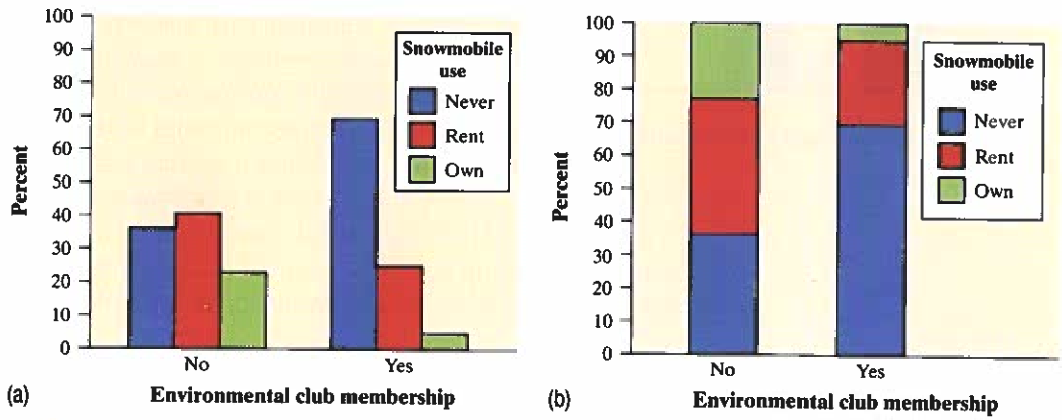


FIGURE 1.3 (a) Side-by-side bar graph and (b) segmented bar graph displaying the distribution of snowmobile use among environmental club members and among non-environmental club members from the 1526 randomly selected winter visitors to Yellowstone National Park.

DEFINITION Side-by side bar graph, Segmented bar graph

A **side-by-side bar graph** displays the distribution of a categorical variable for each value of another categorical variable. The bars are grouped together based on the values of one of the categorical variables and placed side by side.

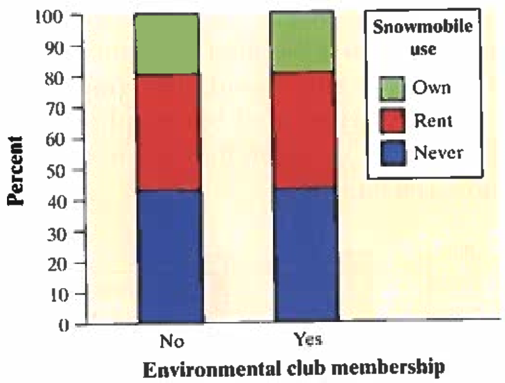
A **segmented bar graph** displays the distribution of a categorical variable as segments of a rectangle, with the area of each segment proportional to the percent of individuals in the corresponding category.

Both graphs in Figure 1.3 show a clear **association** between environmental club membership and snowmobile use in this random sample of 1526 winter visitors to Yellowstone National Park. The environmental club members were much less likely to rent (25.2% versus 40.7%) or own (5.2% versus 29.0%) snowmobiles than non-club-members and more likely to never use a snowmobile (69.5% versus 36.4%). Knowing whether or not a person in the sample is an environmental club member helps us predict that individual's snowmobile use.

DEFINITION Association

There is an **association** between two variables if knowing the value of one variable helps us predict the value of the other. If knowing the value of one variable does not help us predict the value of the other, then there is no association between the variables.

What would the graphs in Figure 1.3 look like if there was *no association* between environmental club membership and snowmobile use in the sample? The blue segments would be the same height for both the "Yes" and "No" groups.



So would the green segments and the red segments, as shown in the graph at left. In that case, knowing whether a survey respondent is an environmental club member would *not* help us predict his or her snowmobile use.

Which distributions should we compare? Our goal all along has been to analyze the relationship between environmental club membership and snowmobile use for this random sample of 1526 Yellowstone National Park visitors. We decided to calculate conditional relative frequencies of snowmobile use among environmental club members and among non-club-members. Why? Because we wanted to see if environmental club membership helped us predict snowmobile use. What if we had wanted to determine whether snowmobile use helps us predict whether a person is an environmental club member? Then we would have calculated conditional relative frequencies of environmental club membership among snowmobile owners, renters, and non-users. *In general, you should calculate the distribution of the variable that you want to predict for each value of the other variable.*

Can we say that there is an association between environmental club membership and snowmobile use in the *population* of all winter visitors to Yellowstone National Park? Making this determination requires formal inference, which will have to wait until Chapter 11.

EXAMPLE

A *Titanic* disaster

Conditional relative frequencies and association

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


Universal History Archive/Getty Images

		Class of travel			
		First	Second	Third	Total
Survival status	Survived	197	94	151	442
	Died	122	167	476	765
	Total	319	261	627	1207

(a) Find the distribution of survival status for each class of travel. Make a segmented bar graph to compare these distributions.

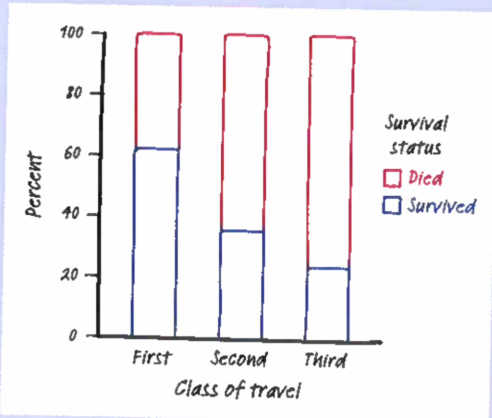
(b) Describe what the graph in part (a) reveals about the association between class of travel and survival status for adult passengers on the *Titanic*.

SOLUTION:

(a) First class Survived: $\frac{197}{319} = 0.618 = 61.8\%$ Died: $\frac{122}{319} = 0.382 = 38.2\%$

Second class Survived: $\frac{94}{261} = 0.360 = 36.0\%$ Died: $\frac{167}{261} = 0.640 = 64.0\%$

Third class Survived: $\frac{151}{627} = 0.241 = 24.1\%$ Died: $\frac{476}{627} = 0.759 = 75.9\%$



To make the segmented bar graph:

- **Draw and label the axes.** Put class of travel on the horizontal axis and percent on the vertical axis.
- **"Scale" the axes.** Use a vertical scale from 0 to 100%, with tick marks every 20%.
- **Draw bars.** Make each bar have a height of 100%. Be sure the bars are equal in width and leave spaces between them. Segment each bar based on the conditional relative frequencies you calculated. Use different colors or shading patterns to represent the two possible statuses—survived and died. Add a key to the graph that tells us which color (or shading) represents which status.

(b) Knowing a passenger's class of travel helps us predict his or her survival status. First class had the highest percentage of survivors (61.8%), followed by second class (36.0%), and then third class (24.1%).

FOR PRACTICE, TRY EXERCISE 29

Bar graphs can be used to compare any set of quantities that can be measured in the same units. See Exercises 33 and 34.

Because the variable "Survival status" has only two possible values, comparing the three distributions displayed in the segmented bar graph amounts to comparing the percent of passengers in each class of travel who survived. The bar graph in Figure 1.4 shows this comparison. Note that the bar heights do *not* add to 100%, because each bar represents a different group of passengers on the *Titanic*.

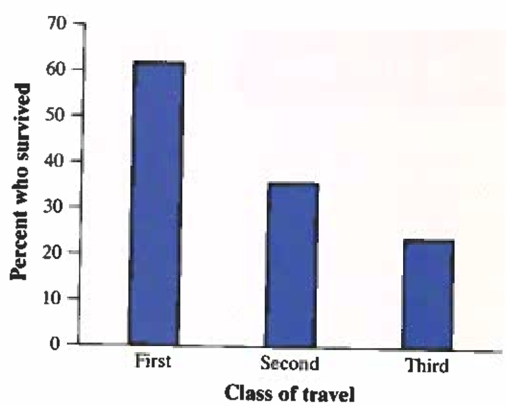


FIGURE 1.4 Bar graph comparing the percents of passengers who survived among each of the three classes of travel on the *Titanic*.

We offer a final caution about studying the relationship between two variables: **association does not imply causation.** It may be true that being in a higher class of travel on the *Titanic* increased a passenger's chance of survival. However, there isn't always a cause-and-effect relationship between two variables even if they are clearly associated. For example, a recent study proclaimed that people who are overweight are less likely to die within a few years than are people of normal weight. Does this mean that gaining weight will cause you to live longer? Not at all. The study included smokers, who tend to be thinner and also much more likely to die in a given period than non-smokers. Smokers increased the death rate for the normal-weight category, making it appear as if being overweight is better.[6] The moral of the story: *beware other variables!*

CHECK YOUR UNDERSTANDING

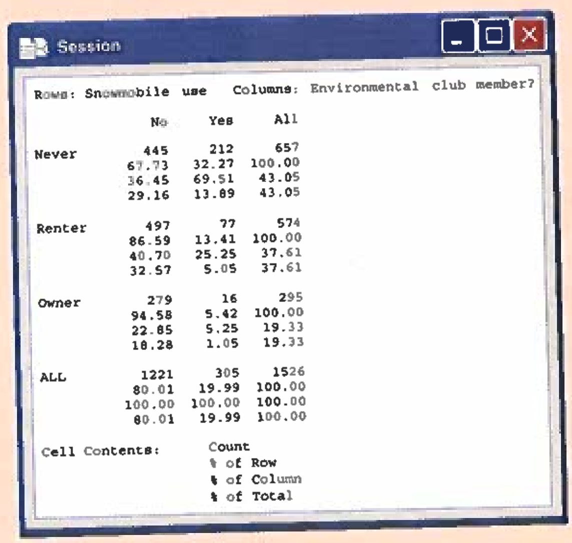
An article in the *Journal of the American Medical Association* reports the results of a study designed to see if the herb St. John's wort is effective in treating moderately severe cases of depression. The study involved 338 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort, Zoloft (a prescription drug), or placebo (an inactive treatment) for an 8-week period. The two-way table summarizes the data from the experiment.

		Treatment		
		St. John's wort	Zoloft	Placebo
Change in depression	Full response	27	27	37
	Partial response	16	26	13
	No response	70	56	66

1. What proportion of subjects who showed a full response took St. John's wort?
2. What percent of subjects who took St. John's wort showed no response?
3. Find the distribution of change in depression for the subjects receiving each of the three treatments. Make a segmented bar graph to compare these distributions.
4. Describe what the graph in Question 3 reveals about the association between treatment and change in depression for these subjects.

1. Technology Corner ANALYZING TWO-WAY TABLES

Statistical software will provide marginal relative frequencies, joint relative frequencies, and conditional relative frequencies for data summarized in a two-way table. Here is output from Minitab for the data on snowmobile use and environmental club membership. Use the information on cell contents at the bottom of the output to help you interpret what each value in the table represents.


```
Session

Rows: Snowmobile use   Columns: Environmental club member?

             No      Yes      All

Never       445      212      657
          67.73    32.27   100.00
          36.45    69.51    43.05
          29.16    13.89    43.05

Renter      497       77      574
          86.59    13.41   100.00
          40.70    25.25    37.61
          32.57     5.05    37.61

Owner       279       16      295
          94.58     5.42   100.00
          22.85     5.25    19.33
          18.28     1.05    19.33

ALL        1221      305     1526
          80.01    19.99   100.00
         100.00   100.00   100.00
          80.01    19.99   100.00

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
```

Section 1.1 Summary

- The distribution of a categorical variable lists the categories and gives the **frequency** (count) or **relative frequency** (percent or proportion) of individuals that fall in each category.
- You can use a **pie chart** or **bar graph** to display the distribution of a categorical variable. When examining any graph, ask yourself, "What do I see?"
- Beware of graphs that mislead the eye. Look at the scales to see if they have been distorted to create a particular impression. Avoid making graphs that replace the bars of a bar graph with pictures whose height and width both change.
- A **two-way table** of counts summarizes data on the relationship between two categorical variables for some group of individuals.
- You can use a two-way table to calculate three types of relative frequencies:
  - A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable. Use the appropriate row total or column total in a two-way table when calculating a marginal relative frequency.
  - A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable. Use the value from the appropriate cell in the two-way table when calculating a joint relative frequency.
  - A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition). Use conditional relative frequencies to compare distributions of a categorical variable for two or more groups.
- Use a **side-by-side bar graph** or a **segmented bar graph** to compare the distribution of a categorical variable for two or more groups.
- There is an **association** between two variables if knowing the value of one variable helps predict the value of the other. To see whether there is an association between two categorical variables, find the distribution of one variable for each value of the other variable by calculating an appropriate set of conditional relative frequencies.

1.1 Technology Corner

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

1. Analyzing two-way tables Page 22

Section 1.1 Exercises

11. **Birth days** The frequency table summarizes data on the numbers of babies born on each day of the week in the United States in a recent week.[7]

Day	Births
Sunday	7374
Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8459

(a) Identify the individuals in this data set.

(b) Make a frequency bar graph to display the data. Describe what you see.

12. **Going up?** As of 2015, there were over 75,000 elevators in New York City. The frequency table summarizes data on the number of elevators of each type.[8]

Type	Count
Passenger elevator	66,602
Freight elevator	4140
Escalator	2663
Dumbwaiter	1143
Sidewalk elevator	943
Private elevator	252
Handicap lift	227
Manlift	73
Public elevator	45

(a) Identify the individuals in this data set.

(b) Make a frequency bar graph to display the data. Describe what you see.

13. **Buying cameras** The brands of the last 45 digital single-lens reflex (SLR) cameras sold on a popular Internet auction site are listed here. Make a relative frequency bar graph for these data. Describe what you see.

Canon	Sony	Canon	Nikon	Fujifilm
Nikon	Canon	Sony	Canon	Canon
Nikon	Canon	Nikon	Canon	Canon
Canon	Nikon	Fujifilm	Canon	Nikon
Nikon	Canon	Canon	Canon	Canon
Olympus	Canon	Canon	Canon	Nikon
Olympus	Sony	Canon	Canon	Sony
Canon	Nikon	Sony	Canon	Fujifilm
Nikon	Canon	Nikon	Canon	Sony

14. **Disc dogs** Here is a list of the breeds of dogs that won the World Canine Disc Championships from 1975 through 2016. Make a relative frequency bar graph for these data. Describe what you see.

Whippet	Mixed breed	Australian shepherd
Whippet	Australian shepherd	Australian shepherd
Whippet	Border collie	Australian shepherd
Mixed breed	Australian shepherd	Border collie
Mixed breed	Mixed breed	Border collie
Other purebred	Mixed breed	Australian shepherd
Labrador retriever	Mixed breed	Border collie
Mixed breed	Border collie	Border collie
Mixed breed	Border collie	Other purebred
Border collie	Australian shepherd	Border collie
Mixed breed	Border collie	Border collie
Mixed breed	Australian shepherd	Border collie
Labrador retriever	Border collie	Mixed breed
Labrador retriever	Mixed breed	Australian shepherd

15. **Cool car colors** The most popular colors for cars and light trucks change over time. Silver advanced past green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. Here is a relative frequency table that summarizes data on the colors of vehicles sold worldwide in a recent year.[9]

Color	Percent of vehicles	Color	Percent of vehicles
Black	19	Red	9
Blue	6	Silver	14
Brown/beige	5	White	29
Gray	12	Yellow/gold	3
Green	1	Other	??

(a) What percent of vehicles would fall in the "Other" category?

(b) Make a bar graph to display the data. Describe what you see.

(c) Would it be appropriate to make a pie chart of these data? Explain.

16. Spam Email spam is the curse of the Internet. Here is a relative frequency table that summarizes data on the most common types of spam:[10]

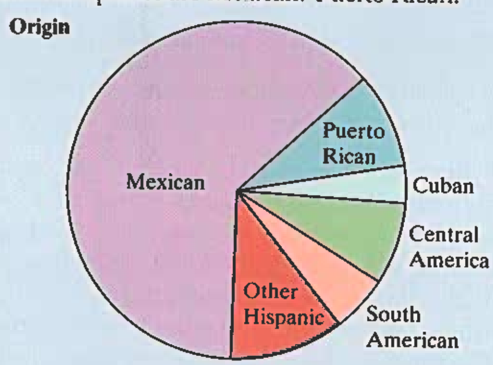
Type of spam	Percent
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

(a) What percent of spam would fall in the "Other" category?

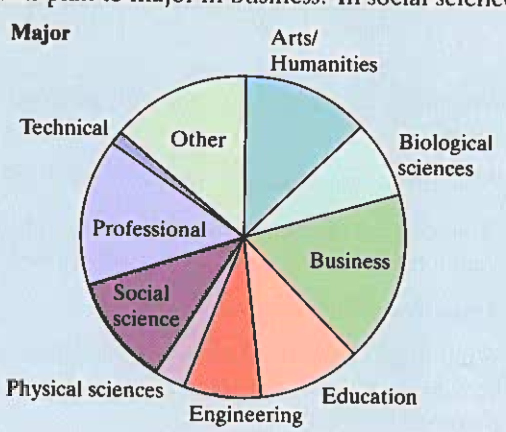
(b) Make a bar graph to display the data. Describe what you see.

(c) Would it be appropriate to make a pie chart of these data? Explain.

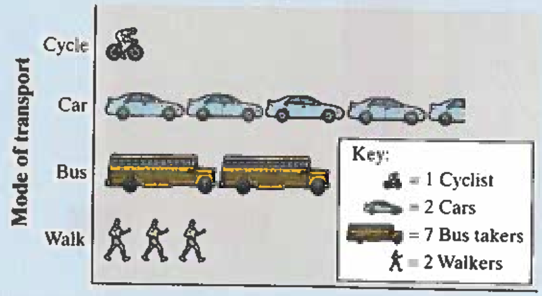
17. Hispanic origins Here is a pie chart prepared by the Census Bureau to show the origin of the more than 50 million Hispanics in the United States in 2010.[11] About what percent of Hispanics are Mexican? Puerto Rican?



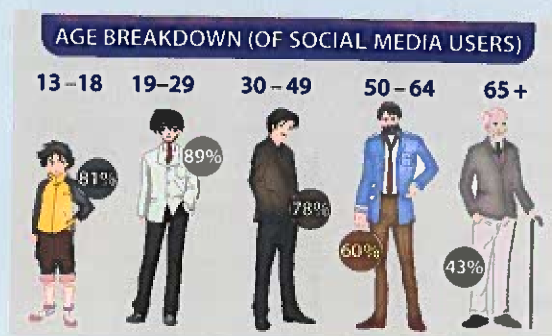
18. Which major? About 3 million first-year students enroll in colleges and universities each year. What do they plan to study? The pie chart displays data on the percent of first-year students who plan to major in several disciplines.[12] About what percent of first-year students plan to major in business? In social science?



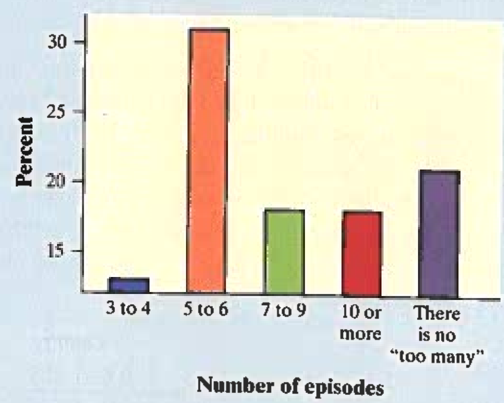
19. Going to school Students in a high school statistics class were given data about the main method of transportation to school for a group of 30 students. They produced the pictograph shown. Explain how this graph is misleading.



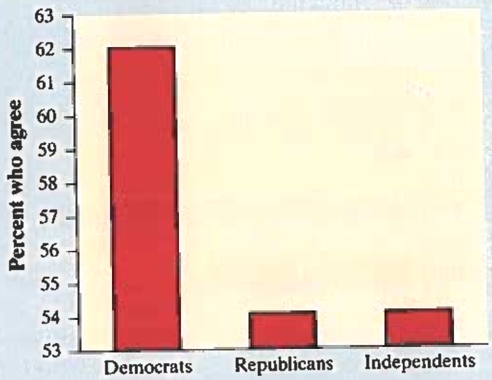
20. Social media The Pew Research Center surveyed a random sample of U.S. teens and adults about their use of social media. The following pictograph displays some results. Explain how this graph is misleading.



21. Binge-watching Do you "binge-watch" television series by viewing multiple episodes of a series at one sitting? A survey of 800 people who binge-watch were asked how many episodes is too many to watch in one viewing session. The results are displayed in the bar graph.[13] Explain how this graph is misleading.



22. Support the court? A news network reported the results of a survey about a controversial court decision. The network initially posted on its website a bar graph of the data similar to the one that follows. Explain how this graph is misleading. (*Note:* When notified about the misleading nature of its graph, the network posted a corrected version.)



23. A smash or a hit? Researchers asked 150 subjects to recall the details of a car accident they watched on video. Fifty subjects were randomly assigned to be asked, "About how fast were the cars going when they smashed into each other?" For another 50 randomly assigned subjects, the words "smashed into" were replaced with "hit." The remaining 50 subjects—the control group—were not asked to estimate speed. A week later, all subjects were asked if they saw any broken glass at the accident (there wasn't any). The table shows each group's response to the broken glass question.[14]

pg 15

		Treatment		
		"Smashed into"	"Hit"	Control
Response	Yes	16	7	6
	No	34	43	44

(a) What proportion of subjects were given the control treatment?

(b) Find the distribution of responses about whether there was broken glass at the accident for the subjects in this study using relative frequencies.

(c) What percent of the subjects were given the "smashed into" treatment and said they saw broken glass at the accident?

24. Superpowers A total of 415 children from the United Kingdom and the United States who completed a survey in a recent year were randomly selected. Each student's country of origin was recorded along with which superpower they would most like to have: the ability to fly, ability to freeze time, invisibility, superstrength, or telepathy (ability to read minds). The data are summarized in the following table.[15]

		Country	
		U.K.	U.S.
Superpower	Fly	54	45
	Freeze time	52	44
	Invisibility	30	37
	Superstrength	20	23
	Telepathy	44	66

(a) What proportion of students in the sample are from the United States?

(b) Find the distribution of superpower preference for the students in the sample using relative frequencies.

(c) What percent of students in the sample are from the United Kingdom and prefer telepathy as their superpower preference?

25. Body image A random sample of 1200 U.S. college students was asked, "What is your perception of your own body? Do you feel that you are overweight, underweight, or about right?" The two-way table summarizes the data on perceived body image by gender.[16]

		Gender		
		Female	Male	Total
Body image	About right	560	295	855
	Overweight	163	72	235
	Underweight	37	73	110
	Total	760	440	1200

(a) What percent of respondents feel that their body weight is about right?

(b) What proportion of the sample is female?

(c) What percent of respondents are males and feel that they are overweight or underweight?

26. Python eggs How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers randomly assigned newly laid eggs to one of three water temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. The two-way table summarizes the data on whether or not the eggs hatched.[17]

		Water temperature			
		Cold	Neutral	Hot	Total
Hatched?	Yes	16	38	75	129
	No	11	18	29	58
	Total	27	56	104	187

(a) What percent of eggs were randomly assigned to hot water?

(b) What proportion of eggs in the study hatched?

(c) What percent of eggs in the study were randomly assigned to cold or neutral water and hatched?

27. A smash or a hit Refer to Exercise 23.

pg 17

(a) What proportion of subjects who said they saw broken glass at the accident received the "hit" treatment?

(b) What percent of subjects who received the "smashed into" treatment said they did not see broken glass at the accident?

28. **Superpower** Refer to Exercise 24.

(a) What proportion of students in the sample who prefer invisibility as their superpower are from the United States?

(b) What percent of students in the sample who are from the United Kingdom prefer superstrength as their superpower?

29. **A smash or a hit** Refer to Exercise 23.

pg 20

(a) Find the distribution of responses about whether there was broken glass at the accident for each of the three treatment groups. Make a segmented bar graph to compare these distributions.

(b) Describe what the graph in part (a) reveals about the association between response about broken glass at the accident and treatment received for the subjects in the study.

30. **Superpower** Refer to Exercise 24.

(a) Find the distribution of superpower preference for the students in the sample from each country (i.e., the United States and the United Kingdom). Make a segmented bar graph to compare these distributions.

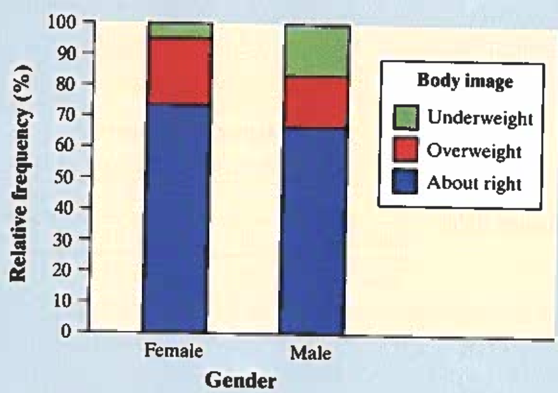
(b) Describe what the graph in part (a) reveals about the association between country of origin and superpower preference for the students in the sample.

31. **Body image** Refer to Exercise 25.

(a) Of the respondents who felt that their body weight was about right, what proportion were female?

(b) Of the female respondents, what percent felt that their body weight was about right?

(c) The segmented bar graph displays the distribution of perceived body image by gender. Describe what this graph reveals about the association between these two variables for the 1200 college students in the sample.

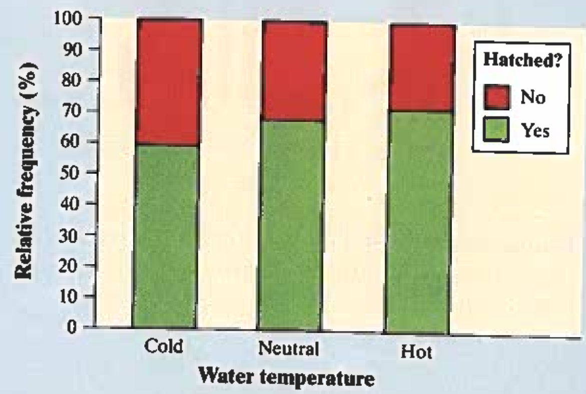


32. **Python eggs** Refer to Exercise 26.

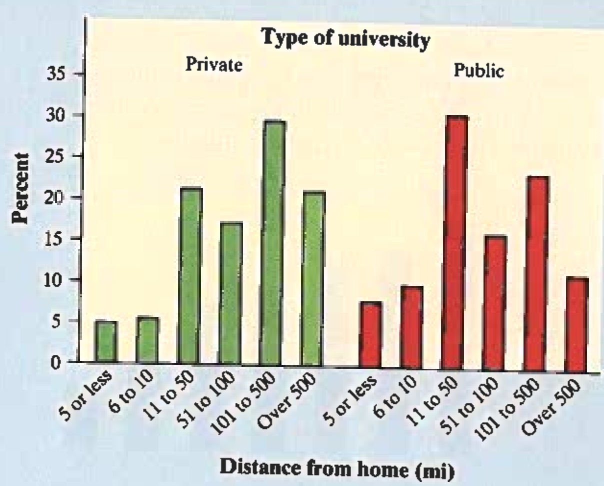
(a) Of the eggs that hatched, what proportion were randomly assigned to hot water?

(b) Of the eggs that were randomly assigned to hot water, what percent hatched?

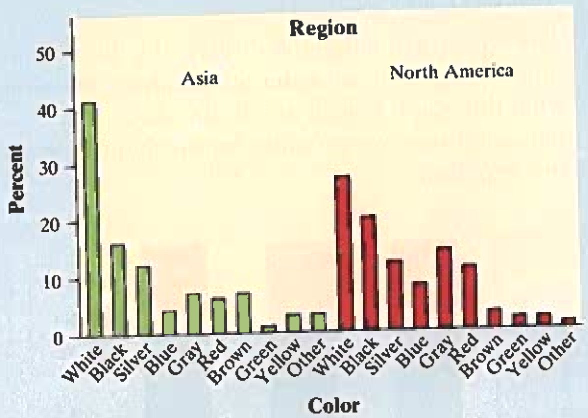
(c) The segmented bar graph displays the distribution of hatching status by water temperature. Describe what this graph reveals about the association between these two variables for the python eggs in this experiment.



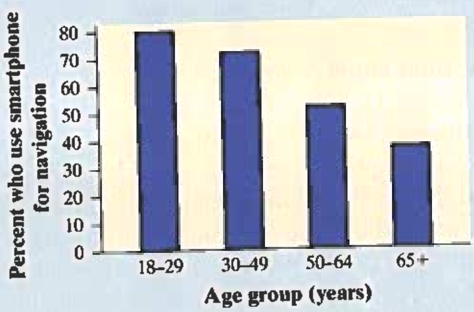
33. **Far from home** A survey asked first-year college students, "How many miles is this college from your permanent home?" Students had to choose from the following options: 5 or fewer, 6 to 10, 11 to 50, 51 to 100, 101 to 500, or more than 500. The side-by-side bar graph shows the percentage of students at public and private 4-year colleges who chose each option.[18] Write a few sentences comparing the distributions of distance from home for students from private and public 4-year colleges who completed the survey.



34. **Popular car colors** Favorite car colors may differ among countries. The side-by-side bar graph displays data on the most popular car colors in a recent year for North America and Asia. Write a few sentences comparing the distributions.[19]



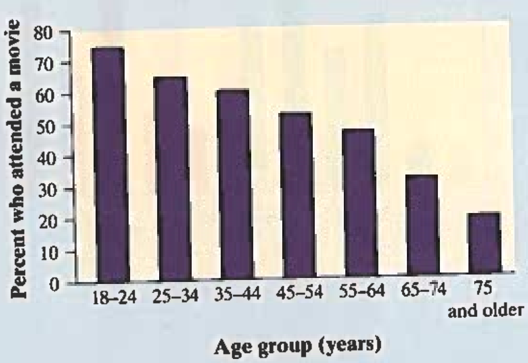
35. **Phone navigation** The bar graph displays data on the percent of smartphone owners in several age groups who say that they use their phone for turn-by-turn navigation.[20]



(a) Describe what the graph reveals about the relationship between age group and use of smartphones for navigation.

(b) Would it be appropriate to make a pie chart of the data? Explain.

36. **Who goes to movies?** The bar graph displays data on the percent of people in several age groups who attended a movie in the past 12 months.[21]



(a) Describe what the graph reveals about the relationship between age group and movie attendance.

(b) Would it be appropriate to make a pie chart of the data? Explain.

37. **Marginal totals aren't the whole story** Here are the row and column totals for a two-way table with two rows and two columns:

a	*b*	50
c	*d*	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.

38. **Women and children first?** Here's another table that summarizes data on survival status by gender and class of travel on the *Titanic*:

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

(a) Find the distributions of survival status for males and for females within each class of travel. Did women survive the disaster at higher rates than men? Explain.

(b) In an earlier example, we noted that survival status is associated with class of travel. First-class passengers had the highest survival rate, while third-class passengers had the lowest survival rate. Does this same relationship hold for both males and females in all three classes of travel? Explain.

39. **Simpson's paradox** Accident victims are sometimes taken by helicopter from the accident scene to a hospital. Helicopters save time. Do they also save lives? The two-way table summarizes data from a sample of patients who were transported to the hospital by helicopter or by ambulance.[22]

		Method of transport		
		Helicopter	Ambulance	Total
Survival status	Died	64	260	324
	Survived	136	840	976
	Total	200	1100	1300

(a) What percent of patients died with each method of transport?

Here are the same data broken down by severity of accident:

Serious accidents

		Method of transport		
		Helicopter	Ambulance	Total
Survival status	Died	48	60	108
	Survived	52	40	92
	Total	100	100	200

Less serious accidents

		Method of transport		
		Helicopter	Ambulance	Total
Survival status	Died	16	200	216
	Survived	84	800	884
	Total	100	1000	1100

(b) Calculate the percent of patients who died with each method of transport for the serious accidents. Then calculate the percent of patients who died with each method of transport for the less serious accidents. What do you notice?

(c) See if you can explain how the result in part (a) is possible given the result in part (b).

Note: This is an example of *Simpson's paradox*, which states that an association between two variables that holds for each value of a third variable can be changed or even reversed when the data for all values of the third variable are combined.

Multiple Choice: *Select the best answer for Exercises 40–43.*

40. For which of the following would it be *inappropriate* to display the data with a single pie chart?

(a) The distribution of car colors for vehicles purchased in the last month

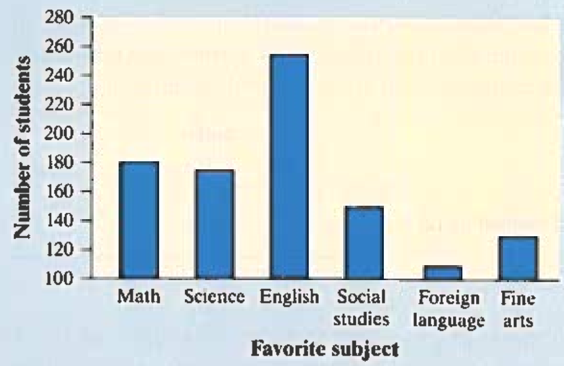
(b) The distribution of unemployment percentages for each of the 50 states

(c) The distribution of favorite sport for a sample of 30 middle school students

(d) The distribution of shoe type worn by shoppers at a local mall

(e) The distribution of presidential candidate preference for voters in a state

41. The following bar graph shows the distribution of favorite subject for a sample of 1000 students. What is the most serious problem with the graph?



(a) The subjects are not listed in the correct order.

(b) This distribution should be displayed with a pie chart.

(c) The vertical axis should show the percent of students.

(d) The vertical axis should start at 0 rather than 100.

(e) The foreign language bar should be broken up by language.

42. The Dallas Mavericks won the NBA championship in the 2010–2011 season. The two-way table displays the relationship between the outcome of each game in the regular season and whether the Mavericks scored at least 100 points.

		Points scored		
		100 or more	Fewer than 100	Total
Outcome of game	Win	43	14	57
	Loss	4	21	25
	Total	47	35	82

Which of the following is the best evidence that there is an association between the outcome of a game and whether or not the Mavericks scored at least 100 points?

(a) The Mavericks won 57 games and lost only 25 games.

(b) The Mavericks scored at least 100 points in 47 games and fewer than 100 points in only 35 games.

(c) The Mavericks won 43 games when scoring at least 100 points and only 14 games when scoring fewer than 100 points.

(d) The Mavericks won a higher proportion of games when scoring at least 100 points (43/47) than when they scored fewer than 100 points (14/35).

(e) The combination of scoring 100 or more points and winning the game occurred more often (43 times) than any other combination of outcomes.

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

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

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

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

Recycle and Review

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

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

SECTION 1.2 Displaying Quantitative Data with Graphs

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

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

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

Dotplots

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

DEFINITION Dotplot

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

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

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

Figure 1.5 shows a dotplot of these data.

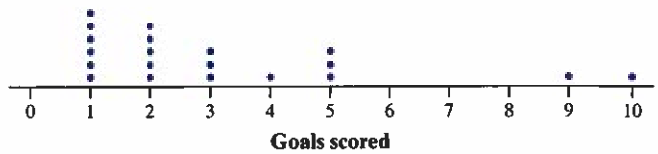


FIGURE 1.5 Dotplot of goals scored in 20 games by the 2016 U.S. women's soccer team.

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

HOW TO MAKE A DOTPLOT

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

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

EXAMPLE

Give it some gas!
Making and interpreting dotplots

Ann Heath

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

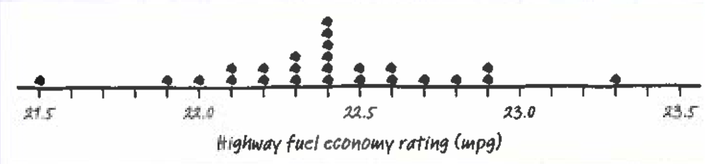
22.4	22.4	22.3	23.3	22.3	22.3	22.5	22.4	22.1	21.5	22.0	22.2	22.7
22.8	22.4	22.6	22.9	22.5	22.1	22.4	22.2	22.9	22.6	21.9	22.4	

(a) Make a dotplot of these data.

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

SOLUTION:

(a)



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

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

FOR PRACTICE, TRY EXERCISE 45

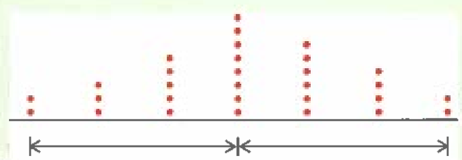
Describing Shape

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

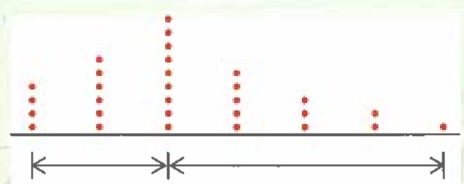
We could also describe a distribution with a long tail to the left as "skewed toward negative values" or "negatively skewed" and a distribution with a long right tail as "positively skewed."

DEFINITION Symmetric and skewed distributions

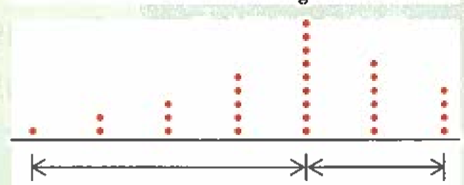
A distribution is roughly **symmetric** if the right side of the graph (containing the half of observations with the largest values) is approximately a mirror image of the left side.


Roughly symmetric

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


Skewed to the right

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


Skewed to the left

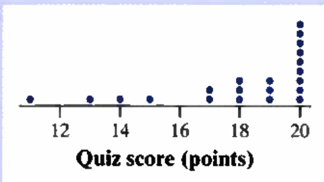
For ease, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed to the right." **The direction of skewness is toward the long tail, not the direction where most observations are clustered.** The drawing is a cute but corny way to help you keep this straight.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

EXAMPLE

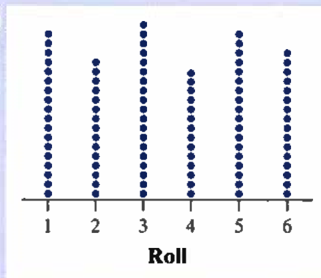
Quiz scores and die rolls
Describing shape

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

(a)



(b)



SOLUTION:

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

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

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

FOR PRACTICE, TRY EXERCISE 49

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

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

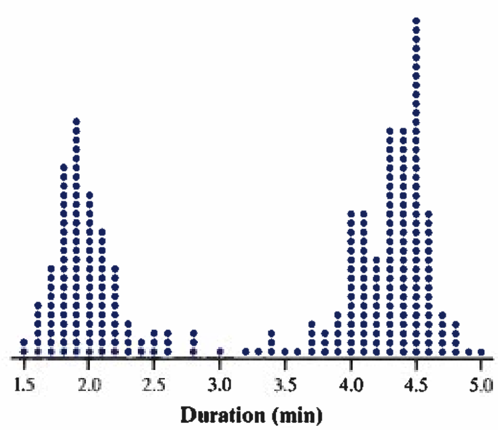


FIGURE 1.6 Dotplot displaying duration (in minutes) of 220 Old Faithful eruptions. This graph has two distinct clusters and two clear peaks.

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

CHECK YOUR UNDERSTANDING

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

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

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

Describing Distributions

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

HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

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

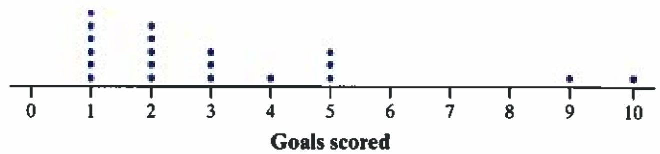
- You can describe the overall pattern of a distribution by its **shape, center,** and **variability.**
- An important kind of departure is an **outlier,** an observation that falls outside the overall pattern.

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

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

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

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



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

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

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

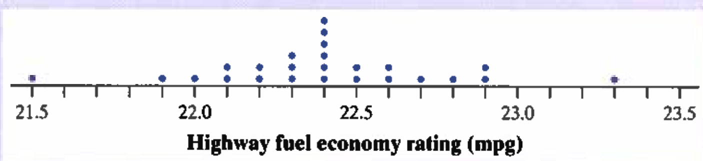
Center: The median is 2 goals scored.

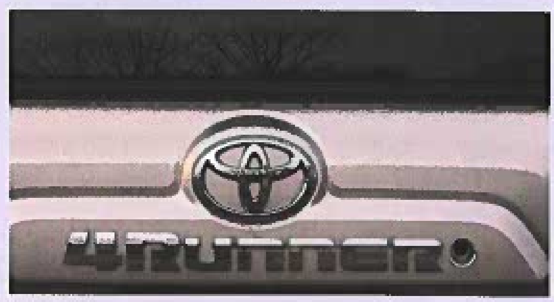
Variability: The data vary from 1 to 10 goals scored.

EXAMPLE

Give it some gas!
Describing a distribution

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





Daren Starnes

SOLUTION:

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

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

Center: **The median rating is 22.4 mpg.**

Variability: **The ratings vary from 21.5 to 23.3 mpg.**

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

FOR PRACTICE, TRY EXERCISE 53

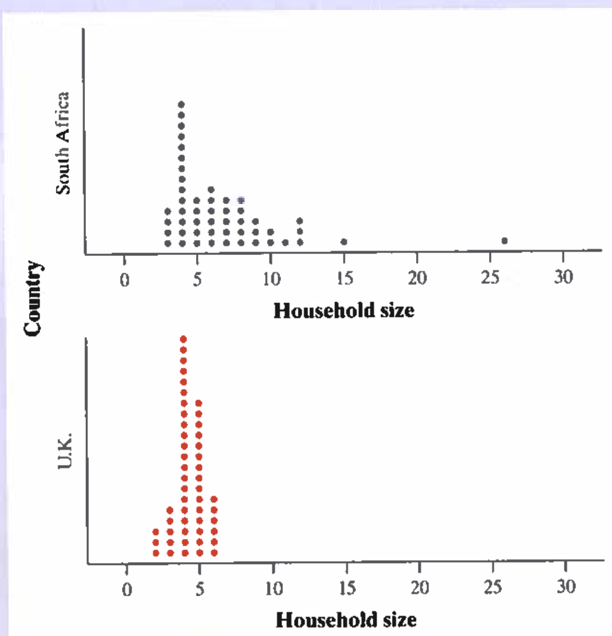
Comparing Distributions

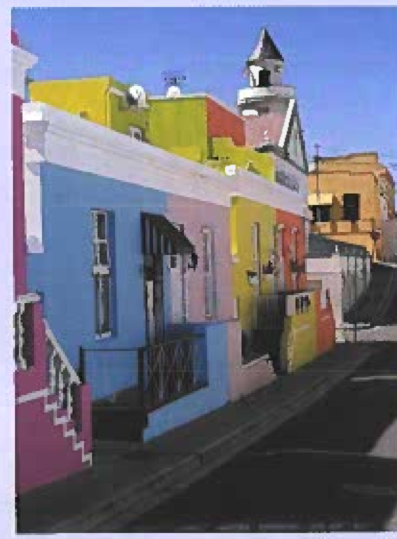
Some of the most interesting statistics questions involve comparing two or more groups. Which of two popular diets leads to greater long-term weight loss? Who texts more—males or females? As the following example suggests, you should always discuss shape, outliers, center, and variability whenever you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

EXAMPLE

Household size: U.K. versus South Africa
Comparing distributions

PROBLEM: 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 Census At School's "Random Data Selector" to choose 50 students from each country. Here are dotplots of the household sizes reported by the survey respondents. Compare the distributions of household size for these two countries.





FrankvandenBergh/Getty Images

AP® EXAM TIP

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

SOLUTION:

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.

Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers: the households with 15 and 26 people.

Center: Household sizes for the South African students tend to be larger (median = 6 people) than for the U.K. students (median = 4 people).

Variability: The household sizes for the South African students vary more (from 3 to 26 people) than for the U.K. students (from 2 to 6 people).

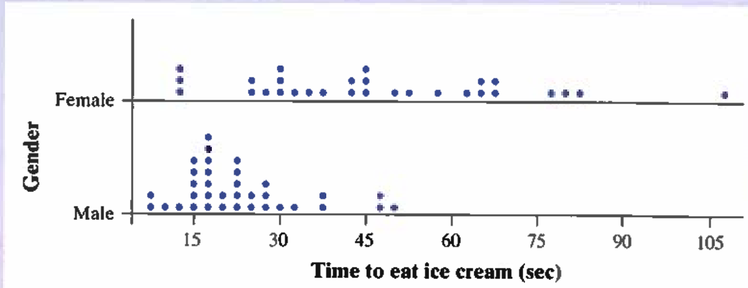
Don't forget to include context! It isn't enough to refer to the U.K. distribution or the South Africa distribution. You need to mention the variable of interest, household size.

FOR PRACTICE, TRY EXERCISE 55

Notice that in the preceding example, we discussed the distributions of household size only for the two *samples* of 50 students. 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.

CHECK YOUR UNDERSTANDING

For a statistics class project, Jonathan and Crystal hosted an ice-cream-eating contest. Each student in the contest was given a small cup of ice cream and instructed to eat it as fast as possible. Jonathan and Crystal then recorded each contestant's gender and time (in seconds), as shown in the dotplots. Compare the distributions of eating times for males and females.



Stemplots

Another simple type of graph for displaying quantitative data is a **stemplot**.

A stemplot is also known as a *stem-and-leaf plot*.

DEFINITION Stemplot

A **stemplot** shows each data value separated into two parts: a *stem*, which consists of all but the final digit, and a *leaf*, the final digit. The stems are ordered from lowest to highest and arranged in a vertical column. The leaves are arranged in increasing order out from the appropriate stems.

Here are data on the resting pulse rates (beats per minute) of 19 middle school students:

71 104 76 88 78 71 68 86 70 90 74 76 69 68 88 96 68 82 120

Figure 1.7 shows a stemplot of these data.

```
 6 | 8889
 7 | 0114668
 8 | 2688
 9 | 06
10 | 4
11 |
12 | 0
```

Key: 8|2 is a student whose resting pulse rate is 82 beats per minute.

FIGURE 1.7 Stemplot of the resting pulse rates of 19 middle school students.

According to the American Heart Association, a resting pulse rate above 100 beats per minute is considered high for this age group. We can see that 2/19 = 0.105 = 10.5%

of these students have high resting pulse rates by this standard. Also, the distribution of pulse rates for these 19 students is skewed to the right (toward the larger values).

Stemplots give us a quick picture of a distribution that includes the individual observations in the graph. It is fairly easy to make a stemplot by hand for small sets of quantitative data.

HOW TO MAKE A STEMPLOT

- **Make stems.** 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. 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.
- **Add leaves.** Write each leaf in the row to the right of its stem.
- **Order leaves.** Arrange the leaves in increasing order out from the stem.
- **Add a key.** Provide a key that identifies the variable and explains what the stems and leaves represent.

EXAMPLE

Wear your helmets!

Making and interpreting stemplots

Pete Saloutos/AGE Fotostock

PROBLEM: Many athletes (and their parents) worry about the risk of concussions when playing sports. A football coach plans to obtain specially made helmets for his players that are designed to reduce the chance of getting a concussion. Here are the measurements of head circumference (in inches) for the 30 players on the team:

23.0	22.2	21.7	22.0	22.3	22.6	22.7	21.5	22.7	25.6	20.8	23.0	24.2	23.5	20.8
24.0	22.7	22.6	23.9	22.5	23.1	21.9	21.0	22.4	23.5	22.5	23.9	23.4	21.6	23.3

(a) Make a stemplot of these data.

(b) Describe the shape of the distribution. Are there any obvious outliers?

SOLUTION:

(a)

20	88
21	05679
22	02345566777
23	001345599
24	02
25	6

Key: 23|5 is a player with a head circumference of 23.5 inches.

To make the stemplot:
- **Make stems.** The smallest head circumference is 20.8 inches and the largest is 25.6 inches. We use the first two digits as the stem and the final digit as the leaf. So we need stems from 20 to 25.
- **Add leaves.**
- **Order leaves.**
- **Add a key.**

(b) The distribution of head circumferences for the 30 players on the team is roughly symmetric, with a single peak on the 22-inch stem. There are no obvious outliers.

FOR PRACTICE, TRY EXERCISE 59

We can get a better picture of the head circumference data by *splitting stems.* In Figure 1.8(a), leaf values from 0 to 9 are placed on the same stem. Figure 1.8(b) shows another stemplot of the same data. This time, values with leaves from 0 to 4 are placed on one stem, while those with leaves from 5 to 9 are placed on another stem. Now we can see the shape of the distribution even more clearly—including the possible outlier at 25.6 inches.

20	88
21	05679
22	02345566777
23	001345599
24	02
25	6

(a)

Key: 23|5 is a player with a head circumference of 23.5 inches.

20	88
21	0
21	5679
22	0234
22	5566777
23	00134
23	5599
24	02
24	
25	
25	6

(b)

FIGURE 1.8 Two stemplots showing the head circumference data. The graph in (b) improves on the graph in (a) by splitting stems.

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

- There is no magic number of stems to use. Too few or too many stems will make it difficult to see the distribution's shape. Five stems is a good minimum.
- If you split stems, be sure that each stem is assigned an equal number of possible leaf digits.
- When the data have too many digits, you can get more flexibility by rounding or truncating the data. See Exercises 61 and 62 for an illustration of rounding data before making a stemplot.

You can use a *back-to-back stemplot* with common stems to compare the distribution of a quantitative variable in two groups. The leaves are placed in order on each side of the common stem. For example, Figure 1.9 shows a back-to-back stemplot of the 19 middle school students' resting pulse rates and their pulse rates after 5 minutes of running.

Resting		After exercise
9888	6	
8664110	7	
8862	8	6788
60	9	02245899
4	10	044
	11	8
0	12	44
	13	
	14	6

Key: 8|2 is a student whose pulse rate is 82 beats per minute.

FIGURE 1.9 Back-to-back stemplot of 19 middle school students' resting pulse rates and their pulse rates after 5 minutes of running.

CHECK YOUR UNDERSTANDING

1. Write a few sentences comparing the distributions of resting and after-exercise pulse rates in Figure 1.9.

Multiple Choice: *Select the best answer for Questions 2–4.*

Here is a stemplot of the percent of residents aged 65 and older in the 50 states and the District of Columbia:

6	8
7	
8	8
9	79
10	08
11	15566
12	0122234444578889999
13	01233333444899
14	02666
15	23
16	8

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

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.
   (b) skewed to the left.
   (c) skewed to the middle.
   (d) double-peaked.
   (e) roughly symmetric.
4. The center of the distribution is close to
   (a) 11.6%. (b) 12.0%. (c) 12.8%. (d) 13.3%. (e) 6.8% to 16.8%.

Histograms

You can use a dotplot or stemplot to display quantitative data. Both graphs show every individual data value. For large data sets, this can make it difficult to see the overall pattern in the graph. We often get a clearer picture of the distribution by grouping together nearby values. Doing so allows us to make a new type of graph: a **histogram**.

DEFINITION Histogram

A **histogram** shows each interval of values as a bar. The heights of the bars show the frequencies or relative frequencies of values in each interval.

Figure 1.10 shows a dotplot and a histogram of the durations (in minutes) of 220 eruptions of the Old Faithful geyser. Notice how the histogram groups together nearby values.

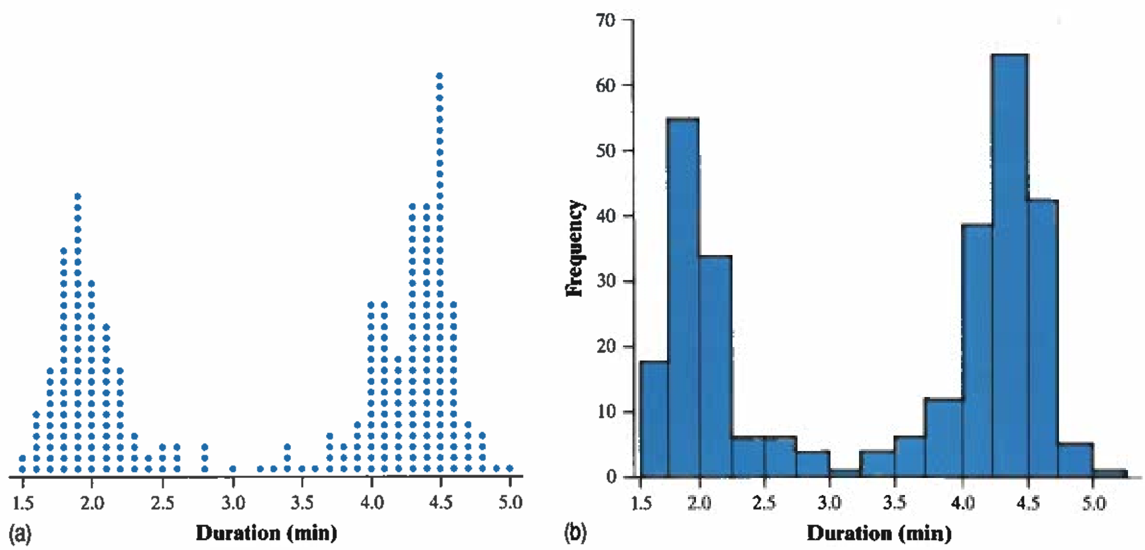


FIGURE 1.10 (a) Dotplot and (b) histogram of the duration (in minutes) of 220 eruptions of the Old Faithful geyser.

It is fairly easy to make a histogram by hand. Here's how you do it.

HOW TO MAKE A HISTOGRAM

- **Choose equal-width intervals** that span the data. Five intervals is a good minimum.
- **Make a table** that shows the frequency (count) or relative frequency (percent or proportion) of individuals in each interval. Put values that fall on an interval boundary in the interval containing larger values.
- **Draw and label the axes.** Draw horizontal and vertical axes. Put the name of the quantitative variable under the horizontal axis. To the left of the vertical axis, indicate whether the graph shows the frequency (count) or relative frequency (percent or proportion) of individuals in each interval.
- **Scale the axes.** Place equally spaced tick marks at the smallest value in each interval along the horizontal axis. On the vertical axis, start at 0 and place equally spaced tick marks until you exceed the largest frequency or relative frequency in any interval.
- **Draw bars** above the intervals. Make the bars equal in width and leave no gaps between them. Be sure that the height of each bar corresponds to the frequency or relative frequency of individuals in that interval. An interval with no data values will appear as a bar of height 0 on the graph.

It is possible to choose intervals of unequal widths when making a histogram. Such graphs are beyond the scope of this book.

EXAMPLE

How much tax?

Making and interpreting histograms

PROBLEM: Sales tax rates vary widely across the United States. Four states charge no state or local sales tax: Delaware, Montana, New Hampshire, and Oregon. The table shows data on the average total tax rate for each of the remaining 46 states and the District of Columbia.[23]

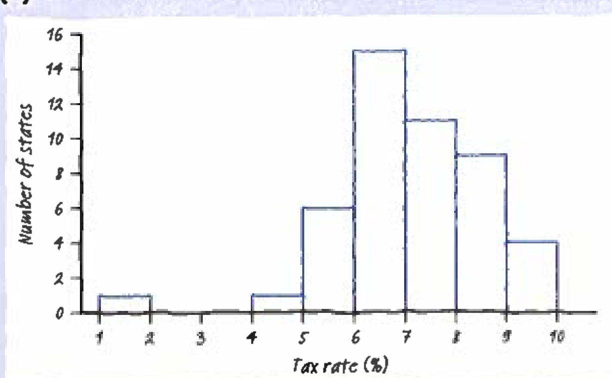
State	Tax rate (%)	State	Tax rate (%)	State	Tax rate (%)
Alabama	9.0	Louisiana	9.0	Oklahoma	8.8
Alaska	1.8	Maine	5.5	Pennsylvania	6.3
Arizona	8.3	Maryland	6.0	Rhode Island	7.0
Arkansas	9.3	Massachusetts	6.3	South Carolina	7.2
California	8.5	Michigan	6.0	South Dakota	5.8
Colorado	7.5	Minnesota	7.3	Tennessee	9.5
Connecticut	6.4	Mississippi	7.1	Texas	8.2
Florida	6.7	Missouri	7.9	Utah	6.7
Georgia	7.0	Nebraska	6.9	Vermont	6.2
Hawaii	4.4	Nevada	8.0	Virginia	5.6
Idaho	6.0	New Jersey	7.0	Washington	8.9
Illinois	8.6	New Mexico	7.5	West Virginia	6.2
Indiana	7.0	New York	8.5	Wisconsin	5.4
Iowa	6.8	North Carolina	6.9	Wyoming	5.4
Kansas	8.6	North Dakota	6.8	District of Columbia	5.8
Kentucky	6.0	Ohio	7.1		

(a) Make a frequency histogram to display the data.

(b) What percent of values in the distribution are less than 6.0? Interpret this result in context.

SOLUTION:

(a)



Interval	Frequency
1.0 to <2.0	1
2.0 to <3.0	0
3.0 to <4.0	0
4.0 to <5.0	1
5.0 to <6.0	6
6.0 to <7.0	15
7.0 to <8.0	11
8.0 to <9.0	9
9.0 to <10.0	4

To make the histogram:

- **Choose equal-width intervals** that span the data. The data vary from 1.8 percent to 9.5 percent. So we choose intervals of width 1.0, starting at 1.0%.
- **Make a table.** Record the number of states in each interval to make a frequency histogram.
- **Draw and label the axes.** Don't forget units (percent) for the variable (tax rate).
- **Scale the axes.**
- **Draw bars.**

(b) $8/47 = 0.170 = 17.0\%$; 17% of the states (including the District of Columbia) have tax rates less than 6%.

FOR PRACTICE, TRY EXERCISE 67

Figure 1.11 shows two different histograms of the state sales tax data. Graph (a) uses the intervals of width 1% from the preceding example. The distribution has a single peak in the 6.0 to <7.0 interval. Graph (b) uses intervals half as wide: 1.0 to <1.5, 1.5 to <2.0, and so on. Now we see a distribution with more than one distinct peak. **The choice of intervals in a histogram can affect the appearance of a distribution.** Histograms with more intervals show more detail but may have a less clear overall pattern.

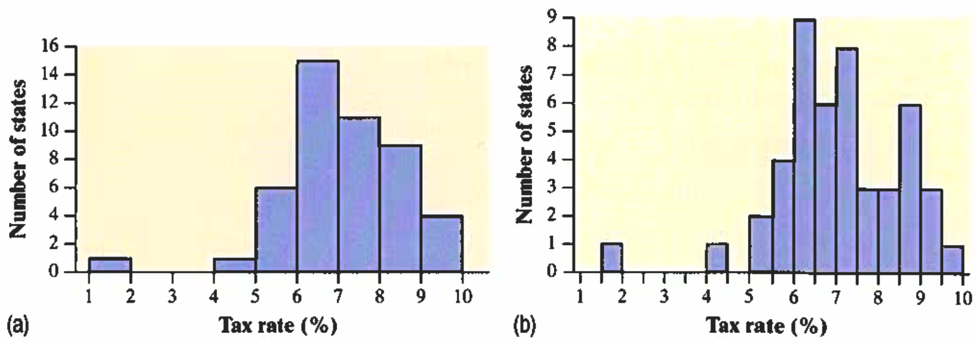


FIGURE 1.11 (a) Frequency histogram of the sales tax rate in the states that have local or state sales taxes and the District of Columbia with intervals of width 1.0%, from the preceding example. (b) Frequency histogram of the data with intervals of width 0.5%.

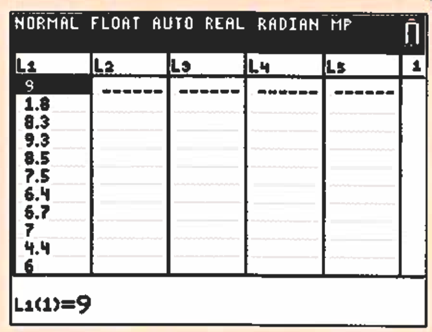
You can use a graphing calculator, statistical software, or an applet to make a histogram. The technology's default choice of intervals is a good starting point, but you should adjust the intervals to fit with common sense.

2. Technology Corner

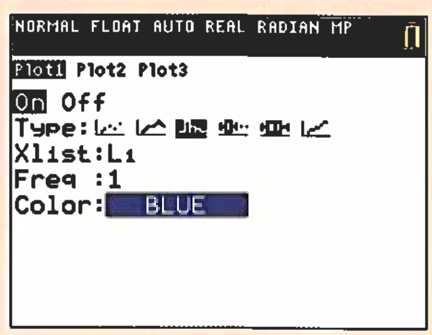
MAKING HISTOGRAMS

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

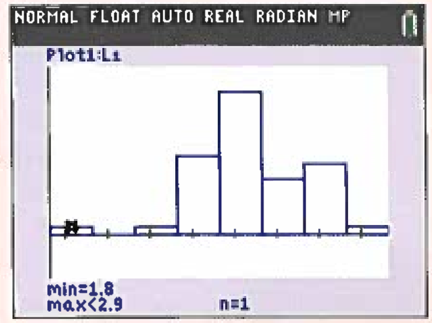
1. Enter the data from the sales tax example in your Statistics/List Editor.
   - Press STAT and choose Edit...
   - Type the values into list L1.



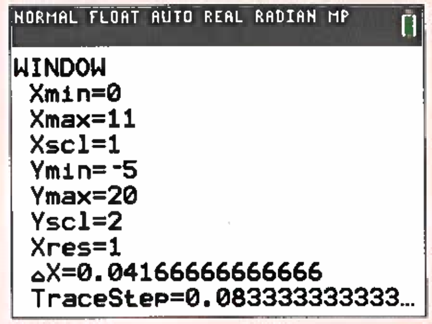
2. Set up a histogram in the Statistics Plots menu.
   - Press 2nd Y= (STAT PLOT).
   - Press ENTER or 1 to go into Plot1.
   - Adjust the settings as shown.



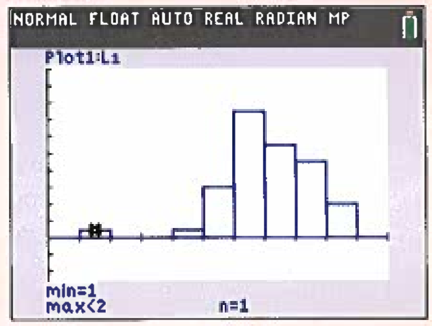
3. Use ZoomStat to let the calculator choose intervals and make a histogram.
   - Press ZOOM and choose ZoomStat.
   - Press TRACE to examine the intervals.



4. Adjust the intervals to match those in Figure 1.11(a), and then graph the histogram.
   - Press WINDOW and enter the values shown for Xmin, Xmax, Xscl, Ymin, Ymax, and Yscl.
   - Press GRAPH.
   - Press TRACE to examine the intervals.



5. See if you can match the histogram in Figure 1.11(b).

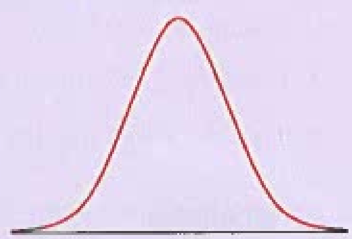


AP® EXAM TIP

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. 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 here.[24]



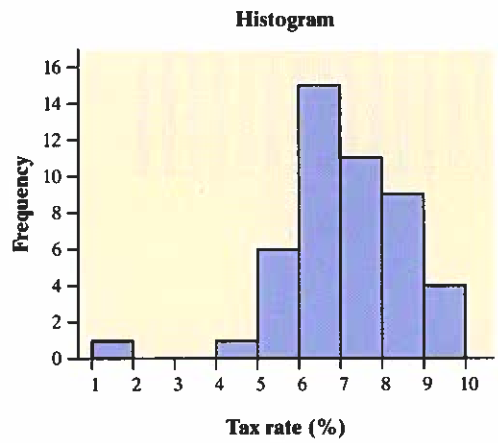
145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

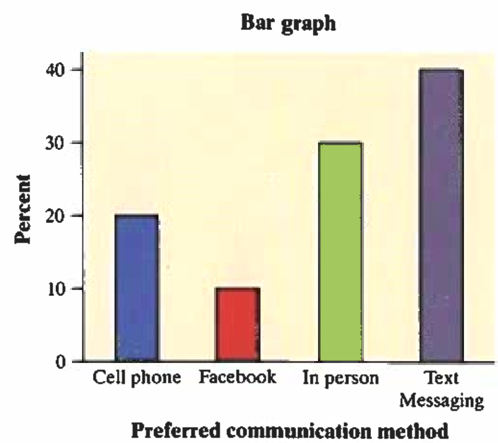
1. Construct a histogram that displays the distribution of IQ scores effectively.
2. Describe what you see. Is the distribution bell-shaped?

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. Its horizontal axis identifies intervals of values that the variable takes. A bar graph displays the distribution of a categorical variable. Its horizontal axis identifies the categories. Be sure to draw bar graphs with blank space between the bars to separate the categories. Draw histograms with no space between bars for adjacent intervals. For comparison, here is one of each type of graph from earlier examples:





2. **Use percents or proportions 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 biology journal and an airline magazine. She counted the number of letters in the first 400 words of an article in the journal and of the first 100 words of an article in the airline magazine. Mary then used statistical software to produce the histograms shown in Figure 1.12(a). This figure is misleading—it compares frequencies, but the two samples were of very different sizes (400 and 100). Using the same data, Mary's teacher produced the histograms in Figure 1.12(b). By using relative frequencies, this figure makes the comparison of word lengths in the two samples much easier.

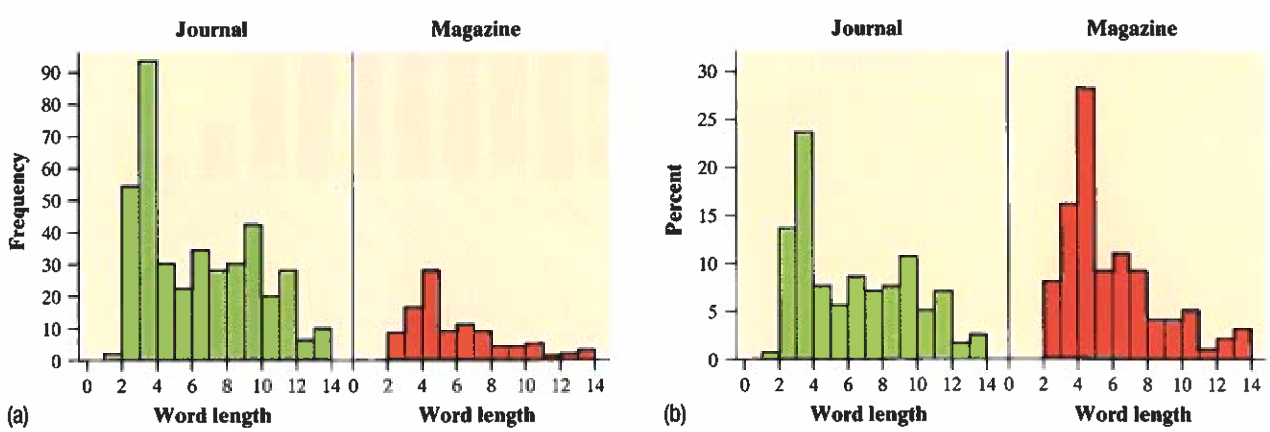
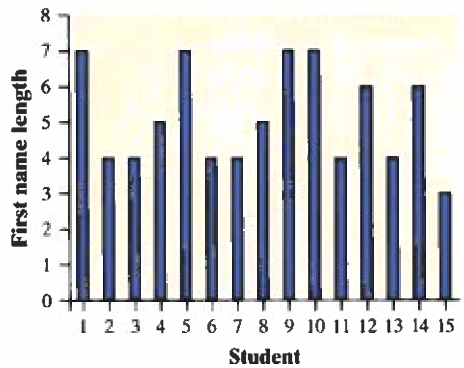
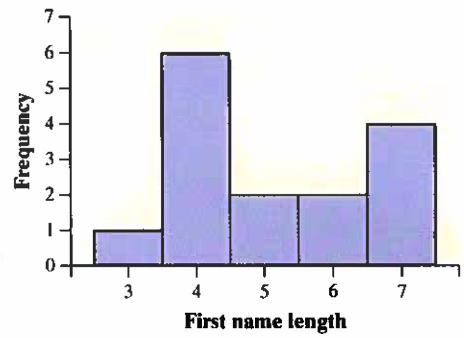


FIGURE 1.12 Two sets of histograms comparing word lengths in articles from a biology journal and from an airline magazine. In graph (a), the vertical scale uses frequencies. Graph (b) fixes the problem of different sample sizes by using percents (relative frequencies) on the vertical scale.

3. **Just because a graph looks nice doesn't make it a meaningful display of data.** The 15 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 on the left. What kind of graph is this? It's a bar graph that compares the raw data values. But first-name length is a quantitative variable, so a bar graph is not an appropriate way to display its distribution. The histogram on the right is a much better choice because the graph makes it easier to identify the shape, center, and variability of the distribution of name length.

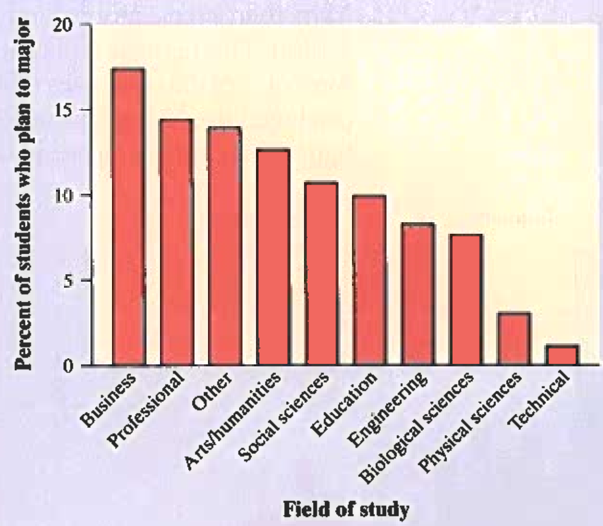




CHECK YOUR UNDERSTANDING

1. Write a few sentences comparing the distributions of word length shown in Figure 1.12(b).

Questions 2 and 3 refer to the following setting. About 3 million first-year students enroll in colleges and universities each year. What do they plan to study? The graph displays data on the percent of first-year students who plan to major in several disciplines.[25]



2. Is this a bar graph or a histogram? Explain.
3. 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 frequencies (counts) or relative frequencies (proportions or percents) of values in equal-width intervals.
- Some distributions have simple shapes, such as **symmetric, skewed to the left,** or **skewed to the right.** The number of peaks is another aspect of overall shape. So are distinct clusters and gaps.
- When examining any graph of quantitative data, look for an *overall pattern* and for clear *departures* from that pattern. **Shape, center,** and **variability** describe the overall pattern of the distribution of a quantitative variable. **Outliers** are observations that lie outside the overall pattern of a distribution.
- When comparing distributions of quantitative data, be sure to compare shape, center, variability, and possible outliers.
- Remember: histograms are for quantitative data; bar graphs are for categorical data. Be sure to use relative frequencies when comparing data sets of different sizes.

1.2 Technology Corner

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

2. Making histograms **Page 43**

Section 1.2 Exercises

45. Feeling sleepy? Students in a high school statistics class 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):

pg 31

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

(a) Make a dotplot to display the data.

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

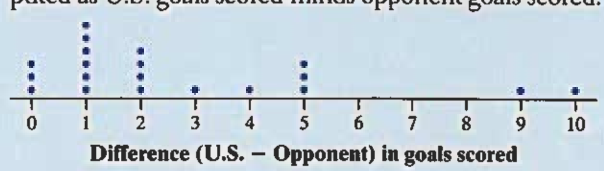
46. Easy reading? Here are data on the lengths of the first 25 words on a randomly selected page from Toni Morrison's *Song of Solomon*:

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

(a) Make a dotplot of these data.

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

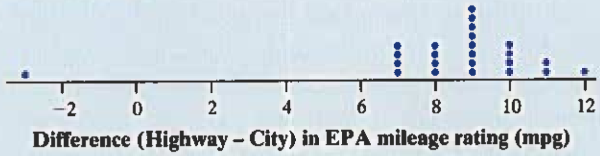
47. U.S. women's soccer—2016 Earlier, we examined data on the number of goals scored by the 2016 U.S. women's soccer team in 20 games played. The following dotplot displays the goal differential for those same games, computed as U.S. goals scored minus opponent goals scored.



(a) Explain what the dot above 3 represents.

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

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

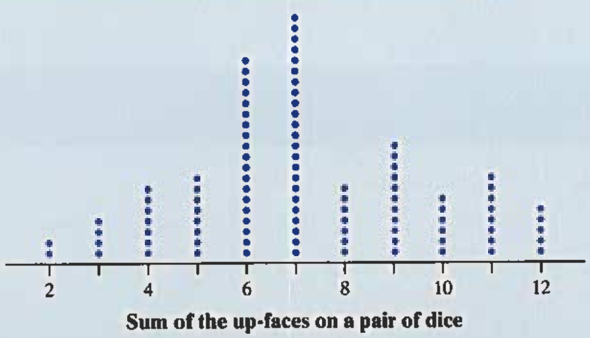


(a) Explain what the dot above −3 represents.

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

49. **Getting older** How old is the oldest person you know? pg 33 Prudential Insurance Company asked 400 people to place a blue sticker on a huge wall next to the age of the oldest person they have ever known. An image of the graph is shown here. Describe the shape of the distribution.

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



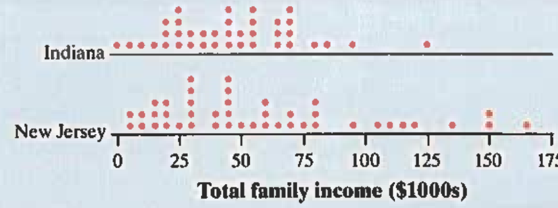
51. **Feeling sleepy?** Refer to Exercise 45. Describe the shape of the distribution.

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

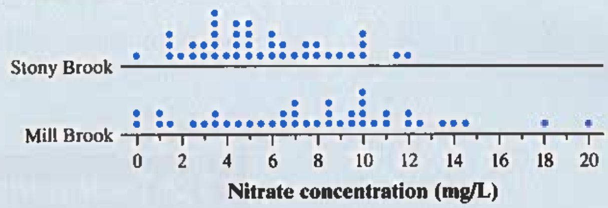
53. **U.S. women's soccer—2016** Refer to Exercise 47. pg 35 Describe the distribution.

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

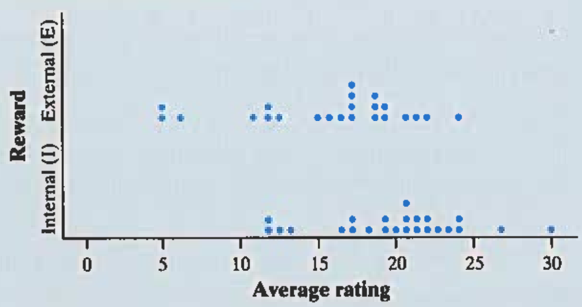
55. **Making money** The parallel dotplots show the total pg 36 family income of randomly chosen individuals from Indiana (38 individuals) and New Jersey (44 individuals). Compare the distributions of total family incomes in these two samples.



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



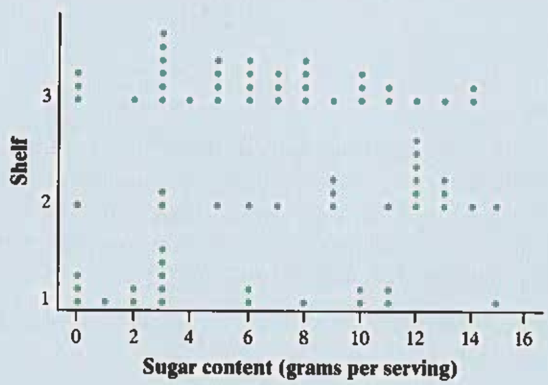
57. **Enhancing creativity** Do external rewards—things like money, praise, fame, and grades—promote creativity? Researcher Teresa Amabile recruited 47 experienced creative writers who were college students and divided them at random into two groups. The students in one group were given a list of statements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



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

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

58. Healthy cereal? Researchers collected data on 76 brands of cereal at a local supermarket.[27] 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 here.



(a) Is the variability in sugar content of the cereals on the three shelves similar or different? Justify your answer.

(b) Critics claim that supermarkets tend to put sugary cereals where kids can see them. Do the data from this study support this claim? Justify your answer. (Note that Shelf 2 is at about eye level for kids in most supermarkets.)

59. Snickers® are fun! Here are the weights (in grams) of 17 Snickers Fun Size bars from a single bag:

pg 38

17.1	17.4	16.6	17.4	17.7	17.1	17.3	17.7	17.8
19.2	16.0	15.9	16.5	16.8	16.5	17.1	16.7	

(a) Make a stemplot of these data.

(b) What interesting feature does the graph reveal?

(c) The advertised weight of a Snickers Fun Size bar is 17 grams. What proportion of candy bars in this sample weigh less than advertised?

60. Eat your beans! Beans and other legumes are a great source of protein. The following data give the protein content of 30 different varieties of beans, in grams per 100 grams of cooked beans.[28]

7.5	8.2	8.9	9.3	7.1	8.3	8.7	9.5	8.2	9.1
9.0	9.0	9.7	9.2	8.9	8.1	9.0	7.8	8.0	7.8
7.0	7.5	13.5	8.3	6.8	10.6	8.3	7.6	7.7	8.1

(a) Make a stemplot of these data.

(b) What interesting feature does the graph reveal?

(c) What proportion of these bean varieties contain more than 9 grams of protein per 100 grams of cooked beans?

61. South Carolina counties Here is a stemplot of the areas of the 46 counties in South Carolina. Note that the data have been rounded to the nearest 10 square miles (mi^2).

3	9999
4	0116689
5	01115566778
6	47899
7	01245579
8	0011
9	13
10	8
11	233
12	2

Key: 6|4 represents a county with an area of 635 to 644.99 square miles.

(a) What is the area of the largest South Carolina county?

(b) Describe the distribution of area for the 46 South Carolina counties.

62. Shopping spree The stemplot displays data on the amount spent by 50 shoppers at a grocery store. Note that the values have been rounded to the nearest dollar.

0	399
1	1345677889
2	000123455668888
3	25699
4	1345579
5	0359
6	1
7	0
8	366
9	3

Key: 9|3 = $92.50 to $93.49 spent

(a) What was the smallest amount spent by any of the shoppers?

(b) Describe the distribution of amount spent by these 50 shoppers.

63. Where do the young live? Here is a stemplot of the percent of residents aged 25 to 34 in each of the 50 states:

11	44
11	66778
12	0134
12	666778888
13	0000001111444
13	7788999
14	0044
14	567
15	11
15	
16	0

(a) Why did we split stems?

(b) Give an appropriate key for this stemplot.

(c) Describe the shape of the distribution. Are there any outliers?

64. **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.

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 of the distribution. Are there any outliers?

65. **Acorns and oak trees** Of the many species of oak trees in the United States, 28 grow on the Atlantic Coast and 11 grow in California. The back-to-back stemplot displays data on the average volume of acorns (in cubic centimeters) for these 39 oak species.[29] Write a few sentences comparing the distributions of acorn size for the oak trees in these two regions.

Atlantic Coast		California
998643	0	4
88864211111	1	06
50	2	06
6640	3	
8	4	1
	5	59
8	6	0
	7	1
1	8	
1	9	
5	10	
	11	
	12	
	13	
	14	
	15	
	16	
	17	1

Key: 2|6 = An oak species whose acorn volume is 2.6 cm³.

66. **Who studies more?** Researchers asked the students in a large first-year college class how many minutes they studied on a typical weeknight. The back-to-back stemplot displays the responses from random samples of 30 women and 30 men from the class, rounded to the nearest 10 minutes. Write a few sentences comparing the male and female distributions of study time.

Women		Men
	0	03333
96	0	56668999
22222222	1	02222222
888888888875555	1	558
4440	2	00344
	2	
	3	0
6	3	

Key: 2|3 = 230 minutes

67. **Carbon dioxide emissions** Burning fuels in power plants and motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. The table displays CO_2 emissions per person from countries with populations of at least 20 million.[30]

pg 42

(a) Make a histogram of the data using intervals of width 2, starting at 0.

(b) Describe the shape of the distribution. Which countries appear to be outliers?

Country	CO_2	Country	CO_2
Algeria	3.3	Mexico	3.8
Argentina	4.5	Morocco	1.6
Australia	16.9	Myanmar	0.2
Bangladesh	0.4	Nepal	0.1
Brazil	2.2	Nigeria	0.5
Canada	14.7	Pakistan	0.9
China	6.2	Peru	2.0
Colombia	1.6	Philippines	0.9
Congo	0.5	Poland	8.3
Egypt	2.6	Romania	3.9
Ethiopia	0.1	Russia	12.2
France	5.6	Saudi Arabia	17.0
Germany	9.1	South Africa	9.0
Ghana	0.4	Spain	5.8
India	1.7	Sudan	0.3
Indonesia	1.8	Tanzania	0.2
Iran	7.7	Thailand	4.4
Iraq	3.7	Turkey	4.1
Italy	6.7	Ukraine	6.6
Japan	9.2	United Kingdom	7.9
Kenya	0.3	United States	17.6
Korea, North	11.5	Uzbekistan	3.7
Korea, South	2.9	Venezuela	6.9
Malaysia	7.7	Vietnam	1.7

68. **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.[31]

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 to display the travel time data using intervals of width 2 minutes, starting at 14 minutes.

(b) Describe the shape of the distribution. What is the most common interval of travel times?

69. **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.[32] Their scores were as follows.

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

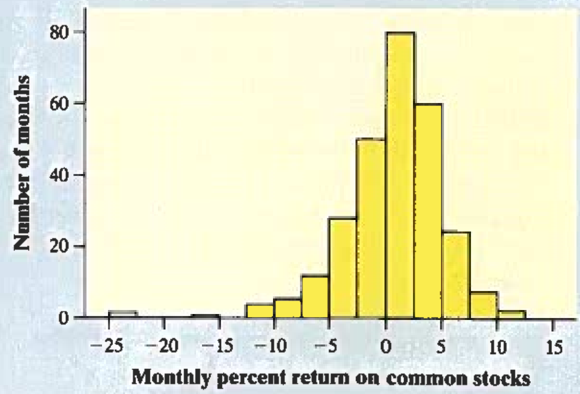
Make a histogram to display the data. Write a few sentences describing the distribution of DRP scores.

70. **Country music** The lengths, in minutes, of the 50 most popular mp3 downloads of songs by country artist Dierks Bentley are given here.

4.2	4.0	3.9	3.8	3.7	4.7
3.4	4.0	4.4	5.0	4.6	3.7
4.6	4.4	4.1	3.0	3.2	4.7
3.5	3.7	4.3	3.7	4.8	4.4
4.2	4.7	6.2	4.0	7.0	3.9
3.4	3.4	2.9	3.3	4.0	4.2
3.2	3.4	3.7	3.5	3.4	3.7
3.9	3.7	3.8	3.1	3.7	3.6
4.5	3.7				

Make a histogram to display the data. Write a few sentences describing the distribution of song lengths.

71. **Returns on common stocks** The return on a stock is the change in its market price plus any dividend payments made. Return is usually expressed as a percent of the beginning price. The figure shows a histogram of the distribution of monthly returns for the U.S. stock market over a 273-month period.[33]



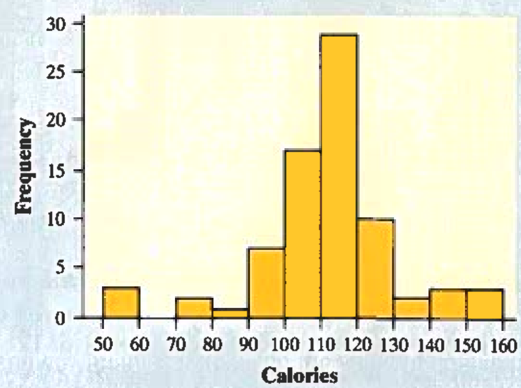
(a) Describe the overall shape of the distribution of monthly returns.

(b) What is the approximate center of this distribution?

(c) Explain why you cannot find the exact value for the minimum return. Between what two values does it lie?

(d) A return less than 0 means that stocks lost value in that month. About what percent of all months had returns less than 0?

72. **Healthy cereal?** Researchers collected data on calories per serving for 77 brands of breakfast cereal. The histogram displays the data.[34]



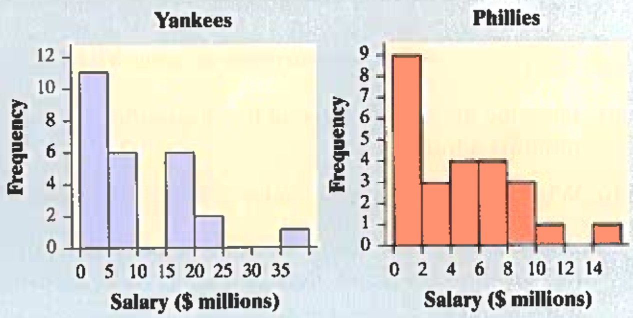
(a) Describe the overall shape of the distribution of calories.

(b) What is the approximate center of this distribution?

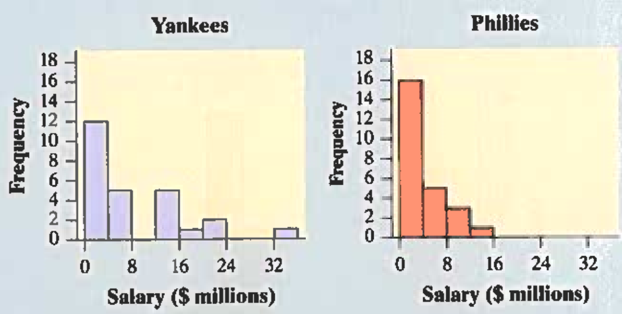
(c) Explain why you cannot find the exact value for the maximum number of calories per serving for these cereal brands. Between what two values does it lie?

(d) About what percent of the cereal brands have 130 or more calories per serving?

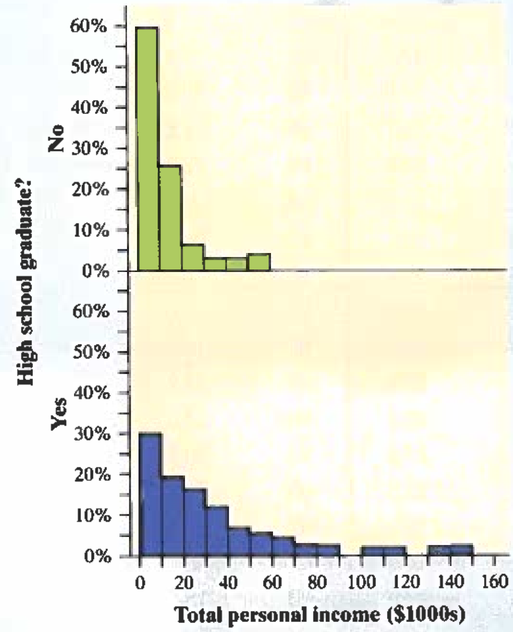
73. **Paying for championships** Does paying high salaries lead to more victories in professional sports? The New York Yankees have long been known for having Major League Baseball's highest team payroll. And over the years, the team has won many championships. This strategy didn't pay off in 2008, when the Philadelphia Phillies won the World Series. Maybe the Yankees didn't spend enough money that year. The figure shows histograms of the salary distributions for the two teams during the 2008 season. Why can't you use these graphs to effectively compare the team payrolls?



74. **Paying for championships** Refer to Exercise 73. Here is a better graph of the 2008 salary distributions for the Yankees and the Phillies. Write a few sentences comparing these two distributions.



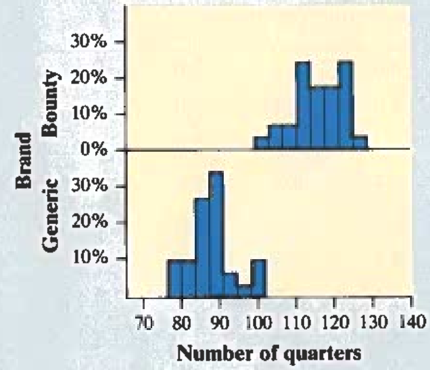
75. **Value of a diploma** Do students who graduate from high school earn more money than students who do not? To find out, we took a random sample of 371 U.S. residents aged 18 and older. The educational level and total personal income of each person were recorded. The data for the 57 non-graduates (No) and the 314 graduates (Yes) are displayed in the relative frequency histograms.



(a) Would it be appropriate to use frequency histograms instead of relative frequency histograms in this setting? Explain why or why not.

(b) Compare the distributions of total personal income for the two groups.

76. **Strong paper towels** In commercials for Bounty paper towels, the manufacturer claims that they are the "quicker picker-upper," but are they also the stronger picker-upper? Two of Mr. Tabor's statistics students, Wesley and Maverick, decided to find out. They selected a random sample of 30 Bounty paper towels and a random sample of 30 generic paper towels and measured their strength when wet. To do this, they uniformly soaked each paper towel with 4 ounces of water, held two opposite edges of the paper towel, and counted how many quarters each paper towel could hold until ripping, alternating brands. The data are displayed in the relative frequency histograms. Compare the distributions.



(a) Would it be appropriate to use frequency histograms instead of relative frequency histograms in this setting? Explain why or why not.

(b) Compare the distributions of number of quarters until breaking for the two paper towel brands.

77. **Birth months** Imagine asking a random sample of 60 students from your school about their birth months. Draw a plausible (believable) graph of the distribution of birth months. Should you use a bar graph or a histogram to display the data?

78. **Die rolls** Imagine rolling a fair, six-sided die 60 times. Draw a plausible graph of the distribution of die rolls. Should you use a bar graph or a histogram to display the data?

79. **AP® exam scores** The table gives the distribution of grades earned by students taking the AP® Calculus AB and AP® Statistics exams in 2016.[35]

	Grade					
	5	4	3	2	1	Total
Calculus AB	76,486	53,467	53,533	30,017	94,712	308,215
Statistics	29,627	44,884	51,367	32,120	48,565	206,563

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

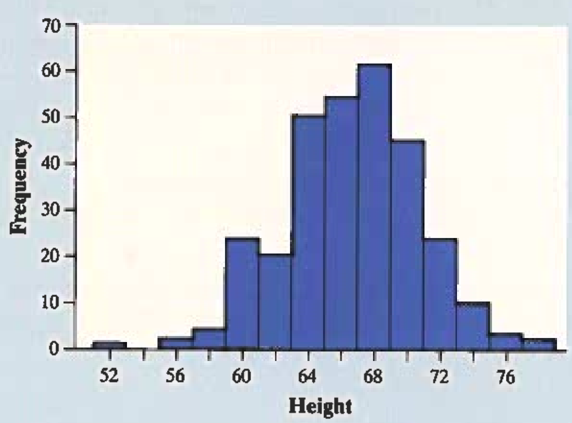
(b) Write a few sentences comparing the two distributions of exam grades.

Multiple Choice: *Select the best answer for Exercises 80–85.*

80. Here are the amounts of money (cents) in coins carried by 10 students in a statistics class: 50, 35, 0, 46, 86, 0, 5, 47, 23, 65. To make a stemplot of these data, you would use stems

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

81. The histogram shows the heights of 300 randomly selected high school students. Which of the following is the best description of the shape of the distribution of heights?

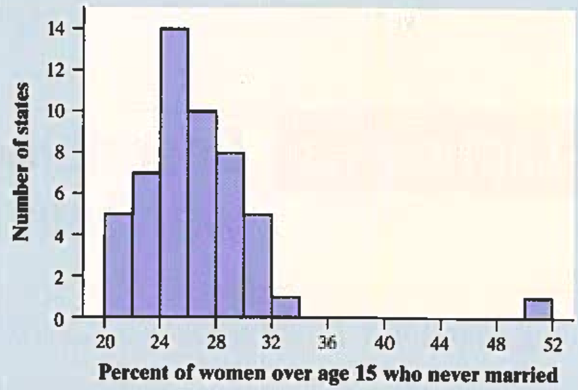


(a) Roughly symmetric and single-peaked
(b) Roughly symmetric and double-peaked
(c) Roughly symmetric and multi-peaked
(d) Skewed to the left
(e) Skewed to the right

82. 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, are priced 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) single-peaked.
(e) too high.

83. The histogram 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. Which of the following statements about the histogram is correct?



(a) The center (median) of the distribution is about 36%.
(b) There are more states with percentages above 32 than there are states with percentages less than 24.
(c) It would be better if the values from 34 to 50 were deleted on the horizontal axis so there wouldn't be a large gap.
(d) There was one state with a value of exactly 33%.
(e) About half of the states had percentages between 24% and 28%.

84. When comparing two distributions, it would be best to use relative frequency histograms rather than frequency histograms when

(a) the distributions have different shapes.
(b) the distributions have different amounts of variability.
(c) the distributions have different centers.
(d) the distributions have different numbers of observations.
(e) at least one of the distributions has outliers.

85. Which of the following is the best reason for choosing a stemplot rather than a histogram to display the distribution of a quantitative variable?

(a) Stemplots allow you to split stems; histograms don't.
(b) Stemplots allow you to see the values of individual observations.
(c) Stemplots are better for displaying very large sets of data.
(d) Stemplots never require rounding of values.
(e) Stemplots make it easier to determine the shape of a distribution.

Recycle and Review

86. **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:[36]

		Soccer level		
		Elite	Non-elite	Did not play
Whether person developed arthritis	Yes	10	9	24
	No	61	206	548

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

(b) What percent of the elite soccer players developed arthritis? What percent of those who got arthritis were elite soccer players?

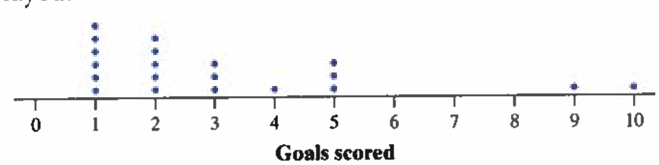
(c) Researchers suspected that the more serious soccer players were more likely to develop arthritis later in life. Do the data confirm this suspicion? Calculate appropriate percentages to support your answer.

SECTION 1.3 Describing Quantitative Data with Numbers

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

- Calculate measures of center (mean, median) for a distribution of quantitative data.
- Calculate and interpret measures of variability (range, standard deviation, *IQR*) for a distribution of quantitative data.
- Explain how outliers and skewness affect measures of center and variability.
- Identify outliers using the 1.5 × *IQR* rule.
- Make and interpret boxplots of quantitative data.
- Use boxplots and numerical summaries to compare distributions of quantitative data.

How much offense did the 2016 U.S. women's soccer team generate? The dotplot (reproduced from Section 1.2) shows the number of goals the team scored in 20 games played.



The distribution is right-skewed and single-peaked. The games in which the team scored 9 and 10 goals appear to be outliers. How can we describe the center and variability of this distribution?

Measuring Center: The Mean

The most common measure of center is the **mean**.

DEFINITION The mean $\bar{x}$

The **mean** $\bar{x}$ (pronounced "*x*-bar") of a distribution of quantitative data is the average of all the individual data values. To find the mean, add all the values and divide by the total number of observations.

If the n observations are $x_1, x_2, \ldots, x_n$, the mean is given by the formula

$$\bar{x} = \frac{\text{sum of data values}}{\text{number of data values}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

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

EXAMPLE

How many goals?
Calculating the mean

Kyodo News/ Getty Images

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

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

(a) Calculate the mean number of goals scored per game by the team. Show your work.

(b) The earlier description of these data (page 35) suggests that the games in which the team scored 9 and 10 goals are possible outliers. Calculate the mean number of goals scored per game by the team in the other 18 games that season. What do you notice?

SOLUTION:

(a) $$\bar{x} = \frac{5+5+1+10+5+2+1+1+2+3+3+2+1+4+2+1+2+1+9+3}{20}$$

$$= \frac{63}{20} = 3.15 \text{ goals}$$

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

(b) The mean for the other 18 games is

$$\bar{x} = \frac{5+5+1+5+2+1+1+2+3+3+2+1+4+2+1+2+1+3}{18}$$

$$= \frac{44}{18} = 2.44 \text{ goals}$$

These two games increased the team's mean number of goals scored per game by 0.71 goals.

FOR PRACTICE, TRY EXERCISE 87

The notation $\bar{x}$ refers to the mean of a *sample*. Most of the time, the data we 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 μ (Greek letter mu, pronounced "mew"). If you have the entire population of data available, then you calculate μ in just the way you'd expect: add the values of all the observations, and divide by the number of observations.

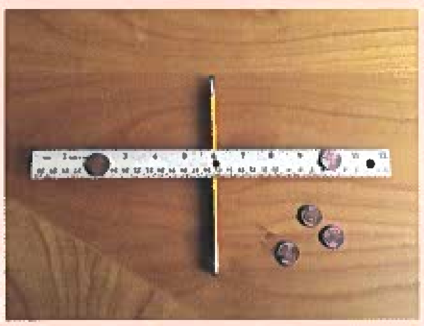
The preceding example illustrates an important weakness of the mean as a measure of center: **the mean is sensitive to extreme values in a distribution.** These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. We say that the mean is not a **resistant** measure of center.

> **DEFINITION Resistant**
>
> A statistical measure is **resistant** if it isn't sensitive to extreme values.

The mean of a distribution also has a physical interpretation, as the following activity shows.

ACTIVITY Mean as a "balance point"

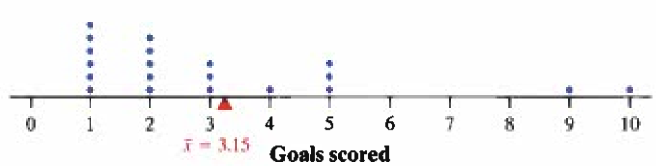
In this activity, you'll investigate an important property of the mean.



Ann Heath

1. Stack 5 pennies on top of the 6-inch mark on a 12-inch ruler. Place a 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, and 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. Discuss with your classmates: Why is the mean called the "balance point" of a distribution?

The activity gives a physical interpretation of the mean as the balance point of a distribution. For the data on goals scored in each of 20 games played by the 2016 U.S. women's soccer team, the dotplot balances at $\bar{x} = 3.15$ goals.



Measuring Center: The Median

We could also report the value in the "middle" of a distribution as its center. That's the idea of the **median**.

DEFINITION Median

The **median** is the midpoint of a distribution, the number such that about half the observations are smaller and about half are larger.

To find the median, arrange the data values from smallest to largest.

- If the number n of data values is odd, the median is the middle value in the ordered list.
- If the number n of data values is even, the median is the average of the two middle values in the ordered list.

The median is easy to find by hand for small sets of data. For instance, here are the data from Section 1.2 on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA:

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

Start by sorting the data values from smallest to largest:

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

There are $n = 25$ data values (an odd number), so the median is the middle (13th) value in the ordered list, the bold **22.4**.

EXAMPLE

How many goals?
Finding the median

Icon Sports Wire/Getty Images

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

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

Find the median.

SOLUTION:

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

The median is $\frac{2+2}{2} = 2$.

To find the median, sort the data values from smallest to largest. Because there are $n = 20$ data values (an even number), the median is the average of the middle two values in the ordered list.

FOR PRACTICE, TRY EXERCISE 89

Comparing the Mean and the Median

Which measure—the mean or the median—should we report as the center of a distribution? That depends on both the shape of the distribution and whether there are any outliers.

- **Shape:** Figure 1.13 shows the mean and median for dotplots with three different shapes. Notice how these two measures of center compare in each case. The mean is pulled in the direction of the long tail in a skewed distribution.

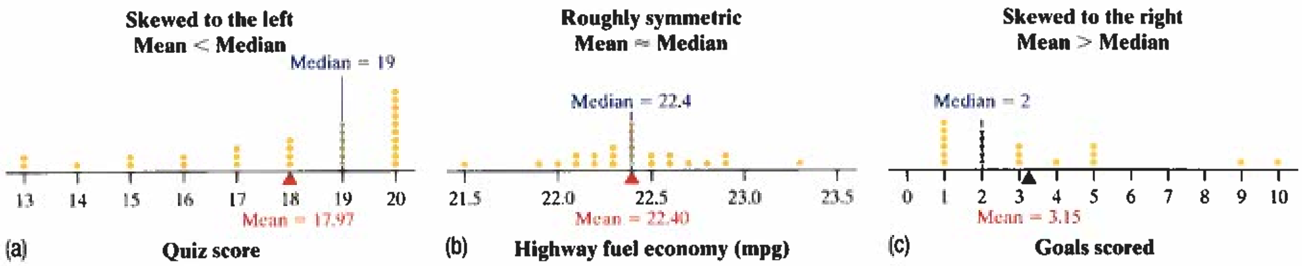


FIGURE 1.13 Dotplots that show the relationship between the mean and median in distributions with different shapes: (a) Scores of 30 statistics students on a 20-point quiz, (b) highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners, and (c) number of goals scored in 20 games played by the 2016 U.S. women's soccer team.

You can compare how the mean and median behave by using the *Mean and Median* applet at the book's website, highschool.bfwpub.com/tps6e.

- **Outliers:** We noted earlier that the mean is sensitive to extreme values. If we remove the two possible outliers (9 and 10) in Figure 1.13(c), the mean number of goals scored per game decreases from 3.15 to 2.44. The median number of goals scored is 2 whether we include these two games or not. The median is a resistant measure of center, but the mean is not.

EFFECT OF SKEWNESS AND OUTLIERS ON MEASURES OF CENTER

- If a distribution of quantitative data is roughly symmetric and has no outliers, the mean and median will be similar.
- If the distribution is strongly skewed, the mean will be pulled in the direction of the skewness but the median won't. For a right-skewed distribution, we expect the mean to be greater than the median. For a left-skewed distribution, we expect the mean to be less than the median.
- The median is resistant to outliers but the mean isn't.

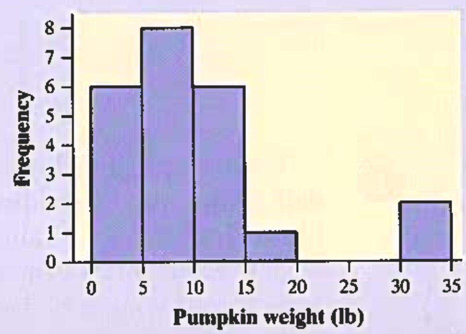
The mean and median measure center in different ways, and both are useful. In Major League Baseball (MLB), the distribution of player salaries is strongly skewed to the right. Most players earn close to the minimum salary (which was $507,500 in 2016), while a few earn more than $20 million. The median salary for MLB players in 2016 was about $1.5 million—but the mean salary was about

\$4.4 million. Clayton Kershaw, Miguel Cabrera, John Lester, and several other highly paid superstars pulled the mean up but that did not affect the median. The median gives us a good idea of what a "typical" MLB salary is. If we want to know the total salary paid to MLB players in 2016, however, we would multiply the mean salary by the total number of players: (\$4.4 million)(862) ≈ \$3.8 billion!

CHECK YOUR UNDERSTANDING

Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights in pounds are shown here, along with a histogram of the data.

3.6	4.0	9.6	14.0	11.0	12.4	13.0	2.0	6.0	6.6	15.0	3.4
12.7	6.0	2.8	9.6	4.0	6.1	5.4	11.9	5.4	31.0	33.0	



1. Calculate the mean weight of the pumpkins.
2. Find the median weight of the pumpkins.
3. Would you use the mean or the median to summarize the typical weight of a pumpkin in this contest? Explain.

Measuring Variability: The Range

Being able to describe the shape and center of a distribution is a great start. However, two distributions can have the same shape and center, but still look quite different.

Figure 1.14 shows comparative dotplots of the length (in millimeters) of separate random samples of PVC pipe from two suppliers, A and B.[37] Both distributions are roughly symmetric and single-peaked, with centers at about 600 mm, but the variability of these two distributions is quite different. The sample of pipes from Supplier A has much more consistent lengths (less variability) than the sample from Supplier B.

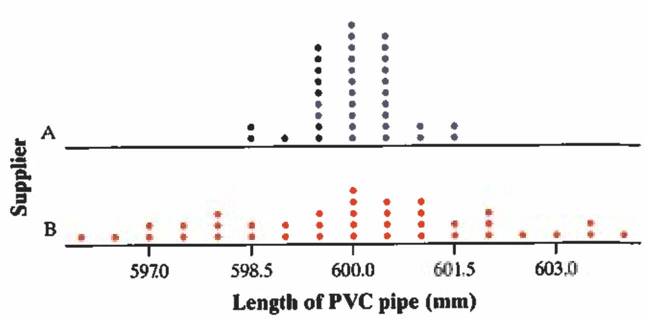


FIGURE 1.14 Comparative dotplots of the length of PVC pipes in separate random samples from Supplier A and Supplier B.

There are several ways to measure the variability of a distribution. The simplest is the **range**.

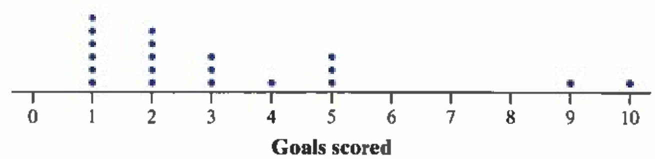
DEFINITION Range

The **range** of a distribution is the distance between the minimum value and the maximum value. That is,

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

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

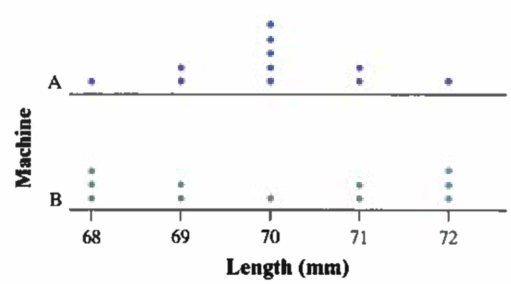
5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3



The range of this distribution is $10 - 1 = 9$ goals. Note that **the range of a data set is a single number.** In everyday language, people sometimes say things like, "The data values range from 1 to 10." A correct statement is "The number of goals scored in 20 games played by the 2016 U.S. women's soccer team varies from 1 to 10, a range of 9 goals."

The range is *not* a resistant measure of variability. It depends on only the maximum and minimum values, which may be outliers. Look again at the data on goals scored by the 2016 U.S. women's soccer team. Without the possible outliers at 9 and 10 goals, the range of the distribution would decrease to $5 - 1 = 4$ goals.

The following graph illustrates another problem with the range as a measure of variability. The parallel dotplots show the lengths (in millimeters) of a sample of 11 nails produced by each of two machines.[38] Both distributions are centered at 70 mm and have a range of $72 - 68 = 4$ mm. But the lengths of the nails made by Machine B clearly vary more from the center of 70 mm than the nails made by Machine A.



Measuring Variability: The Standard Deviation

If we summarize the center of a distribution with the mean, then we should use the **standard deviation** to describe the variation of data values around the mean.

DEFINITION Standard deviation

The **standard deviation** measures the typical distance of the values in a distribution from the mean.

How do we calculate the standard deviation s_x of a quantitative data set with n values? Here are the steps.

HOW TO CALCULATE THE STANDARD DEVIATION s_x

- Find the mean of the distribution.
- Calculate the *deviation* of each value from the mean: deviation = value − mean.
- Square each of the deviations.
- Add all the squared deviations, divide by $n - 1$, and take the square root.

If the values in a data set are $x_1, x_2, \ldots, x_n$, the standard deviation is given by the formula

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

AP® EXAM TIP

The formula sheet provided with the AP® Statistics exam gives the sample standard deviation in the equivalent form

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

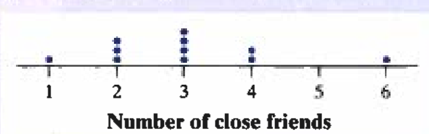
The notation s_x refers to the standard deviation of a *sample*. When we need to refer to the standard deviation of a population, we'll use the symbol σ (Greek lowercase sigma). The population standard deviation is calculated by dividing the sum of squared deviations by n instead of $n - 1$ before taking the square root.

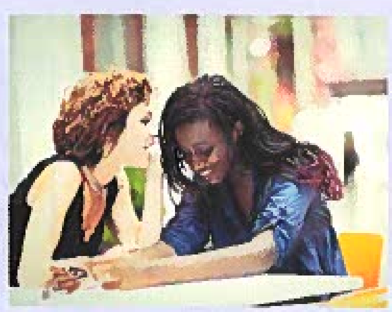
EXAMPLE

How many friends?
Calculating and interpreting standard deviation

PROBLEM: Eleven high school students were asked how many "close" friends they have. Here are their responses, along with a dotplot:

1 2 2 2 3 3 3 3 4 4 6





LaraBelova/Getty Images

Calculate the standard deviation. Interpret this value.

SOLUTION:

$$\bar{x} = \frac{1+2+2+2+3+3+3+3+4+4+6}{11} = 3$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	$1 - 3 = -2$	$(-2)^2 = 4$
2	$2 - 3 = -1$	$(-1)^2 = 1$
2	$2 - 3 = -1$	$(-1)^2 = 1$
2	$2 - 3 = -1$	$(-1)^2 = 1$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
4	$4 - 3 = 1$	$1^2 = 1$
4	$4 - 3 = 1$	$1^2 = 1$
6	$6 - 3 = 3$	$3^2 = 9$
		Sum = 18

To calculate the standard deviation:

- Find the mean of the distribution.
- Calculate the *deviation* of each value from the mean:
  deviation = value − mean
- Square each of the deviations.
- Add all the squared deviations, divide by $n - 1$, and take the square root to return to the original units.

$$s_x = \sqrt{\frac{18}{11-1}} = 1.34 \text{ close friends}$$

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

Interpretation: The number of close friends these students have typically varies by about 1.34 close friends from the mean of 3 close friends.

FOR PRACTICE, TRY EXERCISE 99

The value obtained before taking the square root in the standard deviation calculation is known as the *variance*. In the preceding example, the sample variance is

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{18}{11-1} = 1.80$$

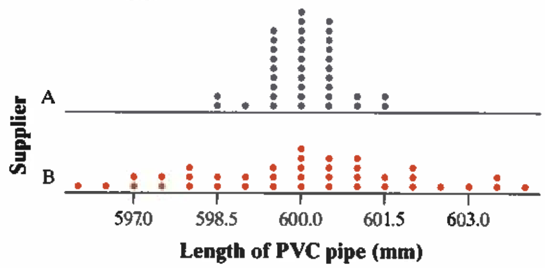
Unfortunately, the units are "squared close friends." Because variance is measured in squared units, it is not a very helpful way to describe the variability of a distribution.

Think About It

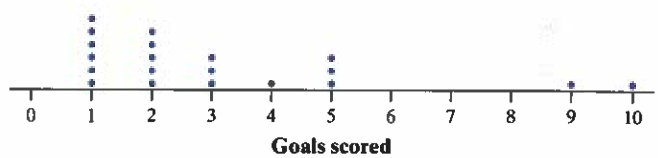
WHY IS THE STANDARD DEVIATION CALCULATED IN SUCH A COMPLEX WAY? Add the deviations from the mean in the preceding example. You should get a sum of 0. Why? Because the mean is the balance point of the distribution. We square the deviations to avoid the positive and negative deviations balancing each other out and adding to 0. It might seem strange to "average" the squared deviations by dividing by $n - 1$. We'll explain the reason for doing this in Chapter 7. It's easier to understand why we take the square root: to return to the original units (close friends).

More important than the details of calculating s_x are the properties of the standard deviation as a measure of variability:

- **s_x is always greater than or equal to 0.** $s_x = 0$ only when there is no variability, that is, when all values in a distribution are the same.
- **Larger values of s_x indicate greater variation** from the mean of a distribution. The comparative dotplot shows the lengths of PVC pipe in random samples from two different suppliers. Supplier A's pipe lengths have a standard deviation of 0.681 mm, while Supplier B's pipe lengths have a standard deviation of 2.02 mm. The lengths of pipes from Supplier B are typically farther from the mean than the lengths of pipes from Supplier A.



- **s_x is not a resistant measure of variability.** The use of squared deviations makes s_x even more sensitive than $\bar{x}$ to extreme values in a distribution. For example, the standard deviation of the number of goals scored in 20 games played by the 2016 U.S women's soccer team is 2.58 goals. If we omit the possible outliers of 9 and 10 goals, the standard deviation drops to 1.46 goals.



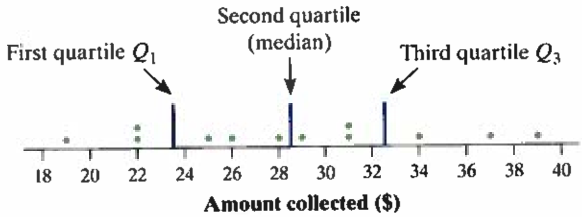
- **s_x measures variation about the mean.** It should be used only when the mean is chosen as the measure of center.

In the close friends example, 11 high school students had an average of $\bar{x} = 3$ close friends with a standard deviation of $s_x = 1.34$. What if a 12th high school student was added to the sample who had 3 close friends? The mean number of close friends in the sample would still be $\bar{x} = 3$. How would s_x be affected? Because the standard deviation measures the typical distance of the values in a distribution from the mean, s_x would *decrease* because this 12th value is at a distance of 0 from the mean. In fact, the new standard deviation would be

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{18}{12-1}} = 1.28 \text{ friends}$$

Measuring Variability: The Interquartile Range (*IQR*)

We can avoid the impact of extreme values on our measure of variability by focusing on the middle of the distribution. Start by ordering the data values from smallest to largest. Then find the **quartiles**, the values that divide the distribution into four groups of roughly equal size. 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. The first and third quartiles mark out the middle half of the distribution.



For example, here are the amounts collected each hour by a charity at a local store: \$19, \$26, \$25, \$37, \$31, \$28, \$22, \$22, \$29, \$34, \$39, and \$31. The dotplot displays the data. Because there are 12 data values, the quartiles divide the distribution into 4 groups of 3 values.

DEFINITION Quartiles, First quartile Q_1, Third quartile Q_3

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.

The **interquartile range** (*IQR*) measures the variability in the middle half of the distribution.

DEFINITION Interquartile range (*IQR*)

The **interquartile range** *(IQR)* is the distance between the first and third quartiles of a distribution. In symbols:

$$IQR = Q_3 - Q_1$$

Notice that the *IQR* is simply the range of the "middle half" of the distribution.

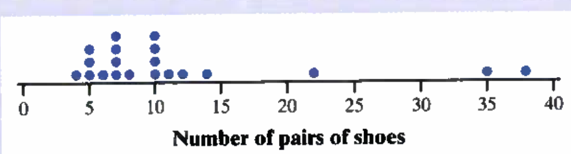
EXAMPLE

Boys and their shoes?
Finding the *IQR*

PROBLEM: How many pairs of shoes does a typical teenage boy own? To find out, two AP® Statistics students surveyed a random sample of 20 male students from their large high school and recorded the number of pairs of shoes that each boy owned. Here are the data, along with a dotplot:

Peter Cade/Getty Images

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



Find the interquartile range.

SOLUTION:

4 5 5 5 6 7 7 7 7 (8) | (10) 10 10 10 11 12 14 22 35 38
Median = 9

Sort the data values from smallest to largest and find the median.

4 5 5 5 (6) | (7) 7 7 7 8 | 10 10 10 10 (11) | (12) 14 22 35 38
$Q_1 = 6.5$ Median $Q_3 = 11.5$

Find the first quartile Q_1 and the third quartile Q_3.

$IQR = 11.5 - 6.5 = 5$ pairs of shoes

$IQR = Q_3 - Q_1$

FOR PRACTICE, TRY EXERCISE 105

The quartiles and the interquartile range are *resistant* because they are not affected by a few extreme values. For the shoe data, Q_3 would still be 11.5 and the IQR would still be 5 if the maximum were 58 rather than 38.

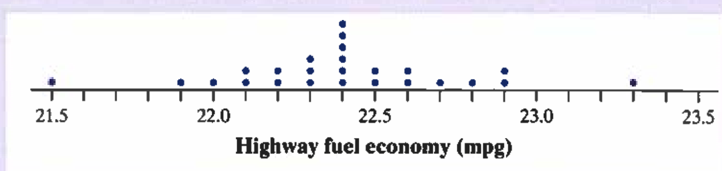
Be sure to leave out the median when you locate the quartiles. In the preceding example, the median was not one of the data values. For the earlier close friends data set, we ignore the circled median of 3 when finding Q_1 and Q_3.

1 2 (2) 2 3 | (3) | 3 3 (4) 4 6
Q_1 Median Q_3

CHECK YOUR UNDERSTANDING

Here are data on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA, along with a dotplot:

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



1. Find the range of the distribution.
2. The mean and standard deviation of the distribution are 22.404 mpg and 0.363 mpg, respectively. Interpret the standard deviation.
3. Find the interquartile range of the distribution.
4. Which measure of variability would you choose to describe the distribution? Explain.

Numerical Summaries with Technology

Graphing calculators and computer software will calculate numerical summaries for you. Using technology to perform calculations will allow you to focus on choosing the right methods and interpreting your results.

3. Technology Corner — COMPUTING NUMERICAL SUMMARIES

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

Let's find numerical summaries for the boys' shoes data from the example on page 64. We'll start by showing you how to compute summary statistics on the TI-83/84 and then look at output from computer software.

I. One-variable statistics on the TI-83/84

1. Enter the data in list L1.
2. Find the summary statistics for the shoe data.

- Press STAT (CALC); choose 1-VarStats.
  OS 2.55 or later: In the dialog box, press 2nd 1 (L1) and ENTER to specify L1 as the List. Leave FreqList blank. Arrow down to Calculate and press ENTER.
  Older OS: Press 2nd 1 (L1) and ENTER.
- Press ▼ to see the rest of the one-variable statistics.

```
NORMAL FLOAT AUTO REAL RADIAN MP
1-Var Stats
x̄=11.65
Σx=233
Σx²=4401
Sx=9.421559822
σx=9.183000599
n=20
minX=4
↓Q1=6.5
```

```
NORMAL FLOAT AUTO REAL RADIAN MP
1-Var Stats
↑Sx=9.421559822
σx=9.183000599
n=20
minX=4
Q1=6.5
Med=9
Q3=11.5
maxX=38
```

II. Output from statistical software We used Minitab statistical software to calculate descriptive statistics for the boys' shoes data. Minitab allows you to choose which numerical summaries are included in the output.

Descriptive Statistics: Shoes

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Shoes	20	11.65	9.42	4.00	6.25	9.00	11.75	38.00

Note: The TI-83/84 gives the first and third quartiles of the boys' shoes distribution as $Q_1 = 6.5$ and $Q_3 = 11.5$. Minitab reports that $Q_1 = 6.25$ and $Q_3 = 11.75$. What happened? Minitab and some other software use slightly different rules for locating quartiles. Results from the various rules are usually close to each other. Be aware of possible differences when calculating quartiles as they may affect more than just the *IQR*.

Identifying Outliers

Besides serving as a measure of variability, the interquartile range (*IQR*) is used as a "ruler" for identifying outliers.

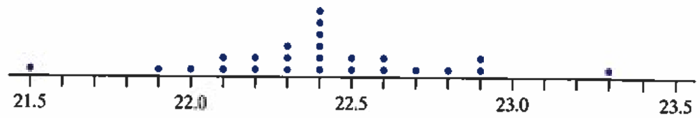
HOW TO IDENTIFY OUTLIERS: THE 1.5 × *IQR* RULE

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

Low outliers $< Q_1 - 1.5 \times IQR$ High outliers $> Q_3 + 1.5 \times IQR$

Here are sorted data on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA, along with a dotplot:

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



Does the $1.5 \times IQR$ rule identify any outliers in this distribution? If you did the preceding Check Your Understanding, you should have found that $Q_1 = 22.2$ mpg, $Q_3 = 22.6$ mpg, and $IQR = 0.4$ mpg. For these data,

$$\text{High outliers} > Q_3 + 1.5 \times IQR = 22.6 + 1.5 \times 0.4 = 23.2$$

and

$$\text{Low outliers} < Q_1 - 1.5 \times IQR = 22.2 - 1.5 \times 0.4 = 21.6$$

The cars with estimated highway fuel economy ratings of 21.5 and 23.3 are identified as outliers.

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.

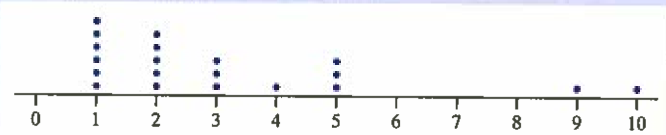
EXAMPLE

How many goals?
Identifying outliers


Icon Sports Wire/Getty Images

PROBLEM: Here are sorted data on the number of goals scored in 20 games played by the 2016 U.S women's soccer team, along with a dotplot:

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



Identify any outliers in the distribution. Show your work.

SOLUTION:

1 1 1 1 (1 1) 2 2 2 (2 | 2) 3 3 3 (4 5) 5 5 9 10

$Q_1 = 1$ Median $= 2$ $Q_3 = 4.5$

$IQR = Q_3 - Q_1 = 4.5 - 1 = 3.5$

$\text{Low outliers} < Q_1 - 1.5 \times IQR = 1 - 1.5 \times 3.5 = -4.25$

$\text{High outliers} > Q_3 + 1.5 \times IQR = 4.5 + 1.5 \times 3.5 = 9.75$

There are no data values less than -4.25, but the game in which the team scored 10 goals is an outlier.

The game in which the team scored 9 goals is not identified as an outlier by the $1.5 \times IQR$ rule.

FOR PRACTICE, TRY EXERCISE 107

It is important to identify outliers in a distribution for several reasons:

1. **They might be inaccurate data values.** Maybe someone recorded a value as 10.1 instead of 101. Perhaps a measuring device broke down. Or maybe someone gave a silly response, like the student in a class survey who claimed to study 30,000 minutes per night! Try to correct errors like these if possible. If you can't, give summary statistics with and without the outlier.
2. **They can indicate a remarkable occurrence.** For example, in a graph of net worth, Bill Gates is likely to be an outlier.
3. **They can heavily influence the values of some summary statistics,** like the mean, range, and standard deviation.

Making and Interpreting Boxplots

You can use a dotplot, stemplot, or histogram to display the distribution of a quantitative variable. Another graphical option for quantitative data is a **boxplot.** A boxplot summarizes a distribution by displaying the location of 5 important values within the distribution, known as its **five-number summary.**

A boxplot is sometimes called a *box-and-whisker* plot.

DEFINITION Five-number summary, Boxplot

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1, the median, the third quartile Q_3, and the maximum.

A **boxplot** is a visual representation of the five-number summary.

Figure 1.15 illustrates the process of making a boxplot. The dotplot in Figure 1.15(a) shows the data on EPA estimated highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners. We have marked the first quartile, the median, and the third quartile with vertical blue lines. The process of testing for outliers with the $1.5 \times IQR$ rule is shown in red. Because the values of 21.5 mpg and 23.3 mpg are outliers, we mark these separately. To get the finished boxplot in Figure 1.15(b), we make a box spanning from Q_1 to Q_3 and then draw "whiskers" to the smallest and largest data values that are not outliers

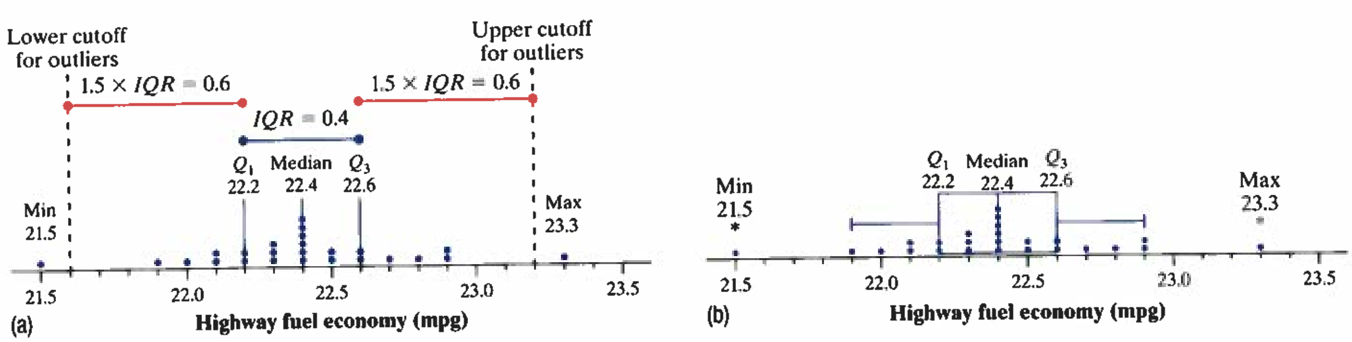


FIGURE 1.15 A visual illustration of how to make a boxplot for the Toyota 4Runner highway gas mileage data. (a) Dotplot of the data with the five-number summary and $1.5 \times IQR$ marked. (b) Boxplot of the data with outliers identified (*).

As you can see, it is fairly easy to make a boxplot by hand for small sets of data. Here's a summary of the steps.

HOW TO MAKE A BOXPLOT

- **Find the five-number summary** for the distribution.
- **Identify outliers** using the $1.5 \times IQR$ rule.
- **Draw and label the axis.** Draw a horizontal axis and put the name of the quantitative variable underneath, including units if applicable.
- **Scale the axis.** Look at the minimum and maximum values in the data set. Start the horizontal axis at a convenient number equal to or below the minimum and place tick marks at equal intervals until you equal or exceed the maximum.
- **Draw a box** that spans from the first quartile (Q_1) to the third quartile (Q_3).
- **Mark the median** with a vertical line segment that's the same height as the box.
- **Draw whiskers**—lines that extend from the ends of the box to the smallest and largest data values that are *not* outliers. Mark any outliers with a special symbol such as an asterisk (*).

We see from the boxplot in Figure 1.15 that the distribution of highway gas mileage ratings for this sample of model year 2018 Toyota 4Runners is roughly symmetric with one high outlier and one low outlier.

EXAMPLE

Picking pumpkins
Making and interpreting boxplots


cscredon/Getty Images

PROBLEM: Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights in pounds are shown here.

3.6	4.0	9.6	14.0	11.0	12.4	13.0	2.0	6.0	6.6	15.0	3.4
12.7	6.0	2.8	9.6	4.0	6.1	5.4	11.9	5.4	31.0	33.0	

(a) Make a boxplot of the data.

(b) Explain why the median and *IQR* would be a better choice for summarizing the center and variability of the distribution of pumpkin weights than the mean and standard deviation.

SOLUTION:

(a)

Min (2.0), Q_1 (4.0), Median (6.6), Q_3 (12.7), Max (33.0)

2.0	2.8	3.4	3.6	4.0	(4.0)	5.4	5.4	6.0	6.0	6.1
9.6	9.6	11.0	11.9	12.4	(12.7)	13.0	14.0	15.0	31.0	33.0

Median: (6.6)

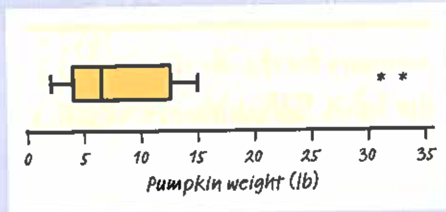
$IQR = Q_3 - Q_1 = 12.7 - 4.0 = 8.7$

Low outliers $< Q_1 - 1.5 \times IQR = 4.0 - 1.5 \times 8.7 = -9.05$

High outliers $> Q_3 + 1.5 \times IQR = 12.7 + 1.5 \times 8.7 = 25.75$

The pumpkins that weighed 31.0 and 33.0 pounds are outliers.

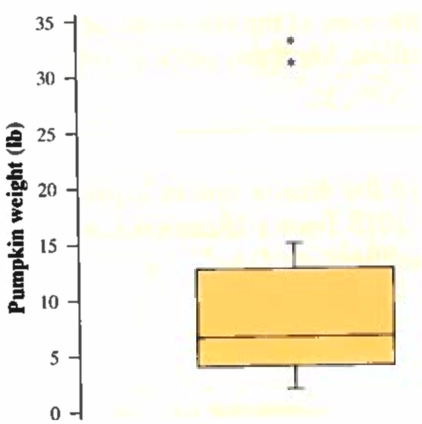
To make the boxplot:
- **Find the five-number summary.**
- **Identify outliers.**
- **Draw and label the axis.**
- **Scale the axis.**
- **Draw a box.**
- **Mark the median.**
- **Draw whiskers** to the smallest and largest data values that are *not* outliers. Mark outliers with an asterisk.



(b) The distribution of pumpkin weights is skewed to the right with two high outliers. Because the mean and standard deviation are sensitive to outliers, it would be better to use the median and *IQR*, which are resistant.

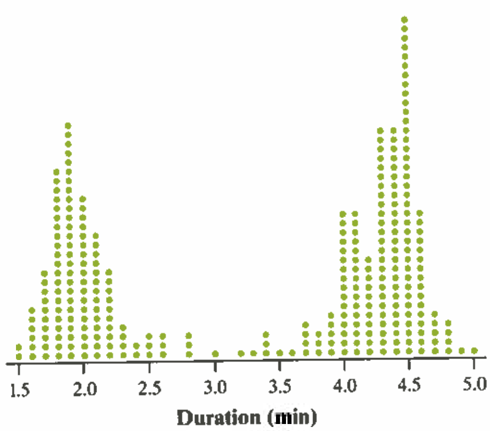
We know the distribution is skewed to the right because the left half of the distribution varies from 2.0 to 6.6 pounds, while the right half of the distribution (excluding outliers) varies from 6.6 to 15.0 pounds.

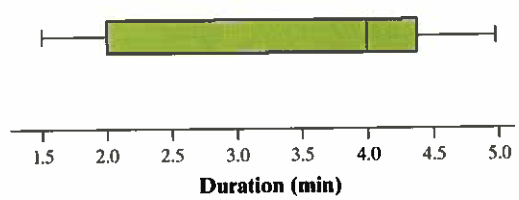
FOR PRACTICE, TRY EXERCISE 111



Boxplots provide a quick summary of the center and variability of a distribution. The median is displayed as a line in the central box, the interquartile range is the length of the box, and the range is the length of the entire plot, including outliers. Note that some statistical software orients boxplots vertically. At left is a vertical boxplot of the pumpkin weight data from the preceding example. You can see that the graph is skewed toward the larger values.

Boxplots do not display each individual value in a distribution. And **boxplots don't show gaps, clusters, or peaks**. For instance, the dotplot below left displays the duration, in minutes, of 220 eruptions of the Old Faithful geyser. The distribution of eruption durations is clearly double-peaked (*bimodal*). But a boxplot of the data hides this important information about the shape of the distribution.





CHECK YOUR UNDERSTANDING

Ryan and Brent were curious about the amount of french fries they would get in a large order from their favorite fast-food restaurant, Burger King. They went to several different Burger King locations over a series of days and ordered a total of 14 large fries. The weight of each order (in grams) is as follows:

165 163 160 159 166 152 166 168 173 171 168 167 170 170

1. Make a boxplot to display the data.
2. According to a nutrition website, Burger King's large fries weigh 160 grams, on average. Ryan and Brent suspect that their local Burger King restaurants may be skimping on fries. Does the boxplot in Question 1 support their suspicion? Explain why or why not.

Comparing Distributions with Boxplots

Boxplots are especially effective for comparing the distribution of a quantitative variable in two or more groups, as seen in the following example.

EXAMPLE

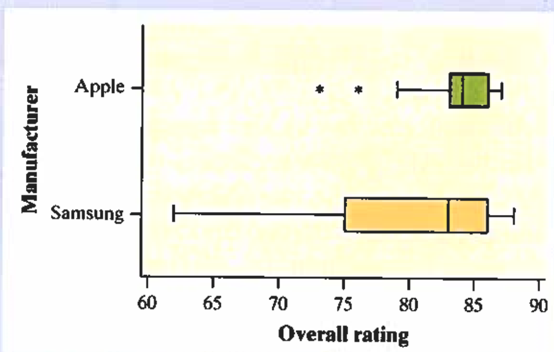
Which company makes better tablets?

Comparing distributions with boxplots

PROBLEM: In a recent year, *Consumer Reports* rated many tablet computers for performance and quality. Based on several variables, the magazine gave each tablet an overall rating, where higher scores indicate better ratings. The overall ratings of the tablets produced by Apple and Samsung are given here, along with parallel boxplots and numerical summaries of the data.[39]


Peter Cade/Getty Images

Apple	87	87	87	87	86	86	86	86	84	84		
	84	84	83	83	83	83	81	79	76	73		
Samsung	88	87	87	86	86	86	86	86	84	84	83	83
	77	76	76	75	75	75	75	75	74	71	62	



	$\bar{x}$	s_x	Min	Q_1	Median	Q_3	Max	*IQR*
Apple	83.45	3.762	73	83	84	86	87	3
Samsung	79.87	6.74	62	75	83	86	88	11

Compare the distributions of overall rating for Apple and Samsung.

SOLUTION:

Shape: Both distributions of overall ratings are skewed to the left.

Outliers: There are two low outliers in the Apple tablet distribution: overall ratings of 73 and 76. The Samsung tablet distribution has no outliers.

Center: The Apple tablets had a slightly higher median overall rating (84) than the Samsung tablets (83). More importantly, about 75% of the Apple tablets had overall ratings that were greater than or equal to the median for the Samsung tablets.

Variability: There is much more variation in overall rating among the Samsung tablets than the Apple tablets. The *IQR* for Samsung tablets (11) is almost four times larger than the *IQR* for Apple tablets (3).

Remember to compare shape, outliers, center, and variability!

Because of the strong skewness and outliers, use the median and *IQR* instead of the mean and standard deviation when comparing center and variability.

FOR PRACTICE, TRY EXERCISE 115

AP® EXAM TIP

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

Here's an activity that gives you a chance to put into practice what you have learned in this section.

ACTIVITY Team challenge: Did Mr. Starnes stack his class?

In this activity, you will work in a team of three or four students to resolve a dispute.

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 following table shows the data for each student in the two classes.

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

Student	Teacher	GPA	SAT-M
1	Starnes	2.900	670
2	Starnes	2.860	520
3	Starnes	2.600	570
4	Starnes	3.600	710
5	Starnes	3.200	600
6	Starnes	2.700	590
7	Starnes	3.100	640
8	Starnes	3.085	570
9	Starnes	3.750	710
10	Starnes	3.400	630
11	Starnes	3.338	630
12	Starnes	3.560	670
13	Starnes	3.800	650
14	Starnes	3.200	660
15	Starnes	3.100	510
16	McGrail	2.900	620
17	McGrail	3.300	590
18	McGrail	3.980	650
19	McGrail	2.900	600
20	McGrail	3.200	620
21	McGrail	3.500	680
22	McGrail	2.800	500
23	McGrail	2.900	502.5
24	McGrail	3.950	640
25	McGrail	3.100	630
26	McGrail	2.850	580
27	McGrail	2.900	590
28	McGrail	3.245	600
29	McGrail	3.000	600
30	McGrail	3.000	620
31	McGrail	2.800	580
32	McGrail	2.900	600
33	McGrail	3.200	600

You can use technology to make boxplots, as the following Technology Corner illustrates.

4. Technology Corner MAKING BOXPLOTS

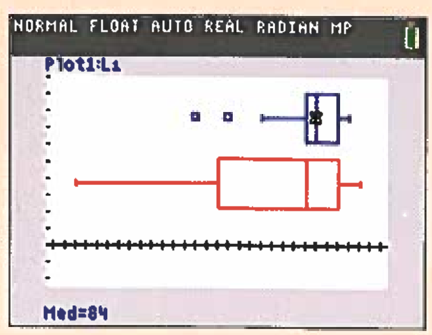
TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

The TI-83/84 can plot up to three boxplots in the same viewing window. Let's use the calculator to make parallel boxplots of the overall rating data for Apple and Samsung tablets.

1. Enter the ratings for Apple tablets in list L1 and for Samsung in list L2.
2. Set up two statistics plots: Plot1 to show a boxplot of the Apple data in list L1 and Plot2 to show a boxplot of the Samsung data in list L2. The setup for Plot1 is shown. When you define Plot2, be sure to change L1 to L2.

Note: The calculator offers two types of boxplots: one that shows outliers and one that doesn't. We'll always use the type that identifies outliers.

3. Use the calculator's Zoom feature to display the parallel boxplots. Then Trace to view the five-number summary.
   - Press ZOOM and select ZoomStat.
   - Press TRACE.



Section 1.3 Summary

- A numerical summary of a distribution should include measures of **center** and **variability.**
- The **mean** $\bar{x}$ and the **median** describe the center of a distribution in different ways. The mean is the average of the observations: $\bar{x} = \frac{\sum x_i}{n}$. The median is the midpoint of the distribution, the number such that about half the observations are smaller and half are larger.
- The simplest measure of variability for a distribution of quantitative data is the **range,** which is the distance from the maximum value to the minimum value.
- When you use the mean to describe the center of a distribution, use the **standard deviation** to describe the distribution's variability. The standard deviation s_x gives the typical distance of the values in a distribution from the mean. In symbols, $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$. The standard deviation s_x is 0 when there is no variability and gets larger as variability from the mean increases.
- When you use the median to describe the center of a distribution, use the **interquartile range** to describe the distribution's variability. 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) measures variability in the middle half of the distribution and is found using $IQR = Q_3 - Q_1$.
- The median is a **resistant** measure of center because it is relatively unaffected by extreme observations. The mean is not resistant. Among measures of variability, the IQR is resistant, but the standard deviation and range are not.
- According to the **$1.5 \times IQR$ rule**, an observation is an outlier if it is less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$.
- **Boxplots** are based on the **five-number summary** of a distribution, consisting of the minimum, Q_1, the median, Q_3, and the maximum. The box shows the variability in the middle 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 with special symbols. Boxplots are especially useful for comparing distributions.

1.3 Technology Corners

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

Section 1.3 Exercises

87. Quiz grades Joey's first 14 quiz grades in a marking period were as follows:

pg 55

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

(a) Calculate the mean. Show your work.

(b) Suppose Joey has an unexcused absence for the 15th quiz, and he receives a score of 0. Recalculate the mean. What property of the mean does this illustrate?

88. Pulse rates Here are data on the resting pulse rates (in beats per minute) of 19 middle school students:

71	104	76	88	78	71	68	86	70	90
74	76	69	68	88	96	68	82	120	

(a) Calculate the mean. Show your work.

(b) The student with a 120 pulse rate has a medical issue. Find the mean pulse rate for the other 18 students. What property of the mean does this illustrate?

89. Quiz grades Refer to Exercise 87.

pg 57

(a) Find the median of Joey's first 14 quiz grades.

(b) Find the median of Joey's quiz grades after his unexcused absence. Explain why the 0 quiz grade does not have much effect on the median.

90. Pulse rates Refer to Exercise 88.

(a) Find the median pulse rate for all 19 students.

(b) Find the median pulse rate excluding the student with the medical issue. Explain why this student's 120 pulse rate does not have much effect on the median.

91. Electing the president 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 College votes. Here is a stemplot of the number of electoral votes in 2016 for each of the 50 states and the District of Columbia:

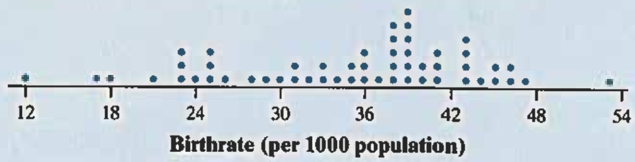
```
0 | 3333333344444
0 | 5556666667778899 9
1 | 00001111234
1 | 5668
2 | 00
2 | 99
3 |
3 | 8
4 |
4 |
5 |
5 | 5
```

Key: 1|5 is a state with 15 electoral votes.

(a) Find the median.

(b) Without doing any calculations, explain how the mean and median compare.

92. Birthrates in Africa One of the important factors in determining population growth rates is the birthrate per 1000 individuals in a population. The dotplot shows the birthrates per 1000 individuals (rounded to the nearest whole number) for 54 African nations.



(a) Find the median.

(b) Without doing any calculations, explain how the mean and median compare.

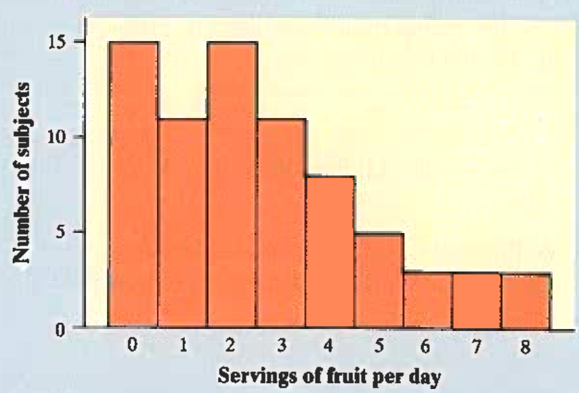
93. House prices The mean and median selling prices of existing single-family homes sold in September 2016 were $276,200 and $234,200.[40] Which of these numbers is the mean and which is the median? Explain your reasoning.

94. Mean salary? Last year a small accounting firm paid each of its five clerks $32,000, two junior accountants $60,000 each, and the firm's owner $280,000.

(a) What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

(b) Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

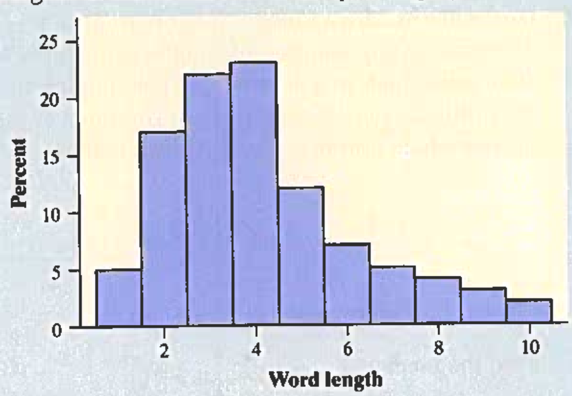
95. Do adolescent girls eat fruit? We all know that fruit is good for us. Here is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania:[41]



(a) Find the median number of servings of fruit per day from the histogram. Explain your method clearly.

(b) Calculate the mean of the distribution. Show your work.

96. **Shakespeare** The histogram shows the distribution of lengths of words used in Shakespeare's plays.[42]



(a) Find the median word length in Shakespeare's plays from the histogram. Explain your method clearly.

(b) Calculate the mean of the distribution. Show your work.

97. **Quiz grades** Refer to Exercise 87.

(a) Find the range of Joey's first 14 quiz grades and the range of Joey's quiz grades after his unexcused absence.

(b) Explain what part (a) suggests about using the range as a measure of variability for a distribution of quantitative data.

98. **Pulse rates** Refer to Exercise 88.

(a) Find the range of the pulse rates for all 19 students and the range of the pulse rates excluding the student with the medical issue.

(b) Explain what part (a) suggests about using the range as a measure of variability for a distribution of quantitative data.

99. **Foot lengths** Here are the foot lengths (in centimeters) for a random sample of seven 14-year-olds from the United Kingdom:

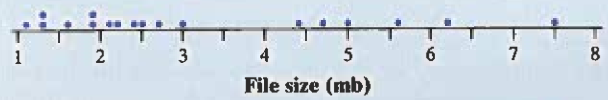
pg 61

25	22	20	25	24	24	28

Calculate the standard deviation. Interpret this value.

100. **Well rested?** A random sample of 6 students in a first-period statistics class was asked how much sleep (to the nearest hour) they got last night. Their responses were 6, 7, 7, 8, 10, and 10. Calculate the standard deviation. Interpret this value.

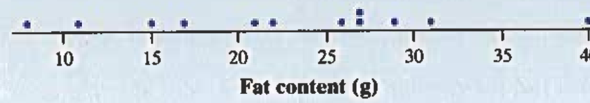
101. **File sizes** How much storage space does your music use? Here is a dotplot of the file sizes (to the nearest tenth of a megabyte) for 18 randomly selected files on Nathaniel's mp3 player:



(a) The distribution of file size has a mean of $\bar{x} = 3.2$ megabytes and a standard deviation of $s_x = 1.9$ megabytes. Interpret the standard deviation.

(b) Suppose the music file that takes up 7.5 megabytes of storage space is replaced with another version of the file that only takes up 4 megabytes. How would this affect the mean and the standard deviation? Justify your answer.

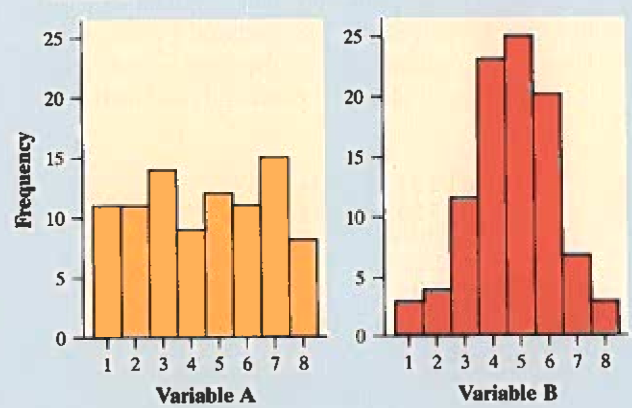
102. **Healthy fast food?** Here is a dotplot of the amount of fat (to the nearest gram) in 12 different hamburgers served at a fast-food restaurant:



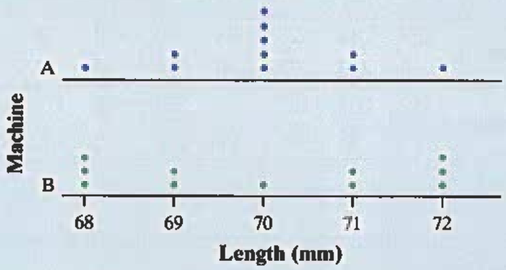
(a) The distribution of fat content has a mean of $\bar{x} = 22.83$ grams and a standard deviation of $s_x = 9.06$ grams. Interpret the standard deviation.

(b) Suppose the restaurant replaces the burger that has 22 grams of fat with a new burger that has 35 grams of fat. How would this affect the mean and the standard deviation? Justify your answer.

103. **Comparing SD** Which of the following distributions has a smaller standard deviation? Justify your answer.



104. Comparing SD The parallel dotplots show the lengths (in millimeters) of a sample of 11 nails produced by each of two machines. Which distribution has the larger standard deviation? Justify your answer.



105. File sizes Refer to Exercise 101. Find the interquartile range of the file size distribution shown in the dotplot.
pg 64

106. Healthy fast food? Refer to Exercise 102. Find the interquartile range of the fat content distribution shown in the dotplot.

107. File sizes Refer to Exercises 101 and 105. Identify any outliers in the distribution. Show your work.
pg 67

108. Healthy fast food? Refer to Exercises 102 and 106. Identify any outliers in the distribution. Show your work.

109. Shopping spree The figure displays computer output for data on the amount spent by 50 grocery shoppers.

	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
Amount spent	34.70	21.70	3.11	19.27	27.86	45.40	93.34

(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.

110. C-sections 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	Med	Q_3	Max
Male doctors	41.333	20.607	20	27	34	50	86
Female doctors	19.1	10.126	5	10	18.5	29	33

(a) Based on the computer output, which distribution would you guess has a more symmetrical shape? Explain your answer.

(b) Explain how the *IQR*s of these two distributions can be so similar even though the standard deviations are quite different.

(c) Does either distribution have any outliers? Justify your answer.

111. Don't call me According to a study by Nielsen Mobile, "Teenagers ages 13 to 17 are by far the most prolific texters, sending 1742 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 they had sent in the past 24 hours. Here are the data:
pg 69

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.

(b) Use the boxplot you created in part (a) to explain how these data seem to contradict the claim in the article.

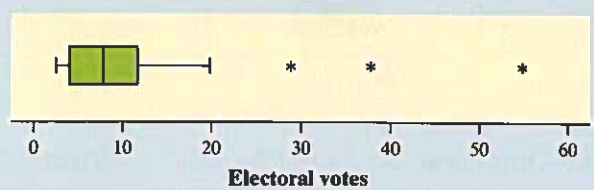
112. Acing the first test Here are the scores of Mrs. Liao's students on their first statistics test:

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 these data.

(b) Use the boxplot you created in part (a) to describe how the students did on Mrs. Liao's first test.

113. Electing the president Refer to Exercise 91. Here are a boxplot and some numerical summaries of the electoral vote data:

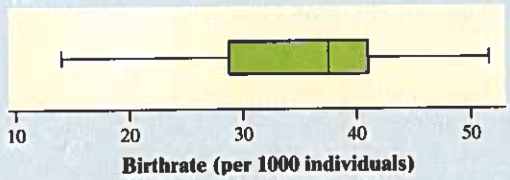


Variable	N	Mean	SD	Min	Q_1	Median	Q_3	Max
Electoral votes	51	10.55	9.69	3	4	8	12	55

(a) Explain why the median and *IQR* would be a better choice for summarizing the center and variability of the distribution of electoral votes than the mean and standard deviation.

(b) Identify an aspect of the distribution that the stemplot in Exercise 91 reveals that the boxplot does not.

114. **Birthrates in Africa** Refer to Exercise 92. Here are a boxplot and some numerical summaries of the birthrate data:

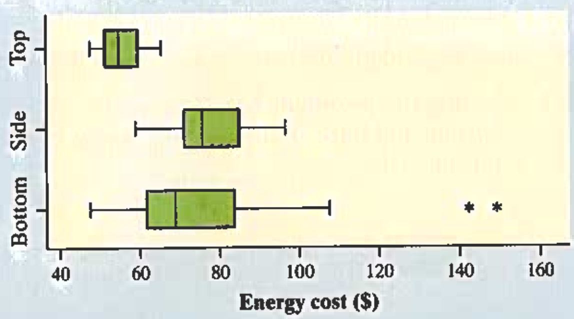


Variable	N	Mean	SD	Min	Q_1	Median	Q_3	Max
Birthrate	54	34.91	8.57	14.00	29.00	37.50	41.00	53.00

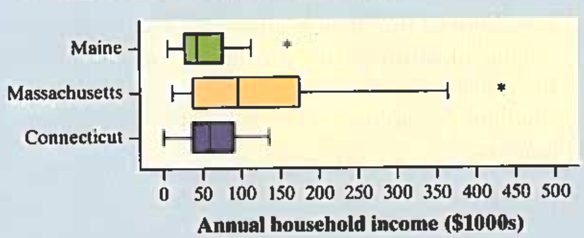
(a) Explain why the median and *IQR* would be a better choice for summarizing the center and variability of the distribution of birthrates in African countries than the mean and standard deviation.

(b) Identify an aspect of the distribution that the dotplot in Exercise 92 reveals that the boxplot does not.

115. **Energetic refrigerators** *Consumer Reports* magazine rated different types of refrigerators, including those with bottom freezers, those with top freezers, and those with side freezers. One of the variables they measured was annual energy cost (in dollars). The following boxplots show the energy cost distributions for each of these types. Compare the energy cost distributions for the three types of refrigerators.



116. **Income in New England** The following boxplots show the total income of 40 randomly chosen households each from Connecticut, Maine, and Massachusetts, based on U.S. Census data from the American Community Survey. Compare the distributions of annual incomes in the three states.



117. **Who texts more?** For their final project, a group of AP® Statistics students wanted to compare the texting habits of males and females. 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	0

(a) Make parallel boxplots of the data.

(b) Use your calculator to compute numerical summaries for both samples.

(c) Do these data suggest that males and females at the school differ in their texting habits? Use the results from parts (a) and (b) to support your answer.

118. **SSHA scores** Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for a random sample of 18 first-year college women:

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

Here are the SSHA scores for a random sample of 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	

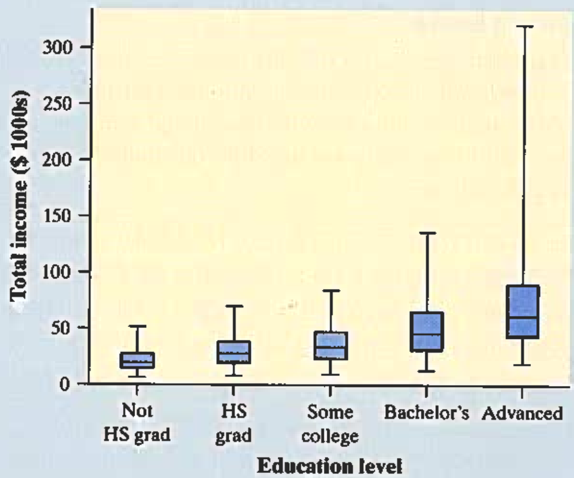
Note that high scores indicate good study habits and attitudes toward learning.

(a) Make parallel boxplots of the data.

(b) Use your calculator to compute numerical summaries for both samples.

(c) Do these data support the belief that men and women differ in their study habits and attitudes toward learning? Use your results from parts (a) and (b) to support your answer.

119. **Income and education level** Each March, the Bureau of Labor Statistics compiles an Annual Demographic Supplement to its monthly Current Population Survey.[43] 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 parallel boxplots 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. Also, the whiskers are drawn all the way to the maximum and minimum values of the remaining data for each distribution.



Use the graph to help answer the following questions.

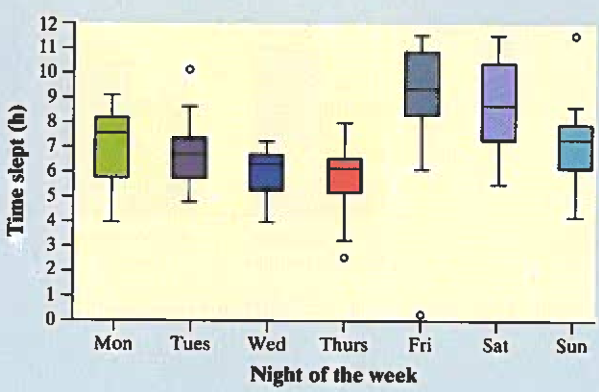
(a) What shape do the distributions of income have?

(b) Explain how you know that there are outliers in the group that earned an advanced degree.

(c) How does the typical income change as the highest education level reached increases? Why does this make sense?

(d) Describe how the variability in income changes as the highest education level reached increases.

120. **Sleepless nights** Researchers recorded data on the amount of sleep reported each night during a week by a random sample of 20 high school students. Here are parallel boxplots comparing the distribution of time slept on all 7 nights of the study:[44]



Use the graph to help answer the following questions.

(a) Which distributions have a clear left-skewed shape?

(b) Which outlier stands out the most, and why?

(c) How does the typical amount of sleep that the students got compare on these seven nights?

(d) On which night was there the most variation in how long the students slept? Justify your answer.

121. **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 part (a) or (b)? Explain.

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

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

Multiple Choice: *Select the best answer for Exercises 123–126.*

123. If a distribution is skewed to the right with no outliers, which expression is correct?

(a) mean < median

(b) mean ≈ median

(c) mean = median

(d) mean > median

(e) We can't tell without examining the data.

124. The scores on a statistics test had a mean of 81 and a standard deviation of 9. One student was absent on the test day, and his score wasn't included in the calculation. If his score of 84 was added to the distribution of scores, what would happen to the mean and standard deviation?

(a) Mean will increase, and standard deviation will increase.

(b) Mean will increase, and standard deviation will decrease.

(c) Mean will increase, and standard deviation will stay the same.

(d) Mean will decrease, and standard deviation will increase.

(e) Mean will decrease, and standard deviation will decrease.

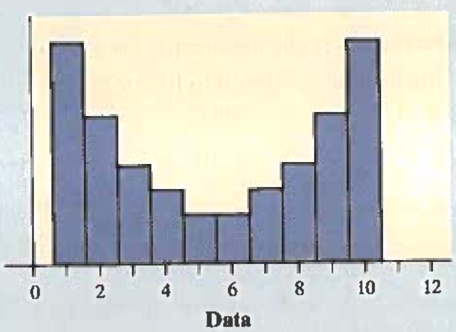
125. The stemplot shows the number of home runs hit by each of the 30 Major League Baseball teams in a single season. Home run totals above what value should be considered outliers?

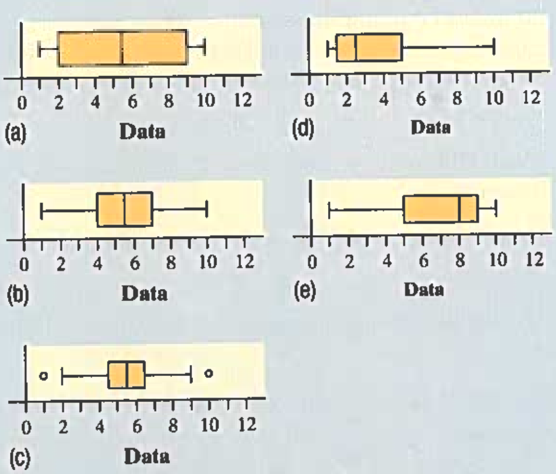
09	15
10	3789
11	47
12	19
13	
14	89
15	34445
16	239
17	223
18	356
19	1
20	3
21	0
22	2

Key: 14|8 is a team with 148 home runs.

(a) 173 (b) 210 (c) 222
(d) 229 (e) 257

126. Which of the following boxplots best matches the distribution shown in the histogram?





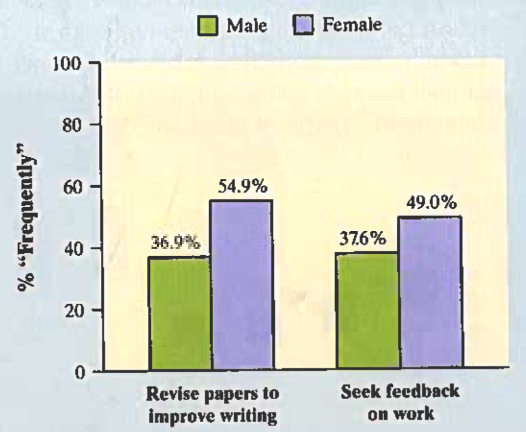
Recycle and Review

127. **How tall are you?** (1.2) We used Census At School's "Random Data Selector" to choose a sample of 50 Canadian students who completed a survey in a recent year. Here are the students' heights (in centimeters):

166.5	170.0	178.0	163.0	150.5	169.0	173.0	169.0	171.0	166.0
190.0	183.0	178.0	161.0	171.0	170.0	191.0	168.5	178.5	173.0
175.0	160.5	166.0	164.0	163.0	174.0	160.0	174.0	182.0	167.0
166.0	170.0	170.0	181.0	171.5	160.0	178.0	157.0	165.0	187.0
168.0	157.5	145.5	156.0	182.0	168.5	177.0	162.5	160.5	185.5

Make an appropriate graph to display these data. Describe the shape, center, and variability of the distribution. Are there any outliers?

128. **Success in college** (1.1) The 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 the percent of males and females who answered "frequently" to these two questions.[45]



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

1.1

(a) The individuals are AP® Statistics students who completed a questionnaire on the first day of class.

(b) The recorded variables were gender (categorical), grade level (categorical), GPA (quantitative), children in family (quantitative), homework last night (min) (categorical), and Type of Phone (categorical).

1.3

The individuals are movies. The variables are year (quantitative), rating (categorical), time (min) (quantitative), genre (categorical), and box office ($) (quantitative). *Note*: Year might be considered categorical if we want to know how many of these movies were made each year, rather than the average year.

1.5

The categorical variables are type of wood, type of water repellent, and paint color. The quantitative variables are paint thickness and weathering time.

1.7

Student answers will vary. Examples of categorical variables could include region of the country and type of institution (2-year college, 4-year college, university). Examples of quantitative variables could include retention rate, graduation rate, class sizes, faculty salaries, student-faculty ratio, percentage of faculty with highest degree in their fields, average ACT/SAT scores, average financial aid, and the percentage of alumni who give to the school.

1.9 b

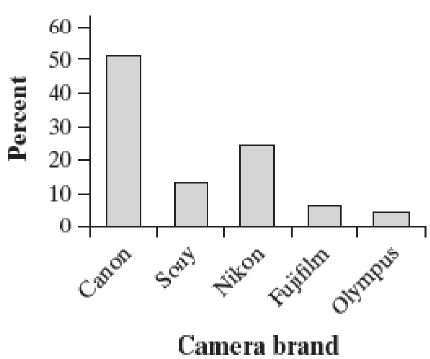
1.10 c

1.13

First, a relative frequency table must be constructed.

Camera brand	Relative Frequency
Canon	23/45 = 0.511 = 51.1%
Sony	6/45 = 0.133 = 13.3%
Nikon	11/45 = 0.244 = 24.4%
Fujifilm	3/45 = 0.067 = 6.7%
Olympus	2/45 = 0.044 = 4.4%

The relative frequency bar graph is given below.

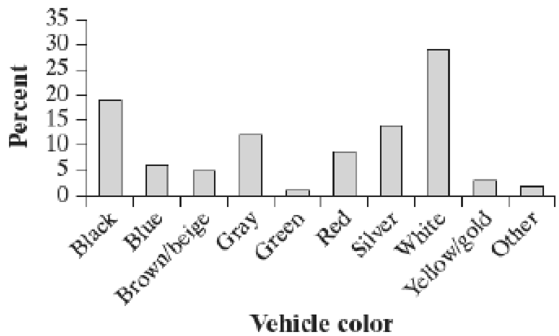


The most popular brand of camera among the 45 most recent purchases on the Internet auction site is Canon, followed by Nikon, Sony, Fujifilm, and Olympus. Canon is the overwhelming favorite with over 50% of the customers purchasing this brand. Also noteworthy is that almost 25% of the customers purchased a Nikon camera.

1.15

(a) The percent of cars with other colors is 100 – 19 – 6 – 5 – 12 – 1 – 9 – 14 – 29 – 3 = 2%.

(b) A bar graph is given below.



The most popular color of vehicles sold that year was white, followed by black, silver, and gray. It appears that a majority of car buyers that year preferred vehicles that were shades of black and white.

(c) It would be appropriate to make a pie chart of these data (including the other category) because the numbers in the table refer to parts of a single whole.

1.17

Estimates will vary, but should be close to 63% Mexican and 9% Puerto Rican.

1.19

The areas of the pictures should be proportional to the numbers of students they represent. As drawn, it appears that most of the students arrived by car but in reality, most came by bus (14 took the bus, 9 came in cars).

1.21

By starting the vertical scale at 12 instead of 0, it looks like the percent of binge-watchers who think that 5 to 6 episodes is too many to watch in one viewing session is almost 20 times higher than the percent of binge-watchers who think that 3 to 4 episodes is too many to watch in one viewing session. In truth, the percent of binge-watchers who think that 5 to 6 episodes is too many to watch in one viewing session (31%) is less than three times higher than the percent of binge-watchers who think that 3 to 4 episodes is too many to watch in one viewing session (13%). Similar arguments can be made for the relative sizes of the other categories represented in the bar graph.

1.23

(a) 50/150 = 0.333. One-third of the 150 subjects were given the control treatment.

(b) 10.7% said they saw broken glass at the accident; 89.3% said they did not; 14% said they saw broken glass at the accident.

(c) Sixteen of the 150 subjects, or 10.67%, were given the “smashed into” treatment and said they saw broken glass at the accident.

1.27

(a) Seven of the 29 subjects who said they saw broken glass at the accident, or a proportion of 0.241, received the "hit" treatment.

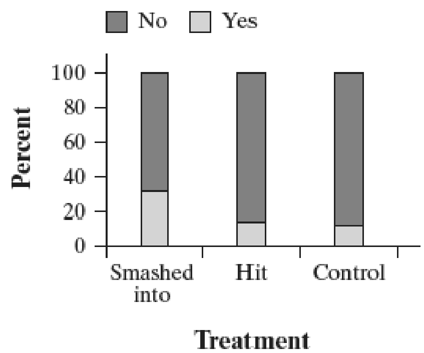
(b) Thirty-four of the 50 subjects who received the "smashed into" treatment, or 68%, said they did not see broken glass at the accident.

1.29

(a) The distributions of responses for the three treatment groups are:

"Smashed Into"	"Hit"	Control
Yes: 16/50 = 32%	Yes: 7/50 = 14%	Yes: 6/50 = 12%
No: 34/50 = 68%	No: 43/50 = 86%	No: 44/50 = 88%

The segmented bar graph is below.



(b) The segmented bar graph reveals that there is an association between opinion about broken glass at the accident and treatment received for subjects in the study. The group that was told that the cars "smashed into" each other was over twice as likely to recall seeing broken glass at the accident, as those who were simply told that the cars "hit" one another or those who were not asked to estimate speed at all were much less likely to recall seeing broken glass at the accident. Knowing which treatment a subject received helps us predict whether or not they will respond that they saw broken glass at the accident.

1.33

Answers may vary. Regardless of whether a student went to a private or public college, most students chose a school that was at least 11 miles from home. Those that went to a public university were most likely to choose a school that was 11 to 50 miles from home (about 30%), while those who went to a private university were most likely to choose a school that was 101 to 500 miles from home (about 29%).

1.35

(a) The graph reveals that as age increases, the percent that use smartphones for navigation decreases.

(b) It would not be appropriate to make a pie chart of this data because the category percentages are not parts of the same whole.

1.40 b

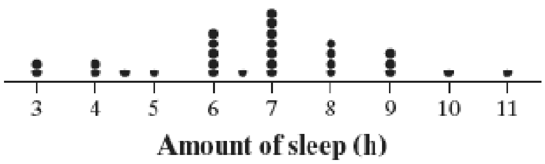
1.41 d

1.42 d

1.43 c

1.45

(a) The graph is shown below.



(b) Five out of 28 students, or a proportion of 0.179, got the recommended amount of sleep.

1.49

The shape of the distribution is left skewed with a peak between 90 and 100 years. There is a small gap around 70 years.

1.51

The shape of the distribution is roughly symmetric with a single peak at 7.

1.55

The distribution of total family income for Indiana is roughly symmetric, while the distribution of total family income for New Jersey is slightly skewed right. The value of $125,000 may be an outlier in the Indiana distribution. There are no obvious outliers in the New Jersey distribution. The median for both distributions is about the same, approximately $49,000. The distribution of total family income in Indiana is less variable than the New Jersey distribution. The incomes in Indiana vary from $0 to about $125,000. The incomes in New Jersey vary from $0 to about $170,000.

1.59

(a) The stemplot is shown below.

```
15 | 9
16 | 0 5 5 6 7 8
17 | 1 1 1 3 4 4 7 7 8
18 |
19 | 2
```

Key: 15 | 9 = 15.9 grams

(b) The graph reveals that there was one Fun Size Snickers® bar that is "gigantic"! It weighs 19.2 grams.

(c) Seven of the 17 candy bars in this sample, or a proportion of 0.412, weigh less than advertised.

1.63

(a) If we had not split the stems, most of the data would appear on just a few stems, making it hard to identify the shape of the distribution.

(b) Key: 16 | 0 means that 16.0% of that state's residents are aged 25 to 34.

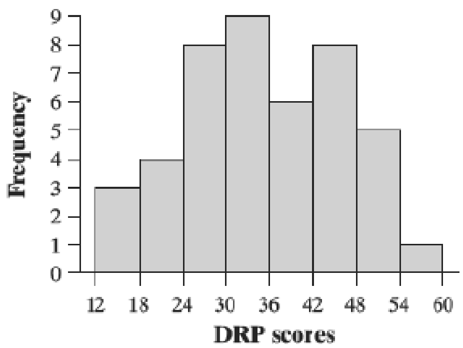
(c) The distribution of percent of residents aged 25-34 is roughly symmetric with a possible outlier at 16.0%.

1.65

The distribution of acorn volume for the Atlantic coast is skewed to the right. The distribution of acorn volume for California is roughly symmetric with one high outlier of 17.1 cubic centimeters. The distribution of volume of acorn for the Atlantic coast has 3 potential outliers: 8.1, 9.1, and 10.5 cubic centimeters. The typical acorn volume for Atlantic coast oak tree species (median = 1.7 cubic centimeters) is less than the typical acorn volume for California oak tree species (median = 4.1 cubic centimeters). The Atlantic coast distribution (with acorn volumes from 0.3 to 10.5 cubic centimeters) varies less than the California distribution (with acorn volumes from 0.4 to 17.1 cubic centimeters.)

1.69

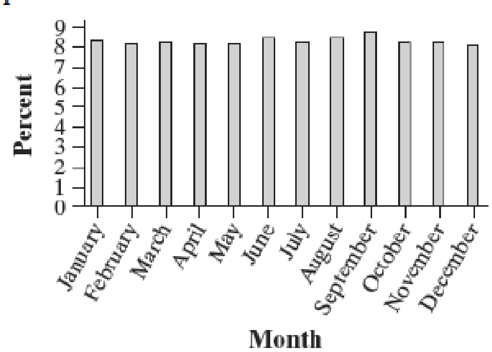
The data vary from 14 to 54. We chose intervals of width 6, beginning at 12.



The distribution of DRP scores is roughly symmetric. There do not appear to be any outliers. The center of the DRP score distribution is between 30 and 36 (with median = 35). The DRP scores vary from 14 to 54.

1.77

A bar graph should be used because birth month is a categorical variable. A possible bar graph is given below.



1.80 b	1.83 e
1.81 a	1.84 d
1.82 c	1.85 b

1.87

(a) The mean of Joey's first 14 quiz scores is $\frac{86+84+...+93}{14} = \frac{1190}{14} = 85$.

(b) Including a 15[th] quiz score of 0, Joey's mean would be $\frac{86+84+...+93+0}{15} = \frac{1190}{15} = 79.3$. This illustrates the property of nonresistance. The mean is not resistant. It is sensitive to extreme values.

1.89

(a) First we must put the scores in order: 74, 75, 76, 78, 80, 82, 84, 86, 87, 90, 91, 93, 96, and 98. Because there are 14 scores, the median is the average of the 7th and 8th scores. Therefore, the median is $\frac{84+86}{2} = 85$.

(b) If Joey had a 0 for the 15th quiz, we add the 0 to the beginning of the list in part (a). Since there are now 15 quiz scores, the median would be the 8th score, which is 84. Notice that the median did not change much. This shows that the median is resistant to outliers.

1.91

(a) Since there are 51 observations, the median is the 26th observation, or 8 electoral votes.

(b) Since the distribution of number of electoral votes is skewed to the right, the mean of this distribution is greater than its median.

1.95

(a) Estimate the frequencies of the bars (from left to right): 15, 11, 15, 11, 8, 5, 3, 3, and 3.

Although the answers may vary slightly, the frequencies must sum to 74. We estimate the median by finding the average of the 37th and 38th values, both of which are 2. The median is 2 servings of fruit per day.

(b) Using these values, we can estimate the mean by adding 0 fifteen times, 1 eleven times, and so on. This gives us a sum of 194. The mean is then calculated by dividing by the number of responses:

$\bar{x} = \frac{194}{74} = 2.62$ servings of fruit per day. Alternatively, we can estimate the mean by finding the balance point of the distribution, which appears to be slightly larger than 2.5.

1.97

(a) The range of Joey's first 14 quiz grades is Range = Max − Min = 98 − 74 = 24.

The range of Joey's quiz grades after his unexcused absence is 98 − 0 = 98.

(b) The range may not be the best way to describe variability for a distribution of quantitative data because the range can be heavily affected by outliers.

1.101

(a) The size of these 18 files typically varies by about 1.9 megabytes from the mean of 3.2 megabytes.

(b) If the music file that takes up 7.5 megabytes of storage space is replaced with another version of the file that only takes up 4 megabytes the mean would decrease slightly. The standard deviation would decrease as well because a file of size 4 megabytes will be closer to the new mean than the file of 7.5 megabytes was to the former mean.

1.103

Variable B has a smaller standard deviation because more of the observations have values closer to the mean than in Variable A's distribution. That is, the typical distance from the mean is smaller for Variable B than for Variable A.

1.105

The file sizes, in order, are 1.1, 1.3, 1.3, 1.6, <u>1.9</u>, 1.9, 2.1, 2.2, 2.4, 2.5, 2.7, 3.0, 4.4, <u>4.7</u>, 5.0, 5.6, 6.2, and 7.5. The first quartile of the distribution of file size is the 5th observation in the ordered list, so $Q_1 = 1.9$ megabytes. The third quartile of the distribution of file size is the 14th observation in the ordered list, so $Q_3 = 4.7$. The interquartile range of the file size distribution shown in the dotplot is $IQR = Q_3 - Q_1 = 4.7 - 1.9 = 2.8$ megabytes.

1.109

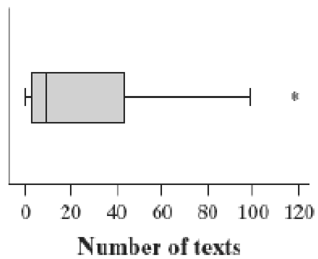
(a) The distribution is skewed to the right because the mean is much larger than the median. Also, Q_3 is much further from the median than Q_1.

(b) The amount of money spent typically varies by about \$21.70 from the mean of \$34.70.

(c) The first quartile is 19.27 and the third quartile is 45.40 so the *IQR* is 45.40 – 19.27 = 26.13. Any points below 19.27 – 1.5(26.13) = –19.925 or above 45.40 + 1.5(26.13) = 84.595 are outliers. Because the maximum of 93.34 is greater than 84.595, there is at least one outlier.

1.111

(a) Putting the data in order: 0, 0, 0, 1, 1, 3, 3, 5, 5, 7, 8, 8, 9, 14, 25, 25, 26, 29, 42, 44, 52, 72, 92, 98, and 118. The median is 9, the first quartile is 3, and the third quartile is 43. The *IQR* is 43 – 3 = 40. An outlier would be any value below $3-1.5(40)=-57$ or above $43+1.5(40)=103$. This means that the value of 118 is an outlier. The boxplot is shown below.



(b) The article claims that teens send 1742 texts a month, which works out to be about 58 texts a day (assuming a 30 day month). Nearly all of the members of the class (21 of 25) sent fewer than 58 texts per day, which seems to contradict the claim in the article.

1.113

(a) the median and *IQR* would be a better choice for summarizing the center and variability of the distribution of electoral votes than the mean and standard deviation because the boxplot reveals that there are three outliers in the data set. The mean and standard deviation are not resistant measures of center and variability, so their values are sensitive to these extreme values.

(b) The stemplot reveals that the distribution has a single peak, which cannot be discerned from the boxplot. Also, the stemplot reveals that there are actually *four* upper outliers rather than three. The value of 29, which is an outlier, gives the number of electoral votes for *two* states. In the boxplot, this appears as one asterisk, however, there are two states that have that many electoral votes, not one.

1.115

Shape: The distribution of energy cost (in dollars) for top freezers looks roughly symmetric.
The distribution of energy cost (in dollars) for side freezers looks roughly symmetric.
The distribution of energy cost (in dollars) for bottom freezers looks skewed to the right.
Outliers: There are no outliers for the top or side freezers. There are at least two bottom freezers with unusually high energy costs (over \$140 per year).
Center: The typical energy cost for the side freezers (median ≈ \$75) is greater than the typical cost for the bottom freezers (median ≈ \$69), which is greater than the typical cost for the top freezers (median ≈ \$56).
Variability: There is much more variability in the energy costs for bottom freezers (*IQR* ≈ \$20), than for side freezers (*IQR* ≈ \$12), than for top freezers (*IQR* ≈ \$8).

1.121

(a) One possible answer is 1, 1, 1, and 1.
(b) 0, 0, 10, 10.
(c) For part (a), any set of four identical numbers will have $s_x = 0$. For part (b), however, there is only 1 possible answer. We want the values to be as far from the mean as possible, so the squared deviations from the mean can be as big as possible. If we choose 0, 10, 10, 10—or 10, 0, 0, 0—we make the first squared deviation 7.5^2, but the other three are only 2.5^2. Our best choice is two values at each extreme, which makes all four squared deviations equal to 5^2.

1.123 d
1.124 b
1.125 e
1.126 a