INTRODUCTION Statistics: The Science and Art of Data

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

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

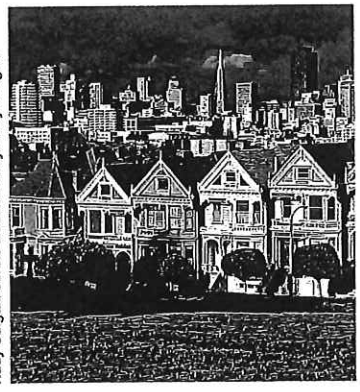
We live in a world of *data*. Every day, the media report poll results, outcomes of medical studies, and analyses of data on everything from stock prices to standardized test scores to global warming. The data are trying to tell us a story. To understand what the data are saying, you need to learn more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

DEFINITION Statistics

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

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

Organizing Data



Rudy Sulgan/Corbis Documentary/Getty Images

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

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

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

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

DEFINITION Individual, Variable

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

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

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

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

DEFINITION Categorical variable, Quantitative variable

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

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

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

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

EXAMPLE Census At School Individuals and Variables

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

Garry Black/Alamy

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

(a) Identify the individuals in this data set.

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

SOLUTION:

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

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

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

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

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

FOR PRACTICE, TRY EXERCISE 1

AP® EXAM TIP

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

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

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

DEFINITION Distribution

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

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

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

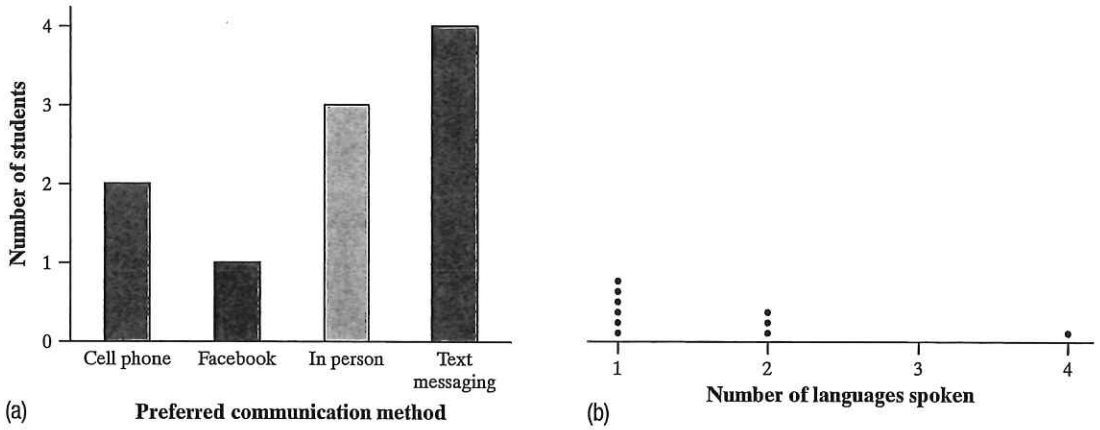


FIGURE 1.1 (a) Bar graph showing the distribution of preferred communication method for the sample of 10 Canadian students. (b) Dotplot showing the distribution of number of languages spoken by these students.

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

HOW TO ANALYZE DATA

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

CHECK YOUR UNDERSTANDING

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

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

From Data Analysis to Inference

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

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

ACTIVITY Hiring discrimination—it just won't fly!

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

Choja/Getty Images

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

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

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

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

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

Introduction Summary

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

Introduction Exercises

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

1. **A class survey** Here is a small part of the data set that pg 3 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.

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%

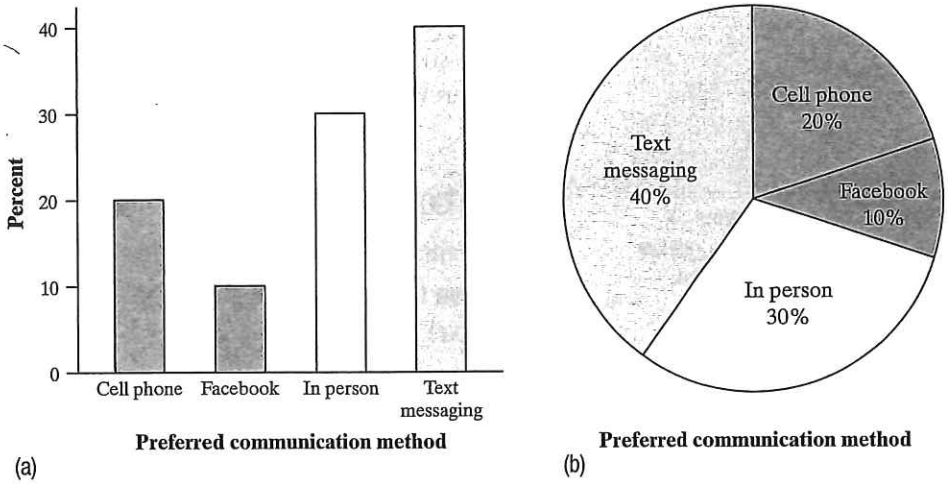


FIGURE 1.2 (a) Bar graph and (b) pie chart of the distribution of preferred communication method for a random sample of 10 Canadian students.

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]

(a) Identify the individuals in this data set.

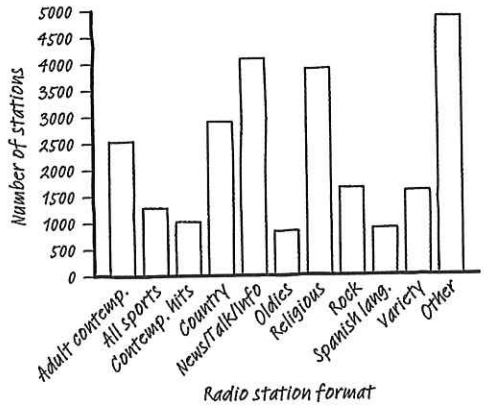
(b) Make a frequency bar graph of the data. Describe what you see.

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**

SOLUTION:

(a) U.S. radio stations

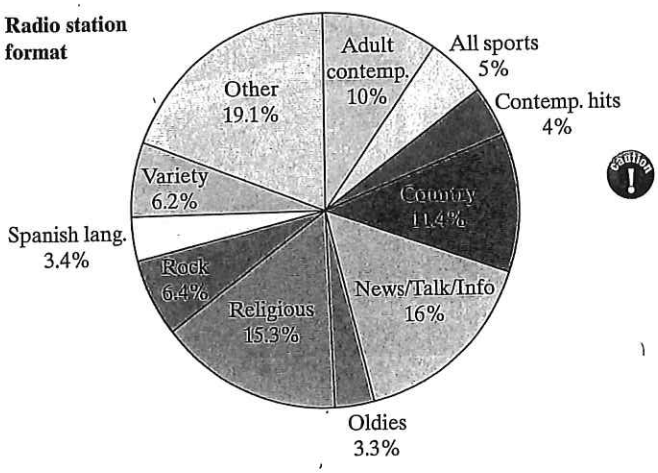
(b)



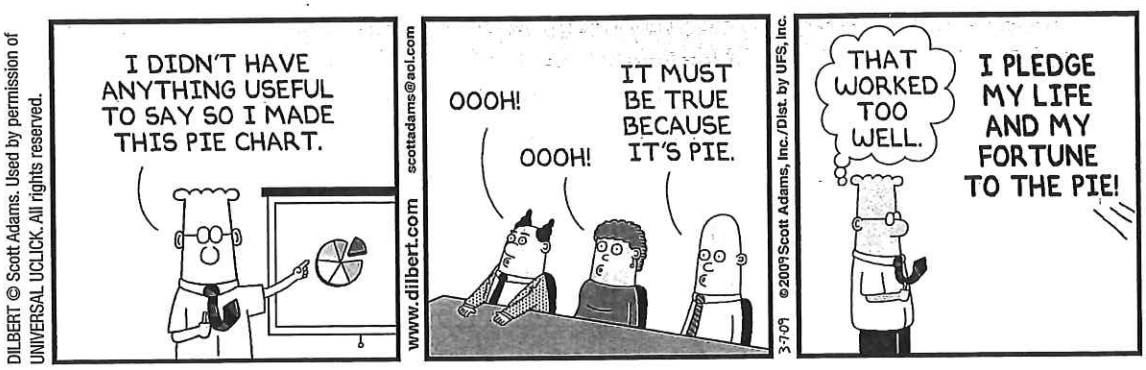
To make the bar graph:
- **Draw and label the axes.**
- **"Scale" the axes.** The largest frequency is 4852. So we choose a vertical scale from 0 to 5000, with tick marks 500 units apart.
- **Draw bars above the category names.**

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

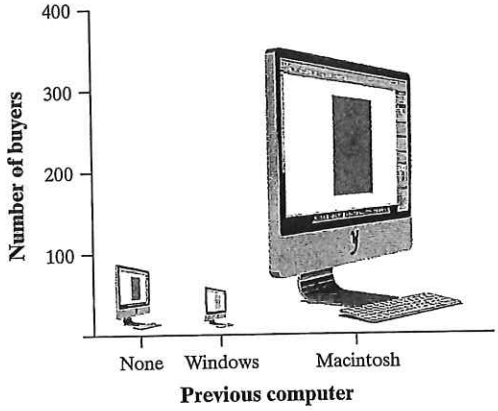
PROBLEM: When Apple, Inc., introduced the iMac, the company wanted to know whether this new computer was expanding Apple's market share. Was the iMac mainly being bought by previous Macintosh owners, or was it being purchased by first-time computer buyers and by previous PC users who were switching over? To find out, Apple hired a firm to conduct a survey of 500

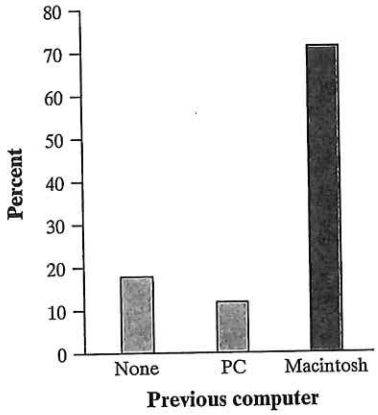
iMac customers. Each customer was categorized as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey results.[4]

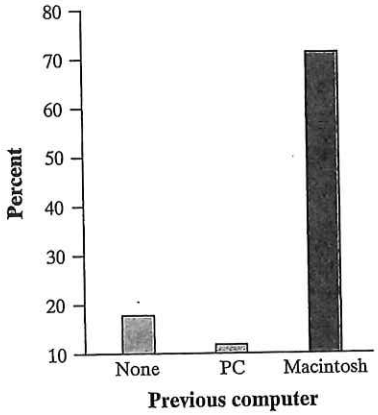
Previous ownership	Count	Percent (%)
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	500	100.0

(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


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

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.

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

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

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

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?

SOLUTION:

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

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.

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.

		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

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%

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

This is the 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).

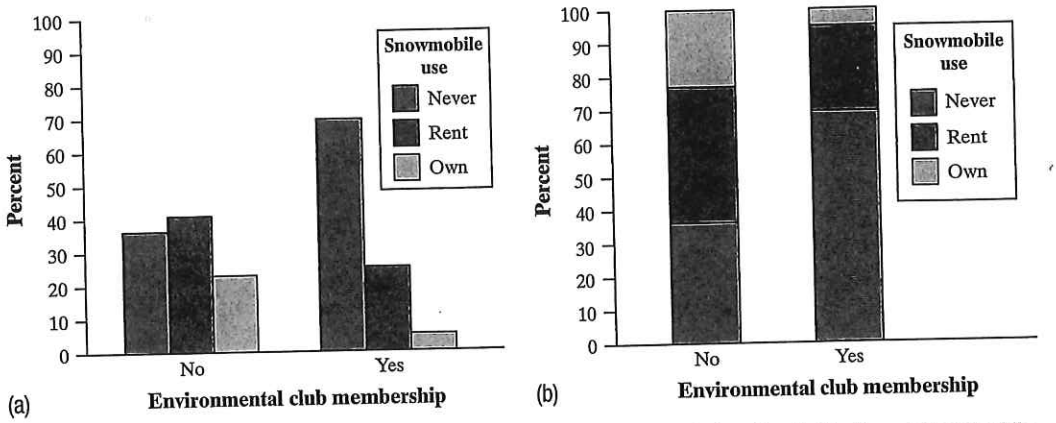


FIGURE 1.3 (a) Side-by-side bar graph and (b) segmented bar graph displaying the distribution of snowmobile use among environmental club members and among non-environmental club members from the 1526 randomly selected winter visitors to Yellowstone National Park.

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.

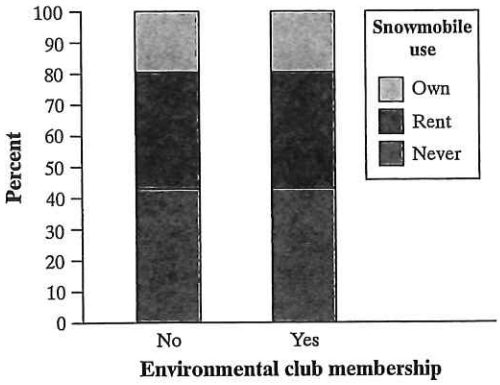
A **segmented bar graph** displays the distribution of a categorical variable as segments of a rectangle, with the area of each segment proportional to the percent of individuals in the corresponding category.

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.

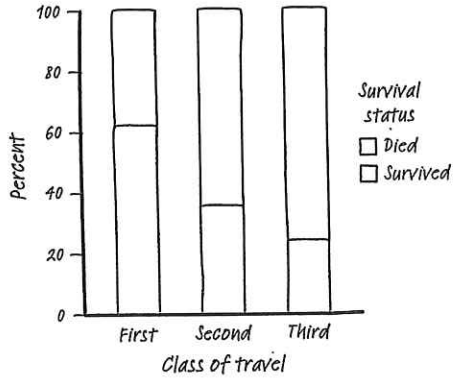
(b) Describe what the graph in part (a) reveals about the association between class of travel and survival status for adult passengers on the *Titanic*.

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: $\frac{476}{627} = 0.759 = 75.9\%$



(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%).

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.

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*.

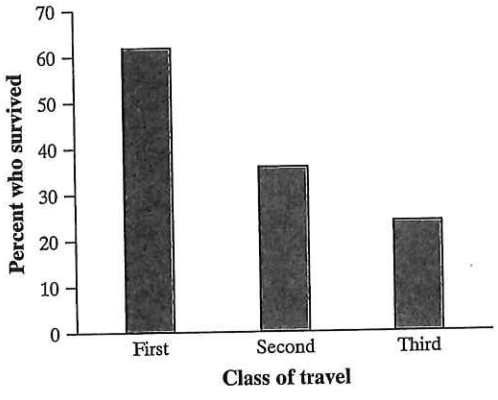


FIGURE 1.4 Bar graph comparing the percents of passengers who survived among each of the three classes of travel on the *Titanic*.

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

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

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

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

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

Recycle and Review

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

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

SECTION 1.2 Displaying Quantitative Data with Graphs

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

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

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

Dotplots

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

DEFINITION Dotplot

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

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

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

Figure 1.5 shows a dotplot of these data.

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

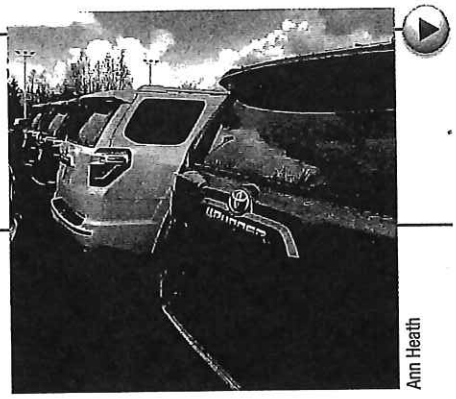
It is fairly easy to make a dotplot by hand for small sets of quantitative data.

HOW TO MAKE A DOTPLOT

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

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

EXAMPLE Give it some gas! Making and interpreting dotplots


Ann Heath

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

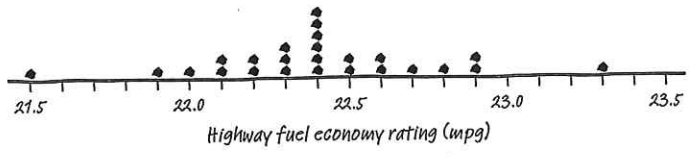
22.4 22.4 22.3 23.3 22.3 22.3 22.5 22.4 22.1 21.5 22.0 22.2 22.7
22.8 22.4 22.6 22.9 22.5 22.1 22.4 22.2 22.9 22.6 21.9 22.4

(a) Make a dotplot of these data.

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

SOLUTION:

(a)



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

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

FOR PRACTICE, TRY EXERCISE 45

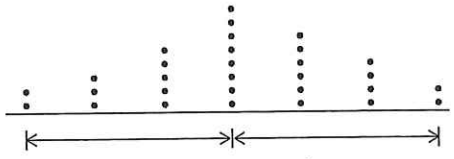
Describing Shape

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

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

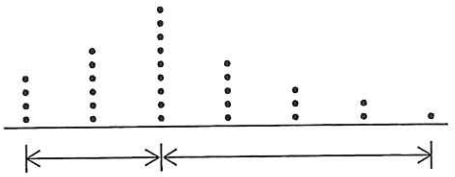
DEFINITION Symmetric and skewed distributions

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



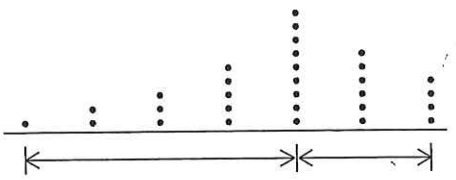
Roughly symmetric

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



Skewed to the right

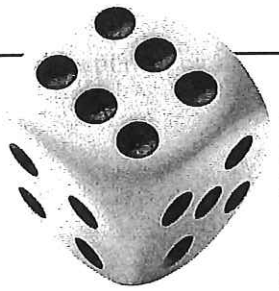
A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.



Skewed to the left

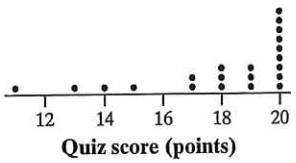
For ease, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed to the right." **The direction of skewness is toward the long tail, not the direction where most observations are clustered.** The drawing is a cute but corny way to help you keep this straight.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

EXAMPLE Quiz scores and die rolls Describing shape

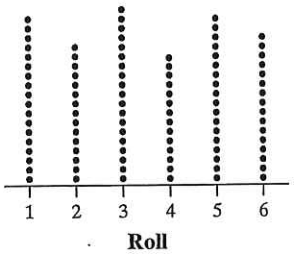

Malerapaso/Getty Images

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

(a)



(b)



SOLUTION:

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

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

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

FOR PRACTICE, TRY EXERCISE 49

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

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

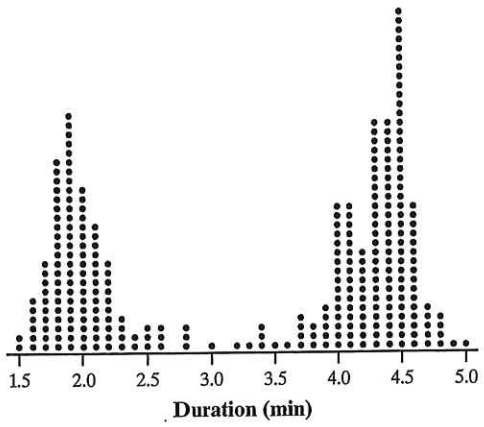


FIGURE 1.6 Dotplot displaying duration (in minutes) of 220 Old Faithful eruptions. This graph has two distinct clusters and two clear peaks.

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

CHECK YOUR UNDERSTANDING

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

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

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

Describing Distributions

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

HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

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

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

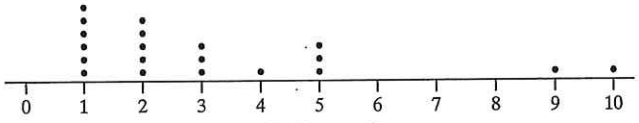
Variability is sometimes referred to as *spread.* We prefer variability because students sometimes think that spread refers only to the distance between the maximum and minimum value of a quantitative data set (the *range*). There are several ways to measure the variability (spread) of a distribution,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

AP® EXAM TIP

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

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

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



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

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

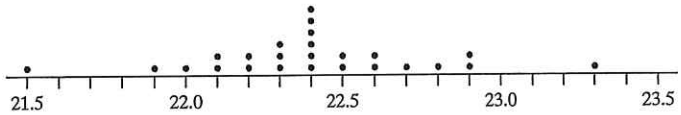
Outliers: The games when the team scored 9 and 10 goals appear to be outliers.

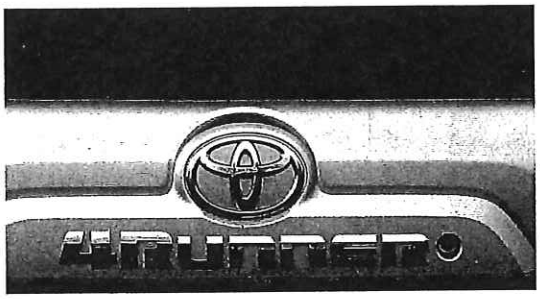
Center: The median is 2 goals scored.

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

EXAMPLE Give it some gas! Describing a distribution

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





Daren Starnes

SOLUTION:

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

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

Center: The median rating is 22.4 mpg.

Variability: The ratings vary from 21.5 to 23.3 mpg.

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

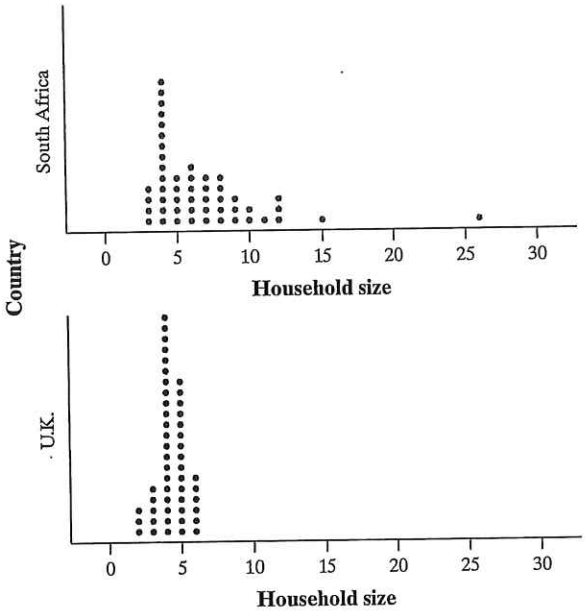
FOR PRACTICE, TRY EXERCISE 53

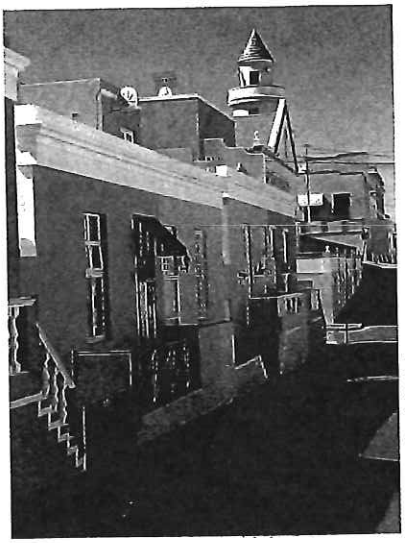
Comparing Distributions

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.

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.

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).

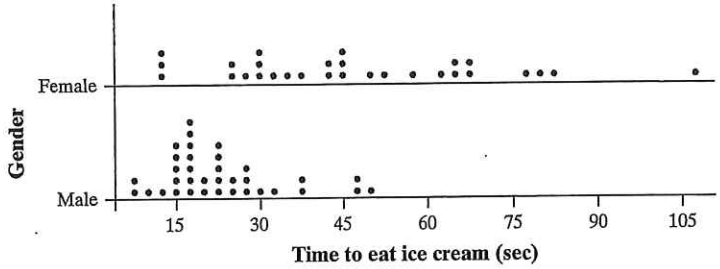
Variability: The household sizes for the South African students vary more (from 3 to 26 people) than for the U.K. students (from 2 to 6 people).

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.

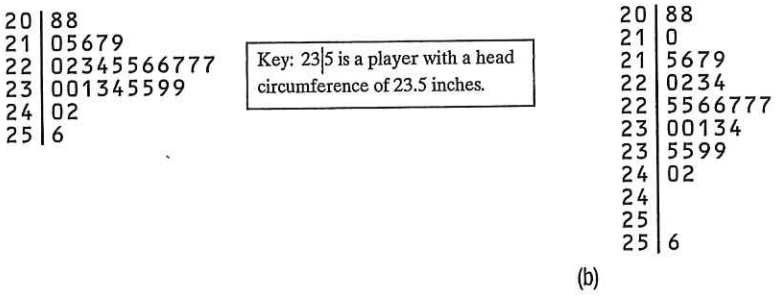


FIGURE 1.8 Two stemplots showing the head circumference data. The graph in (b) improves on the graph in (a) by splitting stems.

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.

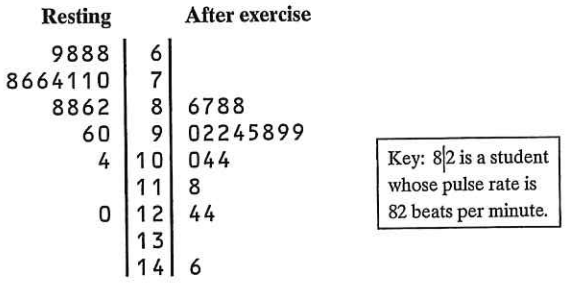


FIGURE 1.9 Back-to-back stemplot of 19 middle school students' resting pulse rates and their pulse rates after 5 minutes of running.

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	012223444457888999
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.

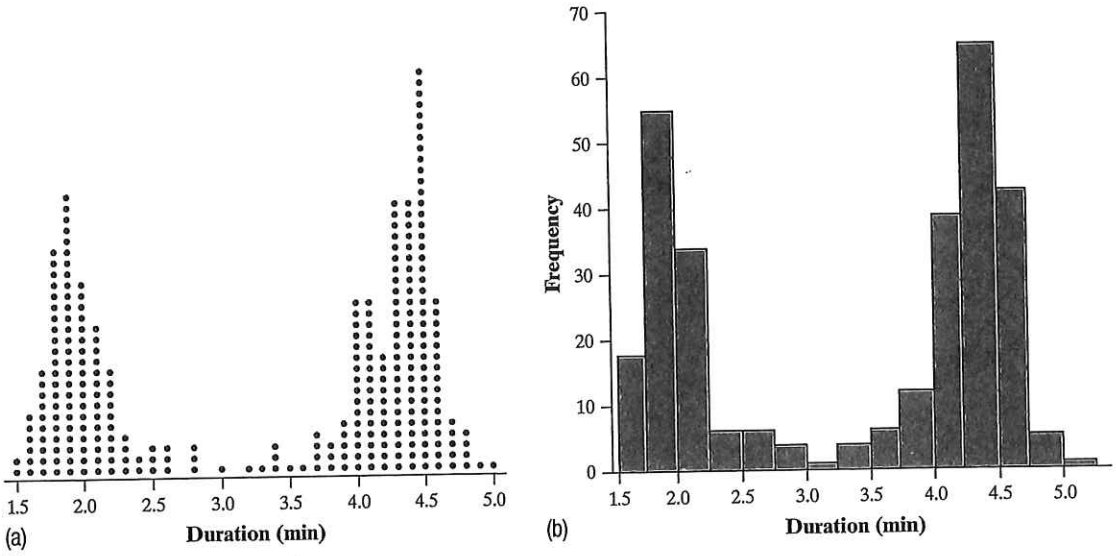


FIGURE 1.10 (a) Dotplot and (b) histogram of the duration (in minutes) of 220 eruptions of the Old Faithful geyser.

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]

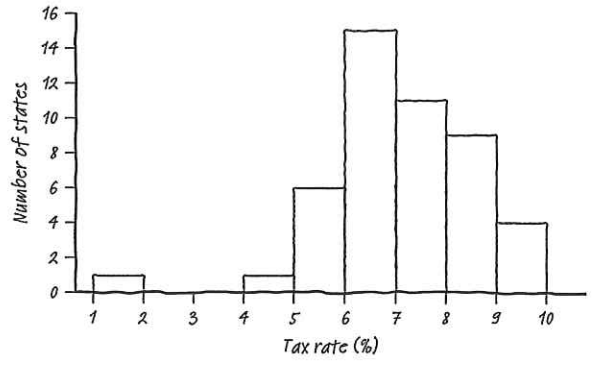
State	Tax rate (%)	State	Tax rate (%)	State	Tax rate (%)
Alabama	9.0	Louisiana	9.0	Oklahoma	8.8
Alaska	1.8	Maine	5.5	Pennsylvania	6.3
Arizona	8.3	Maryland	6.0	Rhode Island	7.0
Arkansas	9.3	Massachusetts	6.3	South Carolina	7.2
California	8.5	Michigan	6.0	South Dakota	5.8
Colorado	7.5	Minnesota	7.3	Tennessee	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.

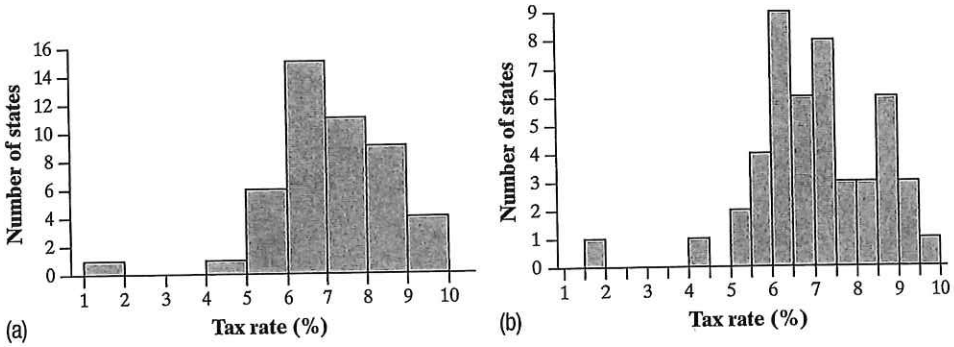


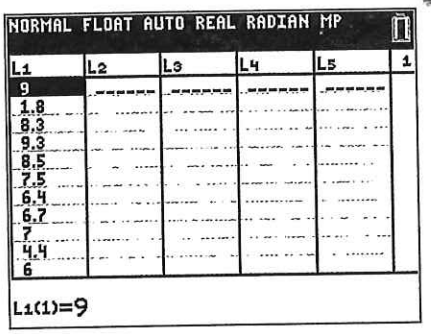
FIGURE 1.11 (a) Frequency histogram of the sales tax rate in the states that have local or state sales taxes and the District of Columbia with intervals of width 1.0%, from the preceding example. (b) Frequency histogram of the data with intervals of width 0.5%.

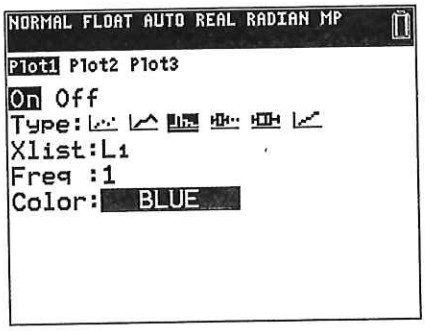
You can use a graphing calculator, statistical software, or an applet to make a histogram. The technology's default choice of intervals is a good starting point, but you should adjust the intervals to fit with common sense.

2. Technology Corner MAKING HISTOGRAMS

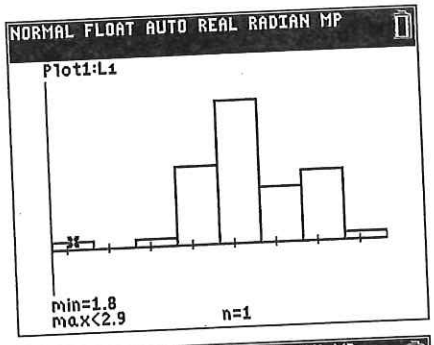
TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

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.

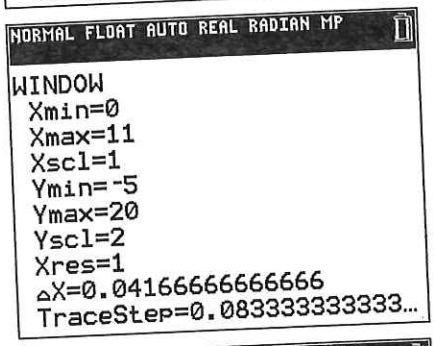




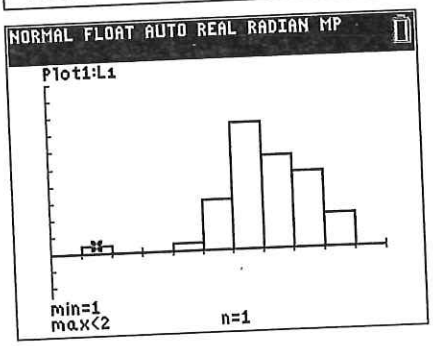
3. Use ZoomStat to let the calculator choose intervals and make a histogram.
   - Press [ZOOM] and choose ZoomStat.
   - Press [TRACE] to examine the intervals.



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).

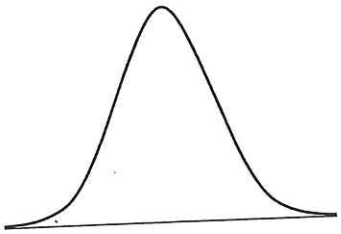


AP® EXAM TIP

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]



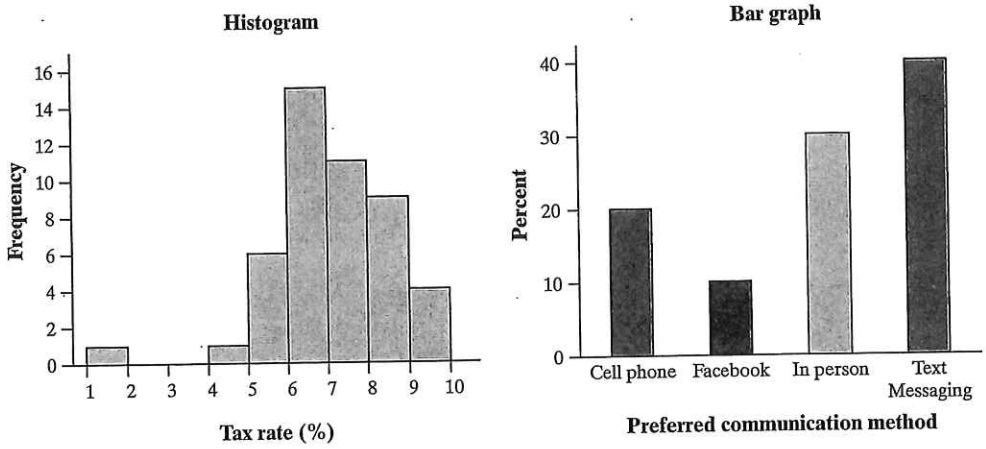
145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

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.

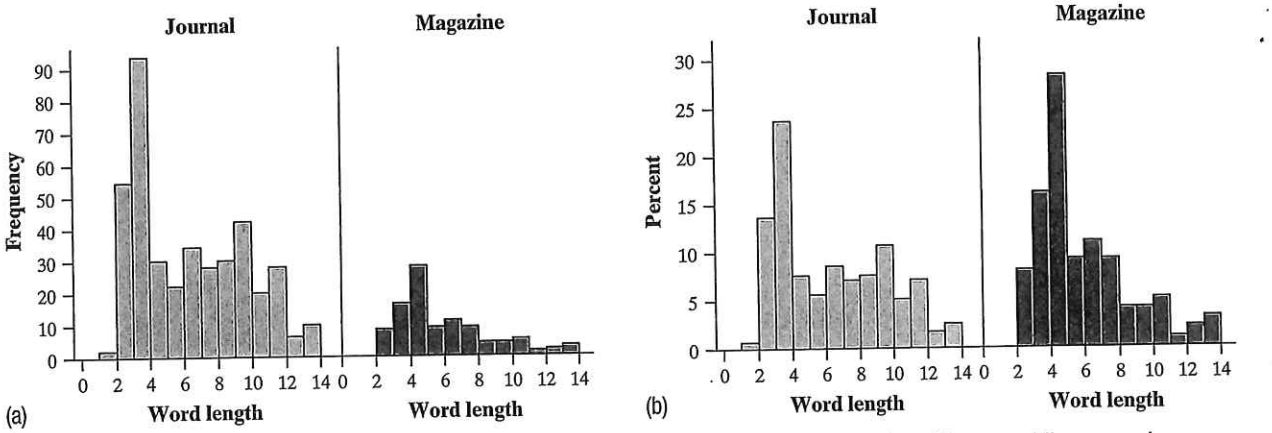
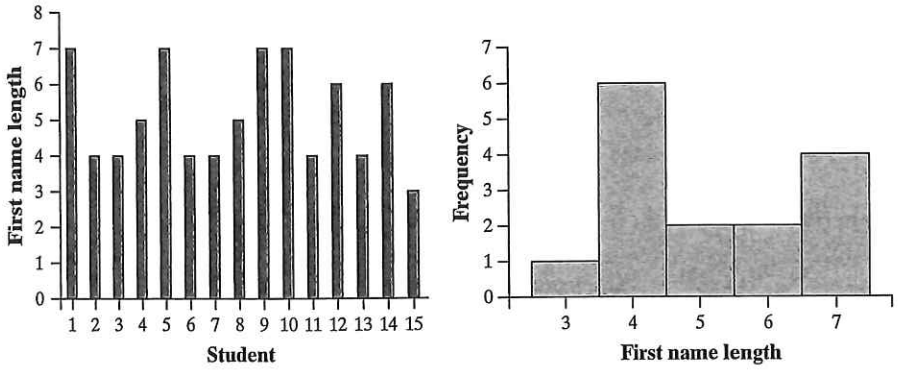


FIGURE 1.12 Two sets of histograms comparing word lengths in articles from a biology journal and from an airline magazine. In graph (a), the vertical scale uses frequencies. Graph (b) fixes the problem of different sample sizes by using percents (relative frequencies) on the vertical scale.

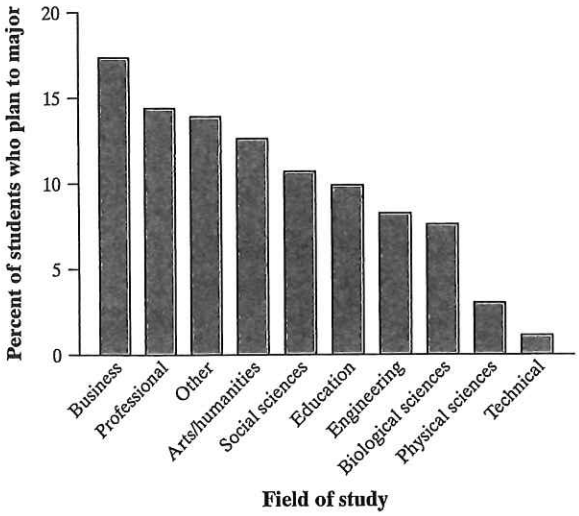
3. **Just because a graph looks nice doesn't make it a meaningful display of data.** The 15 students in a small statistics class recorded the number of letters in their first names. One student entered the data into an Excel spreadsheet and then used Excel's "chart maker" to produce the graph shown on 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?

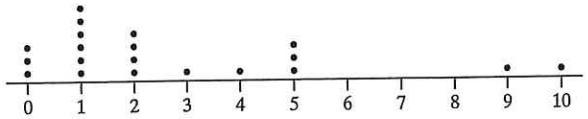
46. **Easy reading?** Here are data on the lengths of the first 25 words on a randomly selected page from Toni Morrison's *Song of Solomon*:

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.



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

(a) Explain what the dot above 3 represents.

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

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?

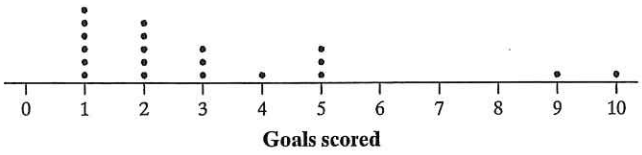
(c) Researchers suspected that the more serious soccer players were more likely to develop arthritis later in life. Do the data confirm this suspicion? Calculate appropriate percentages to support your answer.

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

$= \dfrac{63}{20} = 3.15$ goals

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

(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.

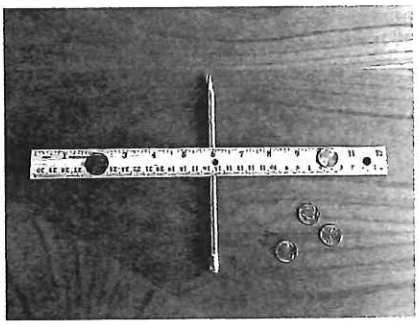
DEFINITION Resistant

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"

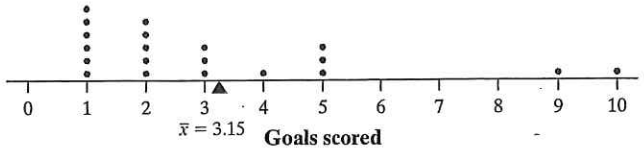
In this activity, you'll investigate an important property of the mean.



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.

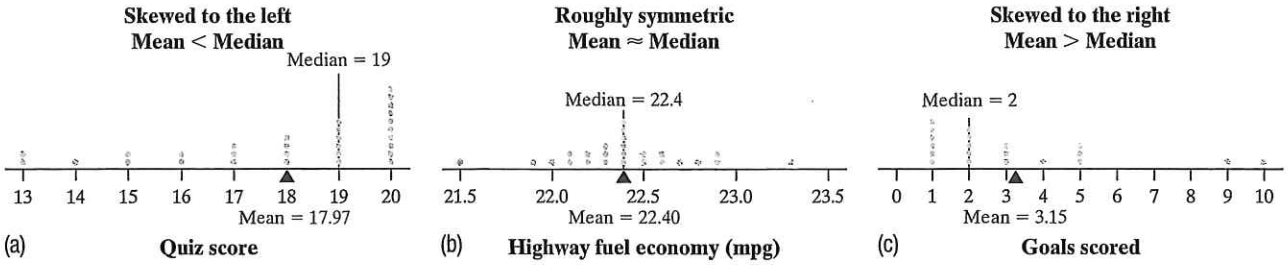


FIGURE 1.13 Dotplots that show the relationship between the mean and median in distributions with different shapes: (a) Scores of 30 statistics students on a 20-point quiz, (b) highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners, and (c) number of goals scored in 20 games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

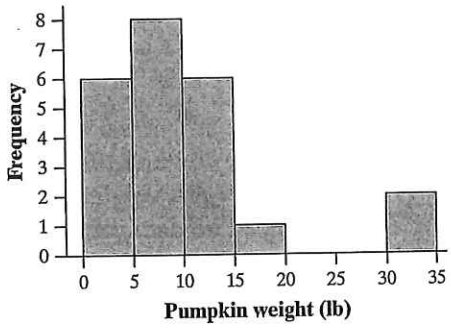
The mean and median measure center in different ways, and both are useful. In Major League Baseball (MLB), the distribution of player salaries is strongly skewed to the right. Most players earn close to the minimum salary (which was $507,500 in 2016), while a few earn more than $20 million. The median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

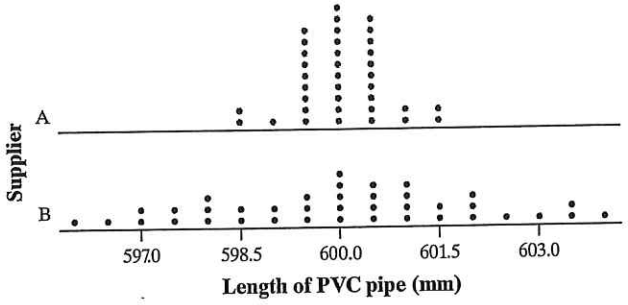


FIGURE 1.14 Comparative dotplots of the length of PVC pipes in separate random samples from Supplier A and Supplier B.

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

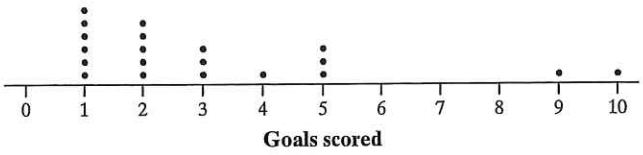
DEFINITION Range

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:

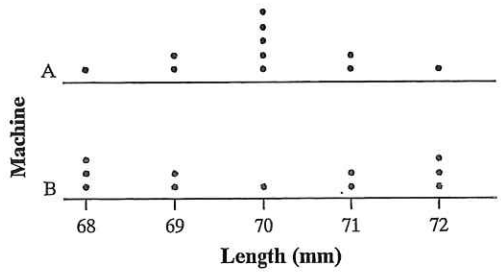
5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3



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

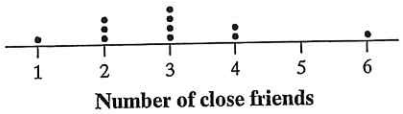
$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}.$$

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

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

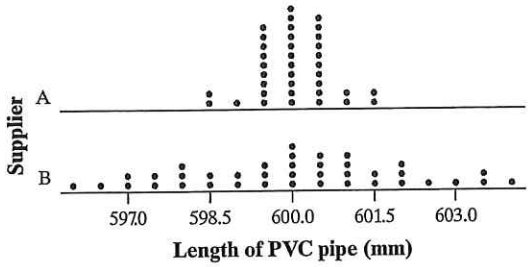
Unfortunately, the units are "squared close friends." Because variance is measured in squared units, it is not a very helpful way to describe the variability of a distribution.

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.

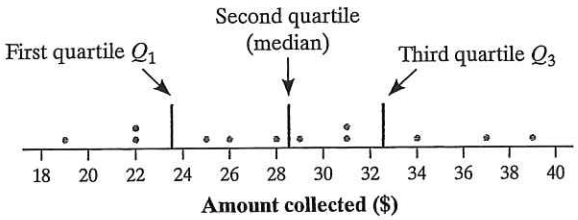
In the close friends example, 11 high school students had an average of $\bar{x} = 3$ close friends with a standard deviation of $s_x = 1.34$. What if a 12th high school student was added to the sample who had 3 close friends? The mean number of close friends in the sample would still be $\bar{x} = 3$. 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:

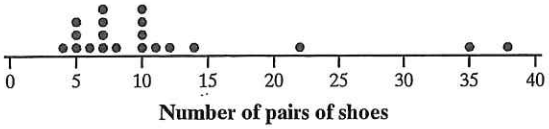
$$IQR = Q_3 - Q_1$$

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:

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



Peter Cade/Getty Images

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

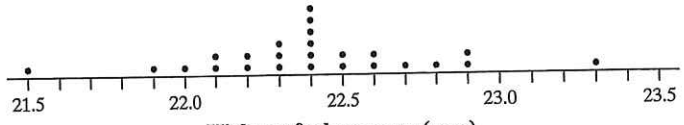
Q_1 Median Q_3

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



Highway fuel economy (mpg)

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.

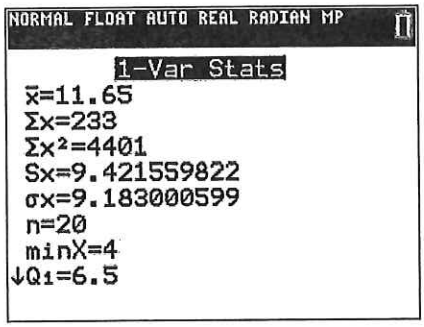
Numerical Summaries with Technology

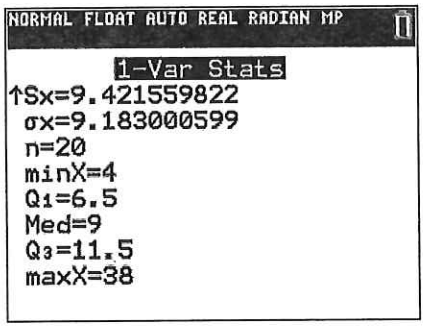
Graphing calculators and computer software will calculate numerical summaries for you. Using technology to perform calculations will allow you to focus on choosing the right methods and interpreting your results.

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.

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.

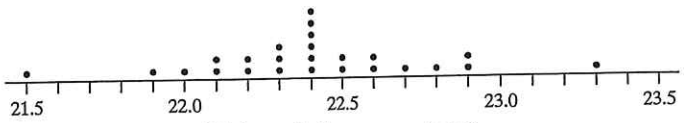
HOW TO IDENTIFY OUTLIERS: THE 1.5 × *IQR* RULE

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

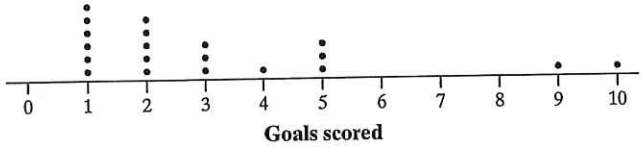
You may be asked to determine whether a quantitative data set has any outliers. Be prepared to state and use the rule for identifying outliers.

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 $Q_3 = 4.5$

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

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

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

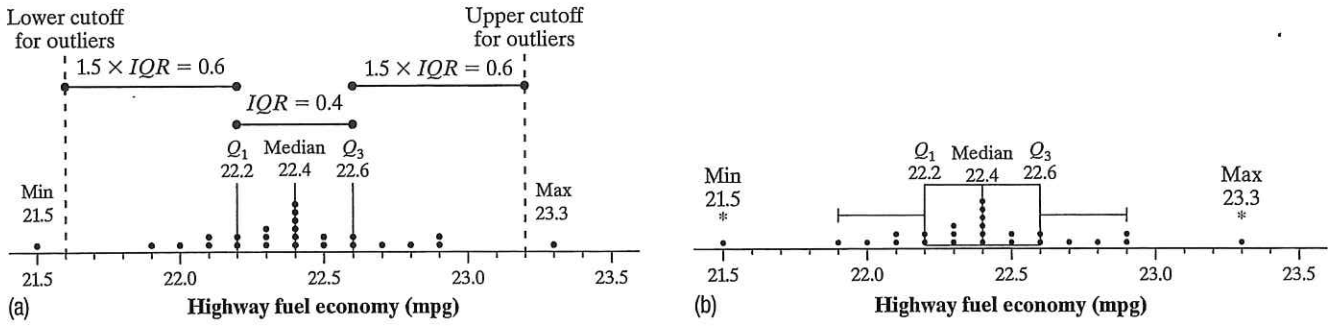


FIGURE 1.15 A visual illustration of how to make a boxplot for the Toyota 4Runner highway gas mileage data. (a) Dotplot of the data with the five-number summary and $1.5 \times IQR$ marked. (b) Boxplot of the data with outliers identified (*).

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 Q_1 Median

2.0 2.8 3.4 3.6 4.0 (4.0) 5.4 5.4 6.0 6.0 6.1 (6.6)

9.6 9.6 11.0 11.9 12.4 (12.7) 13.0 14.0 15.0 31.0 33.0

Q_3 Max

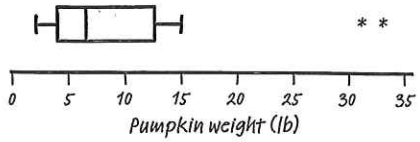
$IQR = Q_3 - Q_1 = 12.7 - 4.0 = 8.7$

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.

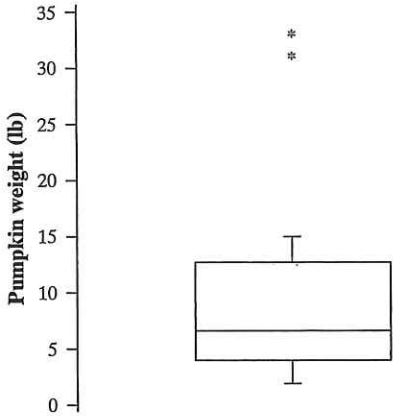
To make the boxplot:
- **Find the five-number summary.**
- **Identify outliers.**
- **Draw and label the axis.**
- **Scale the axis.**
- **Draw a box.**
- **Mark the median.**
- **Draw whiskers** to the smallest and largest data values that are *not* outliers. Mark outliers with an asterisk.



(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.

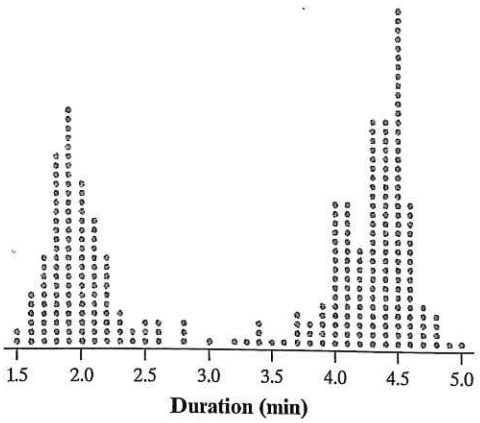
We know the distribution is skewed to the right because the left half of the distribution varies from 2.0 to 6.6 pounds, while the right half of the distribution (excluding outliers) varies from 6.6 to 15.0 pounds.

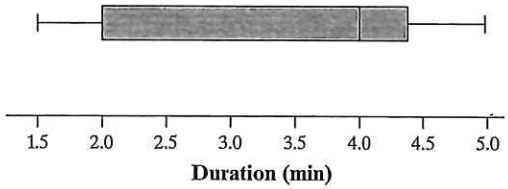
FOR PRACTICE, TRY EXERCISE 111



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

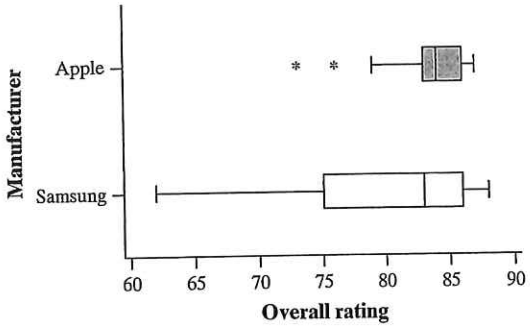
Boxplots are especially effective for comparing the distribution of a quantitative variable in two or more groups, as seen in the following example.

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

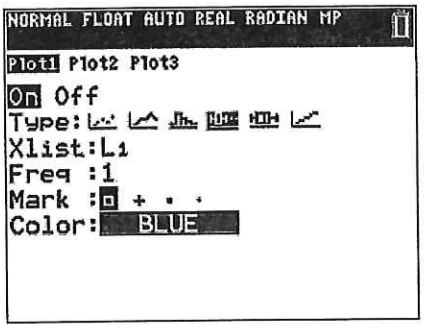
You can use technology to make boxplots, as the following Technology Corner illustrates.

4. Technology Corner MAKING BOXPLOTS

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

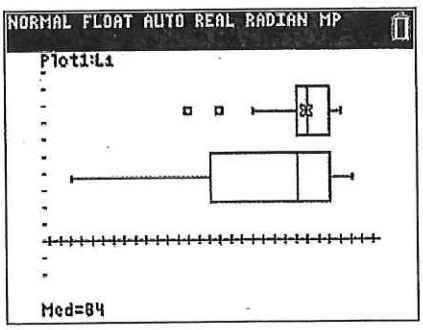
The TI-83/84 can plot up to three boxplots in the same viewing window. Let's use the calculator to make parallel boxplots of the overall rating data for Apple and Samsung tablets.

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 × *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.

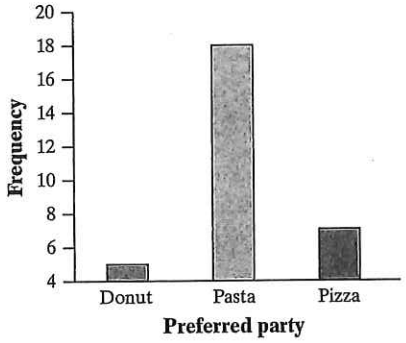
Chapter 1 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T1.1 You record the age, marital status, and earned income of a sample of 1463 women. The number and type of variables you have recorded are

(a) 3 quantitative, 0 categorical.
(b) 4 quantitative, 0 categorical.
(c) 3 quantitative, 1 categorical.
(d) 2 quantitative, 1 categorical.
(e) 1 quantitative, 2 categorical.

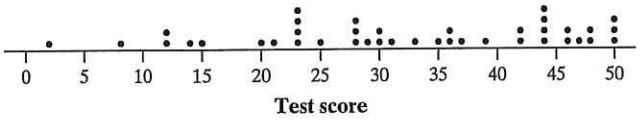
T1.2 The students in Mr. Tyson's high school statistics class were recently asked if they would prefer a pasta party, a pizza party, or a donut party. The following bar graph displays the data.



This graph is misleading because

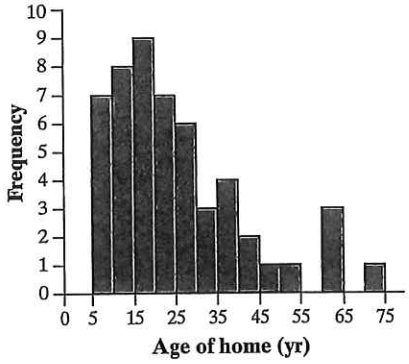
(a) it should be a histogram, not a bar graph.
(b) there should not be gaps between the bars.
(c) the bars should be arranged in decreasing order by height.
(d) the vertical axis scale should start at 0.
(e) preferred party should be on the vertical axis and number of students should be on the horizontal axis.

T1.3 Forty students took a statistics test worth 50 points. The dotplot displays the data. The third quartile is



(a) 45. (b) 44. (c) 43.
(d) 32. (e) 23.

Questions T1.4–T1.6 refer to the following setting. Realtors collect data in order to serve their clients more effectively. In a recent week, data on the age of all homes sold in a particular area were collected and displayed in this histogram.



T1.4 Which of the following could be the median age?

(a) 19 years (b) 24 years (c) 29 years
(d) 34 years (e) 39 years

T1.5 Which of the following is most likely true?

(a) mean > median, range < *IQR*
(b) mean < median, range < *IQR*
(c) mean > median, range > *IQR*
(d) mean < median, range > *IQR*
(e) mean = median, range > *IQR*

T1.6 The standard deviation of the distribution of house age is about 16 years. Interpret this value.

(a) The age of all houses in the sample is within 16 years of the mean.
(b) The gap between the youngest and oldest house is 16 years.
(c) The age of all the houses in the sample is 16 years from the mean.
(d) The gap between the first quartile and the third quartile is 16 years.
(e) The age of the houses in the sample typically varies by about 16 years from the mean age.

T1.7 The mean salary of all female workers is $35,000. The mean salary of all male workers is $41,000. What must be true about the mean salary of all workers?

(a) It must be $38,000.
(b) It must be larger than the median salary.
(c) It could be any number between $35,000 and $41,000.
(d) It must be larger than $38,000.
(e) It cannot be larger than $40,000.

Questions T1.8 and T1.9 refer to the following setting. A survey was designed to study how business operations vary by size. Companies were classified as small, medium, or large. Questionnaires were sent to 200 randomly selected businesses of each size. Because not all questionnaires are returned, researchers decided to investigate the relationship between the response rate and the size of the business. The data are given in the following two-way table.

		Business size		
		Small	Medium	Large
Response?	Yes	125	81	40
	No	75	119	160

T1.8 What percent of all small companies receiving questionnaires responded?

(a) 12.5% (b) 20.8% (c) 33.3%
(d) 50.8% (e) 62.5%

T1.9 Which of the following conclusions seems to be supported by the data?

(a) There are more small companies than large companies in the survey.

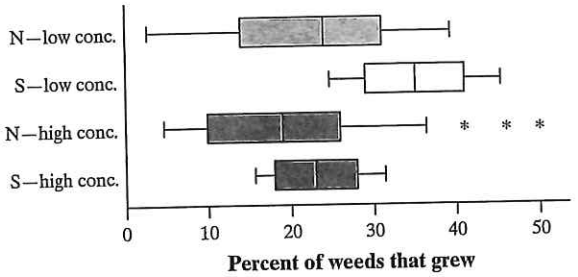
(b) Small companies appear to have a higher response rate than medium or big companies.

(c) Exactly the same number of companies responded as didn't respond.

(d) Overall, more than half of companies responded to the survey.

(e) If we combined the medium and large companies, then their response rate would be equal to that of the small companies.

T1.10 An experiment was conducted to investigate the effect of a new weed killer to prevent weed growth in onion crops. Two chemicals were used: the standard weed killer (S) and the new chemical (N). Both chemicals were tested at high and low concentrations on 50 test plots. The percent of weeds that grew in each plot was recorded. Here are some boxplots of the results.



Which of the following is *not* a correct statement about the results of this experiment?

(a) At both high and low concentrations, the new chemical results in better weed control than the standard weed killer.

(b) For both chemicals, a smaller percentage of weeds typically grew at higher concentrations than at lower concentrations.

(c) The results for the standard weed killer are less variable than those for the new chemical.

(d) High and low concentrations of either chemical have approximately the same effects on weed growth.

(e) Some of the results for the low concentration of weed killer show a smaller percentage of weeds growing than some of the results for the high concentration.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T1.11 You are interested in how many contacts older adults have in their smartphones. Here are data on the number of contacts for a random sample of 30 elderly adults with smartphones in a large city:

7	20	24	25	25	28	28	30	32	35
42	43	44	45	46	47	48	48	50	51
72	75	77	78	79	83	87	88	135	151

(a) Construct a histogram of these data.

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

(c) Would it be better to use the mean and standard deviation or the median and *IQR* to describe the center and variability of this distribution? Why?

T1.12 A study among the Pima Indians of Arizona investigated the relationship between a mother's diabetic status and the number of birth defects in her children. The results appear in the two-way table.

		Diabetic status		
		Nondiabetic	Prediabetic	Diabetic
Number of birth defects	None	754	362	38
	One or more	31	13	9

(a) What proportion of the women in this study had a child with one or more birth defects?

(b) What percent of the women in this study were diabetic or prediabetic, and had a child with one or more birth defects?

(c) Make a segmented bar graph to display the distribution of number of birth defects for the women with each of the three diabetic statuses.

(d) Describe the nature of the association between mother's diabetic status and number of birth defects for the women in this study.

T1.13 The back-to-back stemplot shows the lifetimes of several Brand X and Brand Y batteries.

Brand X		Brand Y
	1	
	1	7
	2	2
	2	6
2110	3	
99775	3	
3221	4	223334
	4	56889
4	5	0
5	5	

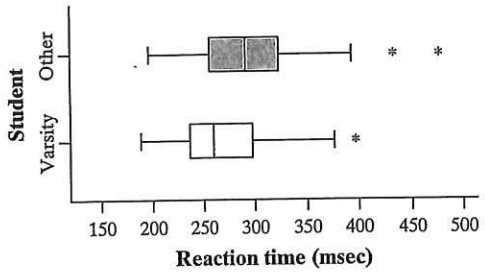
Key: 4|2 represents 420–429 hours.

(a) What is the longest that any battery lasted?

(b) Give a reason someone might prefer a Brand X battery.

(c) Give a reason someone might prefer a Brand Y battery.

T1.14 Catherine and Ana suspect that athletes (i.e., students who have been on at least one varsity team) typically have a faster reaction time than other students. To test this theory, they gave an online reflex test to 33 varsity athletes at their school and 29 other students. Here are parallel boxplots and numerical summaries of the data on reaction times (in milliseconds) for the two groups of students. Write a few sentences comparing the distribution of reaction time for the two types of students.



Student	n	Mean	StDev	Min	Q_1	Med	Q_3	Max
Other	29	297.3	65.9	197.0	255.0	292.0	325.0	478.0
Athlete	33	270.1	57.7	189.6	236.0	261.0	300.0	398.0

Chapter 1 Project American Community Survey

Each month, the U.S. Census Bureau selects a random sample of about 300,000 U.S. households to participate in the American Community Survey (ACS). The chosen households are notified by mail and invited to complete the survey online. The Census Bureau follows up on any uncompleted surveys by phone or in person. Data from the ACS are used to determine how the federal government allocates over $400 billion in funding for local communities.

The file **acs survey ch1 project.xls**, which can be accessed from the book's website at highschool.bfwpub.com/tps6e, contains data for 3000 randomly selected households in one month's ACS survey. Download the file to a computer for further analysis using the application specified by your teacher.

Each row in the spreadsheet describes a household. A serial number that identifies the household is in the first column. The other columns contain values of several variables. See the code sheet on the book's website for details on how each variable is recorded. Note that all the categorical variables have been coded to have numerical values in the spreadsheet.

Use the files provided to answer the following questions.

1. How many variables are recorded? Classify each one as categorical or quantitative.
2. Examine the distribution of location (division or region) for the households in the sample. Make a bar graph to display the data. Then calculate numerical summaries (counts, percents, or proportions). Describe what you see.
3. Explore the relationship between two categorical variables of interest to you. Summarize the data in a two-way table. Then calculate appropriate conditional relative frequencies and make a side-by-side or segmented bar graph. Write a few sentences comparing the distributions.
4. Analyze the distribution of household income (HINCP) using appropriate graphs and numerical summaries.
5. Compare the distribution of a quantitative variable that interests you in two or more groups. For instance, you might compare the distribution of number of people in a family (NPF) by region. Make appropriate graphs and calculate numerical summaries. Then write a few sentences comparing the distributions.