

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.

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.”

		Frequency table		Relative frequency table	
Preferred method	Tally	Preferred method	Frequency	Preferred method	Relative frequency
Cell phone	II	Cell phone	2	Cell phone	$2/10 = 0.20$ or 20%
Facebook	I	Facebook	1	Facebook	$1/10 = 0.10$ or 10%
In person	III	In person	3	In person	$3/10 = 0.30$ or 30%
Text messaging	IIII	Text messaging	4	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.

PD Section 1.1 Overview

The video overview of Section 1.1 (for teachers) can be found by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher's Resources on the book's digital platform.

TRM Section 1.1 Alternate Examples

You can find the Alternate Examples for this section in Microsoft Word format by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher's Resources on the book's digital platform.

Teaching Tip

Frequency tables show counts. Relative frequency tables show the counts *relative* to the total (proportion or percent).

Teaching Tip

It is a good idea for students to practice creating each type of graph by hand to make sure they truly understand what the graph represents. More important, students need to be able to interpret graphs and use them to answer questions.

AP® EXAM TIP

Many students spend lots of time constructing graphs only to forget the labels. This will always result in a deduction in the student's score. It is very important for students to make sure their graphs display data with the proper labels and scaling.

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

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%

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.

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.

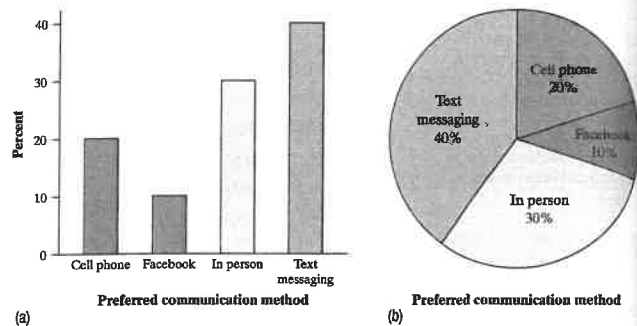


FIGURE 1.2 (a) Bar graph and (b) pie chart of the distribution of preferred communication method for a random sample of 10 Canadian students.

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.

Teaching Tip: Using Technology

Go to the Student Site at highschool.bfwpub.com/updatedtps6e and choose "One Categorical Variable" from the Extra Applets menu. Type in each category and the frequency of each. Click "Begin Analysis." Graph options here are the bar graph and pie chart.

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

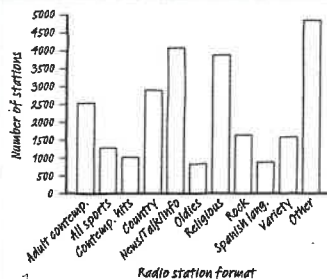
PROBLEM: Arbitron, the rating service for radio audiences, categorizes U.S. radio stations in terms of the kinds of programs they broadcast. The frequency table summarizes the distribution of station formats in a recent year.²

- (a) Identify the individuals in this data set.
(b) Make a frequency bar graph of the data. Describe what you see.

SOLUTION:

(a) U.S. radio stations

(b)



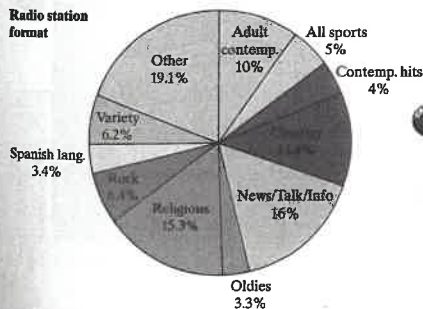
To make the bar graph:

- 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

Radio station format



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.

ALTERNATE EXAMPLE

Skill 2.B

Who listens to the Beatles?**Making and interpreting bar graphs****PROBLEM:**

Spotify tracks data about listeners. The frequency table summarizes the distribution of ages for a sample of 100 Beatles listeners.

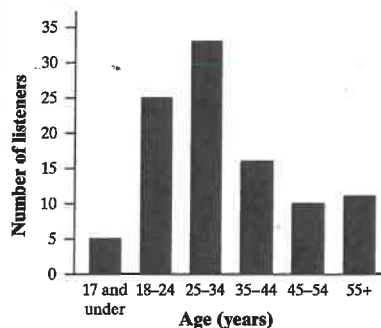
Age (years)	Number of listeners
17 and under	5
18–24	25
25–34	33
35–44	16
45–54	10
55+	11

- (a) Identify the individuals in this data set.
(b) Make a frequency bar graph of the data. Describe what you see.

SOLUTION:

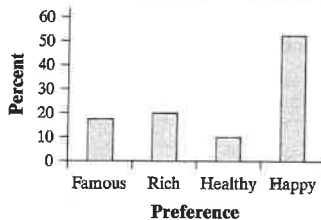
(a) 100 Spotify users who listen to the Beatles.

(b)



Of the sample of 100 Spotify users who listen to the Beatles, the most frequent ages are 25–34 years (33) and 18–24 years (25), while the least frequent are 17 years and under (5), and 45–54 years (10).

✓ Answers to CYU

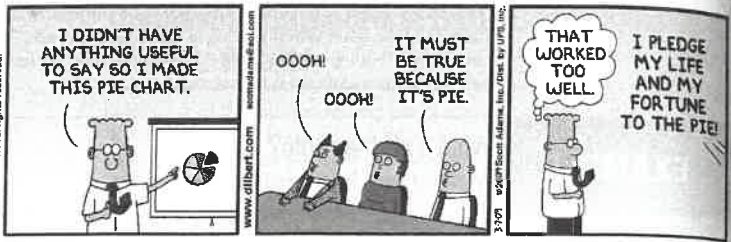


A slight majority (52.5%) of students in the sample said that they would prefer to be Happy. Rich and Famous were preferred about equally (20% and 17.5%). The least popular choice was Healthy (10%).

Teaching Tip

There are many good websites that illustrate deceptive graphs. For example, try www.datavis.ca/gallery/index.php or do an Internet search for “deceptive graphs.” The classic book *How to Lie with Statistics* by Darrell Huff also includes some good examples.

© 1997 Scott Adams. Used by permission of UNIVERSAL UCLICK. All rights reserved.



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 $\times$ 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!

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

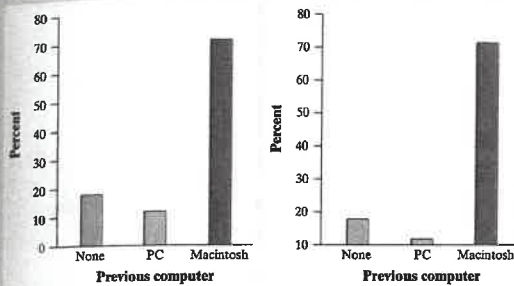


Justin Sullivan/Getty Images

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!

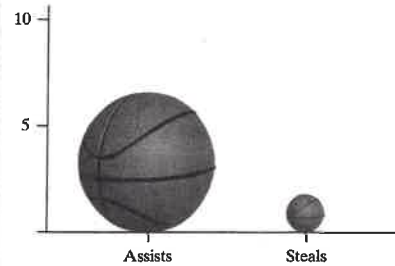
FOR PRACTICE, TRY EXERCISE 19

ALTERNATE EXAMPLE

Is Steph Curry the greatest of all time?
Beware the pictograph!

PROBLEM:

In the 2017–2018 regular season, basketball player Stephen Curry averaged 6.1 assists per game and 1.6 steals per game. Explain what is wrong with the following graph and how to make it better.



SOLUTION:

Although the heights of the basketballs are correct, our eyes respond to the areas of the basketballs. Curry averaged between 3 and 4 times as many assists as steals, but the area of the basketball representing assists is more than 10 times as big. It would be better to use equally wide bars for assists and steals so that only the height of the bars determines the area.



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

Teaching Tip

In *TPS 6e*, we follow a convention for two-way tables that places the values of the explanatory variable in the columns and the values of the response variable in the rows. (A response variable measures an outcome of a study, and an explanatory variable may help explain or predict changes in the response variable.) This matches how we will construct scatterplots in Chapter 3, plotting the explanatory variable on the horizontal axis and the response variable on the vertical axis. Not everyone follows this convention, however, so it isn't wrong to place the values of the explanatory variable in the rows and the values of the response variable in the columns.

Making Connections

Students will use two-way tables in Chapter 5 to calculate probabilities and again in Chapter 12 to perform a chi-square test.

Making Connections

In Chapter 5, these percents and proportions will be identified as probabilities.



Francis Z. Stutts/istock

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
	Rent	497
	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
	Snowmobile renter	497	77
	Snowmobile owner	279	16
	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.

Teaching Tip

Ask students to identify where the margins would be in a Microsoft Word document. Most will recognize that they are on the top, bottom, right, and left. We use the values on the right and the bottom (totals) to calculate marginal relative frequencies.

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

We could call this the marginal distribution of snowmobile use.

In the body, a joint frequency table, a joint frequency table, where a row and a column intersect.

EXA

PROBLEM across the A gers got off table gives a and who dia

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:

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

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

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

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

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

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

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*.

In the body, a joint is where two bones come together. In a two-way table, a *joint frequency* is shown where a row and column come together.

DEFINITION Joint relative frequency

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

EXAMPLE

A Titanic disaster Calculating marginal and joint relative frequencies

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



Making Connections

In Chapter 5, we will use probability to describe these fractions. For a randomly selected member of the sample:

$\frac{16}{1526} = 0.010 = 1.0\%$ will be $P(\text{environmental club AND own snowmobile})$.

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

$P(\text{no environmental club AND never use snowmobile})$.

ALTERNATE EXAMPLE Skills 2.C, 2.D

iPhone or Android?

Calculating marginal and joint relative frequencies

PROBLEM:

The Pew Research Center asked a random sample of 2024 adult cell-phone owners from the United States which type of cell phone they own: iPhone, Android, or other (including non-smartphones). Here are the results, broken down by age category:

		Age			
		18–34	35–54	55 +	Total
Type of phone	iPhone	169	171	127	467
	Android	214	189	100	503
	Other	134	277	643	1054
	Total	517	637	870	2024

- (a) What proportion of the adult cell-phone owners have an iPhone?
 (b) Find the distribution of ages for adult cell-phone owners using relative frequencies.
 (c) What percent of adult cell-phone owners are 55 or older and own an iPhone?

SOLUTION:

$$(a) \frac{467}{2024} = 0.231$$

$$(b) 18-34: \frac{517}{2024} = 0.255 = 25.5\%$$

$$35-54: \frac{637}{2024} = 0.315 = 31.5\%$$

$$55+: \frac{870}{2024} = 0.430 = 43.0\%$$

$$(c) \frac{127}{2024} = 0.063 = 6.3\%$$

- (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?

SOLUTION:

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

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

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

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

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

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

Start by finding the marginal totals.

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

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

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

✓ Answers to CYU

1. $113/338 = 0.334$; this value makes sense because there were three treatments, so we would expect about one-third of the subjects to be assigned to this treatment.

2. The distribution of change in depression is:

Full response: $91/338 = 0.269$

Partial response: $55/338 = 0.163$

No response: $192/338 = 0.568$

3. $27/338 = 0.08 = 8\%$

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).

Teaching Tip

If you are displaying the table with a projector or on the board, physically cover up the "No" column to answer the first question. In the second question, physically cover up the rows with "Never used" and "Snowmobile owner."

Preparing for Inference

The *P*-value of a significance test is a conditional probability. More specifically, it is the probability of getting evidence for the alternative hypothesis as strong as or stronger than the observed evidence, given that the null hypothesis is true.

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

ALTERNATE EXAMPLE

Skill 2.C

Do older people own iPhones?
Conditional relative frequencies

PROBLEM:

The Pew Research Center asked a random sample of 2024 adult cell-phone owners from the United States which type of cell phone they own: iPhone, Android, or other (including non-smartphones). Here are the results, broken down by age category:

Type of phone	Age			Total
	18–34	35–54	55 +	
iPhone	169	171	127	467
Android	214	189	100	503
Other	134	277	643	1054
Total	517	637	870	2024

- (a) What proportion of iPhone owners were 55 or older?
 (b) What percent of 18- to 34-year-olds owned Android phones?

SOLUTION:

$$(a) \frac{127}{467} = 0.272$$

$$(b) \frac{214}{517} = 0.414 = 41.4\%$$

Making Connections

In Chapter 5, these proportions will be referred to as *conditional probabilities* and will have the following notation:

$$P(\text{never} \mid \text{yes}) = 0.695$$

$$P(\text{rent} \mid \text{yes}) = 0.252$$

$$P(\text{own} \mid \text{yes}) = 0.052$$

- (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		Total
	No	Yes	
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:

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

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

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	$\frac{445}{1221} = 0.364 \text{ or } 36.4\%$	$\frac{212}{305} = 0.695 \text{ or } 69.5\%$
Rent	$\frac{497}{1221} = 0.407 \text{ or } 40.7\%$	$\frac{77}{305} = 0.252 \text{ or } 25.2\%$
Own	$\frac{279}{1221} = 0.229 \text{ or } 22.9\%$	$\frac{16}{305} = 0.052 \text{ or } 5.2\%$

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
(b) see
display
among
non-en-
viron-
mental
club members

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, (b) a segmented bar graph, and (c) a mosaic plot. 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). The bar widths in the mosaic plot are proportional to the number of survey respondents who are (305) and are not (1221) environmental club members.

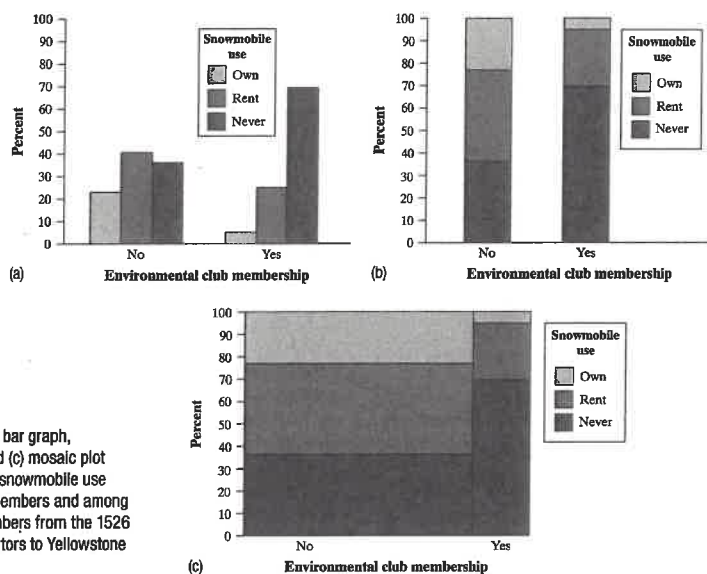


FIGURE 1.3 (a) Side-by-side bar graph, (b) segmented bar graph, and (c) mosaic plot displaying the distribution of snowmobile use among environmental club members and among non-environmental club members from the 1526 randomly selected winter visitors to Yellowstone National Park.

DEFINITION Side-by side bar graph, Segmented bar graph, Mosaic plot

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.

A **mosaic plot** is a modified segmented bar graph in which the width of each rectangle is proportional to the number of individuals in the corresponding category.

Teaching Tip

Segmented bar graphs and mosaic plots were not listed on the previous AP[®] Topic Outline, but they are now specifically mentioned in the new Course Framework.

Teaching Tip: Using Technology

Go to the Student Site at highschool.bfwpub.com/updatedtps6e and choose "Two Categorical Variables" from the Extra Applets. Type in the categories for each variable and the frequency of each. Click "Begin Analysis." Graph options here are the side-by-side bar graph and segmented bar graph.

Teaching Tip

Emphasize the concept of association between two variables. This will be a focus throughout the course. It is OK to use the word *relationship* as a synonym for *association*, but do not use the word *correlation*. It is reserved for a specific measure of the strength of a linear relationship between two quantitative variables.

Preparing for Inference

How do we calculate the percents for these hypothetical distributions? They are the marginal relative frequencies calculated earlier.

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

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

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

These percents will be used in Chapter 12 to calculate the expected values for a chi-square test for independence. In this test, we will assume that there is no association between the two variables in the population.

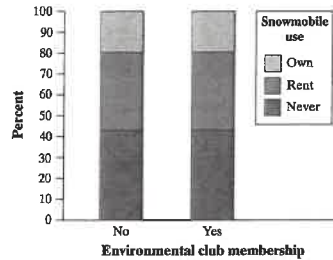
Preparing for Inference

In Chapter 12, the chi-square test for independence will be used to test for an association between two categorical variables in a population. Two variables with no association are said to be independent, while two variables that have an association are not independent.

All three 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 12.

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.

ALTERNATE EXAMPLE

Skills 2.C, 2.D

Do older people own iPhones? Conditional relative frequencies and association

PROBLEM:

The Pew Research Center asked a random sample of 2024 adult cell-phone owners from the United States which type of cell phone they own: iPhone, Android, or other (including non-smartphones). Here are the results, broken down by age category:

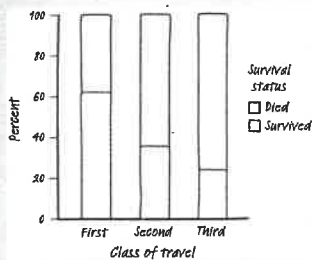
	Age			Total
	18–34	35–54	55 +	
iPhone	169	171	127	467
Android	214	189	100	503
Other	134	277	643	1054
Total	517	637	870	2024

(continues)

- (a) Find the distribution of survival status for each class of travel. Make a segmented bar graph to compare these distributions.
- (b) Describe what the graph in part (a) reveals about the association between class of travel and survival status for adult passengers on the *Titanic*.

SOLUTION:

- (a) First class: Survived: $\frac{197}{319} = 0.618 = 61.8\%$ Died: $\frac{122}{319} = 0.382 = 38.2\%$
- Second class: Survived: $\frac{94}{261} = 0.360 = 36.0\%$ Died: $\frac{167}{261} = 0.640 = 64.0\%$
- Third class: Survived: $\frac{151}{627} = 0.241 = 24.1\%$ Died: $\frac{476}{627} = 0.759 = 75.9\%$



To make the segmented bar graph:

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

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

FOR PRACTICE, TRY EXERCISE 29

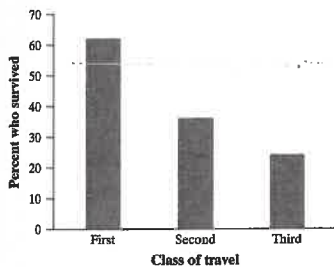


FIGURE 1.4 Bar graph comparing the percents of passengers who survived among each of the three classes of travel on the *Titanic*.

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*.

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

35–54:

$$\text{iPhone: } \frac{171}{637} = 0.268 = 26.8\%$$

$$\text{Android: } \frac{189}{637} = 0.297 = 29.7\%$$

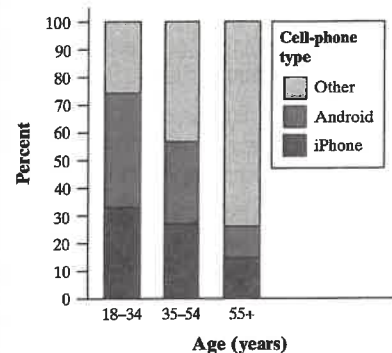
$$\text{Other: } \frac{277}{637} = 0.435 = 43.5\%$$

55+ :

$$\text{iPhone: } \frac{127}{870} = 0.146 = 14.6\%$$

$$\text{Android: } \frac{100}{870} = 0.115 = 11.5\%$$

$$\text{Other: } \frac{643}{870} = 0.739 = 73.9\%$$



- (b) Knowing an adult's age helps us predict his or her cell-phone type. A much higher percentage of people 18–34 years old have iPhones (32.7%) than those who are 35–54 years old (26.8%) or over 55 years old (14.6%). Also, a much higher percentage of people over 55 years old have "Other" phones (73.9%) than do people who are 35–54 years old (43.5%) or 18–34 years old (25.9%).

ALTERNATE EXAMPLE (continued)

- (a) Find the distribution of type of cell phone for each age group. Make a segmented bar graph to compare these distributions.
- (b) Describe what the graph in (a) reveals about the association between age and type of cell phone for adult cell-phone owners.

SOLUTION:

(a) 18–34:

$$\text{iPhone: } \frac{169}{517} = 0.327 = 32.7\%$$

$$\text{Android: } \frac{214}{517} = 0.414 = 41.4\%$$

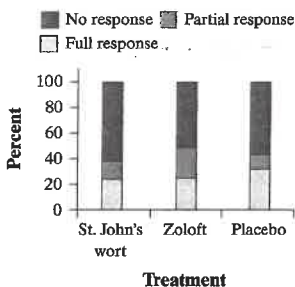
$$\text{Other: } \frac{134}{517} = 0.259 = 25.9\%$$

Preparing for Inference

In Chapter 4, we will make the distinction between an observational study and an experiment (where a treatment is imposed). While an observational study can provide evidence of an association, a properly designed experiment is needed to show causation.

✓ Answers to CYU

- $27/91 = 0.297$
- $70/113 = 0.619 = 61.9\%$
- The distribution of change in depression for the subjects receiving each of the three treatments is:
St. John's wort: Full response: $27/113 = 0.239$, Partial response: $16/113 = 0.142$, No response: $70/113 = 0.619$
Zoloft: Full response: $27/109 = 0.248$, Partial response: $26/109 = 0.239$, No response: $56/109 = 0.514$
Placebo: Full response: $37/116 = 0.319$, Partial response: $13/116 = 0.112$, No response: $66/116 = 0.569$



- There does not appear to be a strong association between treatment and change in depression for these subjects because the distribution of response status is very similar for the three different treatments. Also, note that the treatment with the highest rate of "full response" was the placebo.

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
Full response	27	27	37
Partial response	16	26	13
No response	70	56	66

- What proportion of subjects who showed a full response took St. John's wort?
- What percent of subjects who took St. John's wort showed no response?
- 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.
- 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.

	No	Yes	All
Never	445	212	657
	67.73	32.27	100.00
	36.45	69.51	45.05
	29.16	13.89	43.05
Renter	497	77	574
	66.59	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	15.33
	18.28	1.05	19.33
All	1221	305	1526
	80.03	19.99	100.00
	100.00	100.00	100.00
	85.01	19.99	100.00

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

Teaching Tip

Remind students that they can watch a video of each of the Technology Corners by clicking on the link in the e-Book or viewing the videos on the Student Site at highschool.bfwpub.com/updatedtps6e.

Teaching Tip: Using Technology

Make sure to point out the "key" at the bottom of the computer output. Almost all computer output showing two-way tables will have a key of some kind to explain what each value in the table represents.

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**, a **segmented bar graph**, or a **mosaic plot** 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/updatedtps6e.

1. Analyzing two-way tables

Page 22

TRM Section 1.1 Quiz

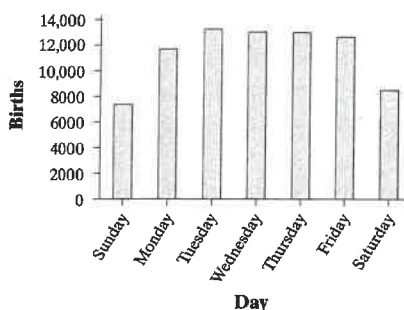
There is one quiz available for this section in the Teacher's Resource Materials. Click on the link in the TE-Book, look on the TRFD, or download from the Teacher's Resources on the book's digital platform. You can also create your own quiz using the ExamView® Assessment Suite that is part of the TPS6 program. Questions are coded by Learning Target to make it easy to build parallel quizzes.

Full Solutions to Section 1.1 Exercises

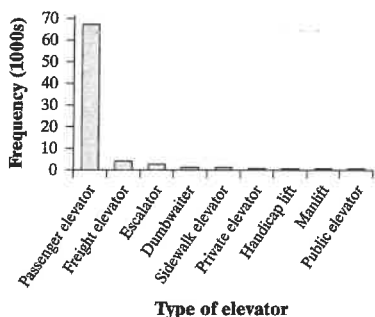
Click on the link in the TE-Book, open the TRFD, or download from the Teacher's Resources on the book's digital platform.

Answers to Section 1.1 Exercises

1.11 (a) The individuals are the babies. **(b)** It appears that births occurred with similar frequencies on weekdays (Monday through Friday), but with noticeably smaller frequencies on the weekend days (Saturday and Sunday).



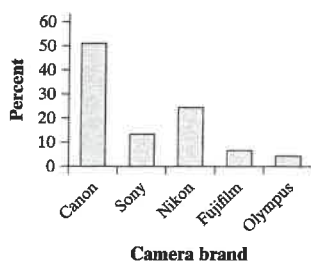
1.12 (a) The individuals are the elevators in New York City. **(b)** A large majority of the elevators in New York City are passenger elevators. All other types of elevators are much less common.



1.13 First, a relative frequency table must be constructed.

Camera brand	Relative frequency
Canon	$23/45 = 51.1\%$
Sony	$6/45 = 13.3\%$
Nikon	$11/45 = 24.4\%$
Fujifilm	$3/45 = 6.7\%$
Olympus	$2/45 = 4.4\%$

The relative frequency bar graph is given below.



The most popular brand of camera among the 45 most recent purchases on the Internet auction site is Canon, followed by Nikon, Sony, Fujifilm, and Olympus. Canon is the overwhelming favorite, with over 50% of the customers purchasing this brand. Also noteworthy is that almost 25% of the customers purchased a Nikon camera.

1.14 First, a relative frequency table must be constructed.

Breed group	Relative frequency
Whippet	$3/42 = 7.1\%$
Mixed breed	$12/42 = 28.6\%$
Other purebred	$2/42 = 4.8\%$
Labrador retriever	$3/42 = 7.1\%$
Border collie	$13/42 = 31.0\%$
Australian shepherd	$9/42 = 21.4\%$

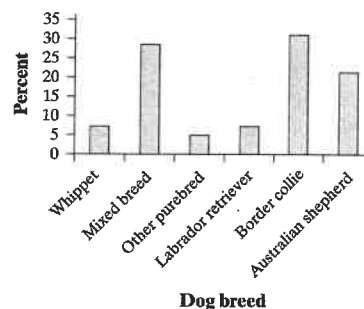
14. Disc dogs Here is a list of the breeds of dogs that won the World Canine Disc Championships from 1975 through 2016. Make a relative frequency bar graph for these data. Describe what you see.

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

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

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

- What percent of vehicles would fall in the "Other" category?
- Make a bar graph to display the data. Describe what you see?
- Would it be appropriate to make a pie chart of these data? Explain.



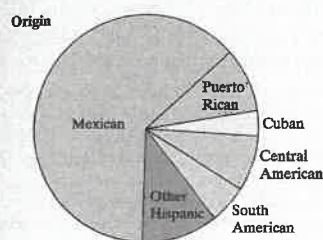
At the World Canine Disc Championships from 1975 to 2016, Border collie dogs won most often (31% of the time), followed by Mixed breed (28.6% of the time) and Australian shepherds (21.4% of the time). Whippets (7.1%), Labrador retrievers (7.1%), and other purebreds (4.8%) were the least frequent winners.

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

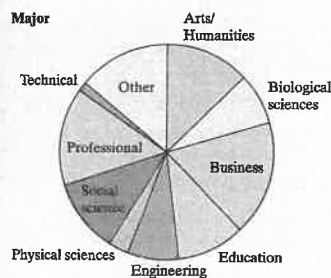
Type of spam	Percent
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

- (a) What percent of spam would fall in the "Other" category?
 (b) Make a bar graph to display the data. Describe what you see.
 (c) Would it be appropriate to make a pie chart of these data? Explain.

17. **Hispanic origins** Here is a pie chart prepared by the Census Bureau to show the origin of the more than 50 million Hispanics in the United States in 2010.¹¹ About what percent of Hispanics are Mexican? Puerto Rican?

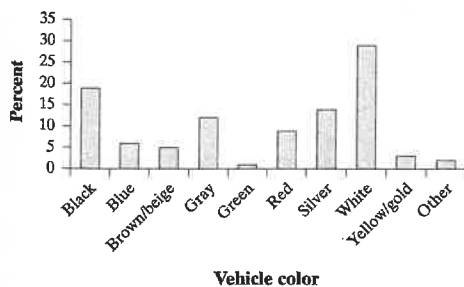


18. **Which major?** About 3 million first-year students enroll in colleges and universities each year. What do they plan to study? The pie chart displays data on the percent of first-year students who plan to major in several disciplines.¹² About what percent of first-year students plan to major in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



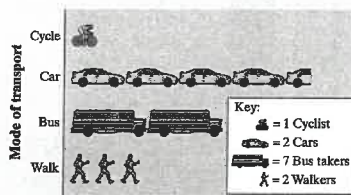
1.15 (a) 2%

(b)

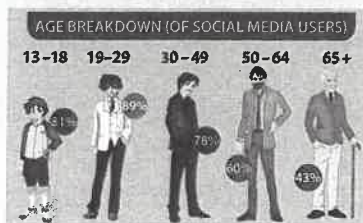


The most popular color of vehicles sold that year was white, followed by black, silver, and gray. It appears that a majority of car buyers that year preferred vehicles that were shades of black and white. (c) It would be appropriate to make a pie

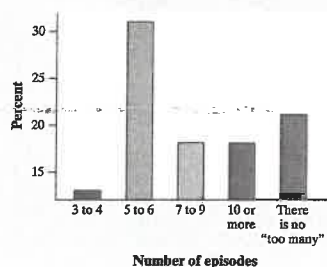
19. **Going to school** Students in a high school statistics class were given data about the main method of transportation to school for a group of 30 students. They produced the pictograph shown. Explain how this graph is misleading.



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



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



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

followed closely by Financial (20%), and Adult (19%). (c) It would be appropriate to make a pie chart of these data (including the other category) because the numbers in the table refer to parts of a single whole.

1.17 Estimates will vary, but should be close to 63% Mexican and 9% Puerto Rican.

1.18 Estimates will vary, but should be close to 17% Business and 10% Social science.

1.19 The areas of the pictures should be proportional to the numbers of students they represent. As drawn, it appears that most of the students arrived by car but, in reality, most came by bus (14 took the bus, 9 came in cars).

1.20 Based on the heights of the people, it appears that the peak of social media use is ages 50-64, but actually it is ages 19-29.

1.21 By starting the vertical scale at 12 instead of 0, it looks like the percent of binge-watchers who think that 5 to 6 episodes are too many to watch in one viewing session is almost 20 times higher than the percent of binge-watchers who think that 3 to 4 episodes are too many to watch in one viewing session. In truth, the percent of binge-watchers who think that 3 to 4 episodes are too many to watch in one viewing session (31%) is less than 3 times higher than the percent of binge-watchers who think that 5 to 6 episodes are too many to watch in one viewing session (13%). Similar arguments can be made for the relative sizes of the other categories represented in the bar graph.

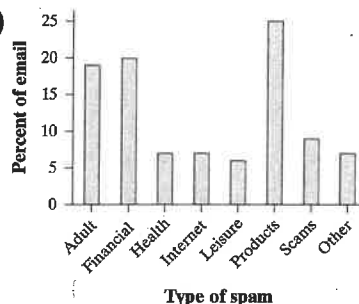
Teaching Tip

Exercises 17 and 18 help students see that it is hard to determine relative frequencies from a pie chart. Bar graphs are much easier to use.

chart of these data (including the other category) because the numbers in the table refer to parts of a single whole.

1.16 (a) 7%

(b)



The bar graph shows that the most common type of spam is Products (25% of spam email),

1.22 By starting the vertical scale at 53 instead of 0, it looks as if the percent of Democrats who agree with the decision is almost 10 times higher than the percent of Republicans and Independents who agree. In truth, the percent of Democrats who agree with the decision (62%) is only slightly higher than the percent of Republicans (54%) and Independents (54%) who agree.

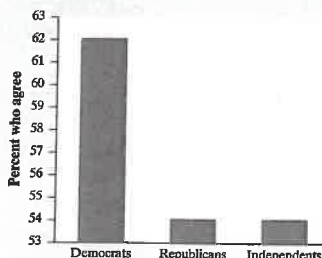
1.23 (a) $50/150 = 0.333$ **(b)** $29/150 = 19.3\%$ said they saw broken glass at the accident; $121/150 = 80.7\%$ said they did not; 14% said they saw broken glass at the accident; 12% said they saw broken glass at the accident. **(c)** 10.67%

1.24 (a) 0.518 **(b)** Fly: $99/415 = 0.239 = 23.9\%$, Freeze time: $96/415 = 0.231 = 23.1\%$, Invisibility: $67/415 = 0.161 = 16.1\%$, Superstrength: $43/415 = 0.104 = 10.4\%$, and Telepathy: $110/415 = 0.265 = 26.5\%$ **(c)** 10.6%

1.25 (a) 71.25% **(b)** 0.633 **(c)** 12.1%

1.26 (a) 55.6% **(b)** 0.690 **(c)** 28.9%

the misleading nature of its graph, the network posted a corrected version.)



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

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

- (a) What proportion of subjects were given the control treatment?
- (b) Find the distribution of responses about whether there was broken glass at the accident for the subjects in this study using relative frequencies.
- (c) What percent of the subjects were given the "smashed into" treatment and said they saw broken glass at the accident?
- 24. Superpowers** A total of 415 children from the United Kingdom and the United States who completed a survey in a recent year were randomly selected. Each student's country of origin was recorded along with which superpower they would most like to have: the ability to fly, ability to freeze time, invisibility, superstrength, or telepathy (ability to read minds). The data are summarized in the following table.¹⁵

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

- (a) What proportion of students in the sample are from the United States?
- (b) Find the distribution of superpower preference for the students in the sample using relative frequencies.
- (c) What percent of students in the sample are from the United Kingdom and prefer telepathy as their superpower preference?
- 25. Body image** A random sample of 1200 U.S. college students was asked, "What is your perception of your own body? Do you feel that you are overweight, underweight, or about right?" The two-way table summarizes the data on perceived body image by gender.¹⁶

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

- (a) What percent of respondents feel that their body weight is about right?
- (b) What proportion of the sample is female?
- (c) What percent of respondents are males and feel that they are overweight or underweight?
- 26. Python eggs** How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers randomly assigned newly laid eggs to one of three water temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. The two-way table summarizes the data on whether or not the eggs hatched.¹⁷

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

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

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

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

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

28. Superpower Refer to Exercise 24.

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

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

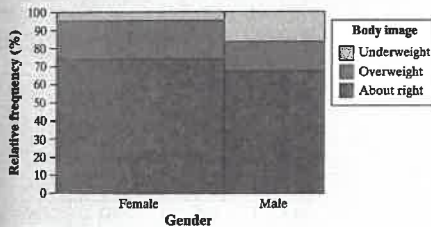
- (a) Find the distribution of responses about whether there was broken glass at the accident for each of the three treatment groups. Make a segmented bar graph to compare these distributions.
- (b) Describe what the graph in part (a) reveals about the association between response about broken glass at the accident and treatment received for the subjects in the study.

30. Superpower Refer to Exercise 24.

- (a) Find the distribution of superpower preference for the students in the sample from each country (i.e., the United States and the United Kingdom). Make a segmented bar graph to compare these distributions.
- (b) Describe what the graph in part (a) reveals about the association between country of origin and superpower preference for the students in the sample.

31. Body image Refer to Exercise 25.

- (a) Of the respondents who felt that their body weight was about right, what proportion were female?
- (b) Of the female respondents, what percent felt that their body weight was about right?
- (c) The mosaic plot displays the distribution of perceived body image by gender. Describe what this graph reveals about the association between these two variables for the 1200 college students in the sample.



1.27 (a) 0.241 (b) 68%

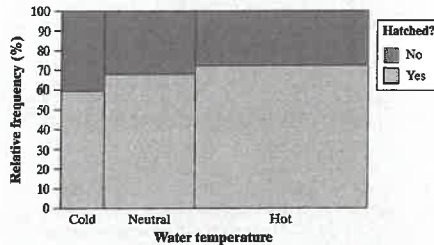
1.28 (a) 0.552 (b) 10%

1.29 (a) The distributions of responses for the three treatment groups are:

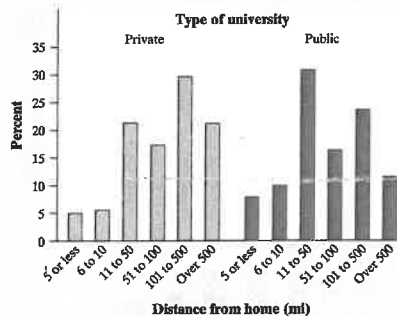
"Smashed into"	"Hit"	Control
Yes: 16/50 = 32%	Yes: 7/50 = 14%	Yes: 6/50 = 12%
No: 34/50 = 68%	No: 43/50 = 86%	No: 44/50 = 88%

32. Python eggs Refer to Exercise 26.

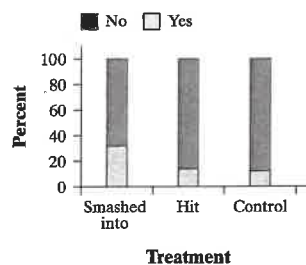
- (a) Of the eggs that hatched, what proportion were randomly assigned to hot water?
- (b) Of the eggs that were randomly assigned to hot water, what percent hatched?
- (c) The mosaic plot displays the distribution of hatching status by water temperature. Describe what this graph reveals about the association between these two variables for the python eggs in this experiment.



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



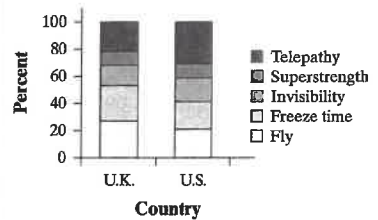
The segmented bar graph appears below.



(b) The segmented bar graph reveals that there is an association between opinion about broken glass at the accident and treatment received for subjects in the study. Knowing which treatment a subject received helps us predict whether or not that person will respond that he or she saw broken glass at the accident.

1.30 (a) The distributions of superpower preference for the students in the sample from each country are:

United Kingdom	United States
Fly: 54/200 = 27%	Fly: 45/215 = 20.9%
Freeze time: 52/200 = 26%	Freeze time: 44/215 = 20.5%
Invisibility: 30/200 = 15%	Invisibility: 37/215 = 17.2%
Superstrength: 20/200 = 10%	Superstrength: 23/215 = 10.7%
Telepathy: 44/200 = 22%	Telepathy: 66/215 = 30.7%



(b) There is an association between country of origin and superpower preference for the students in the sample. Students from the United States were more likely to choose the superpower of telepathy and less likely to choose the superpowers of flying or freezing time than were students from the United Kingdom.

1.31 (a) 0.655 (b) 73.7% (c) Based on the mosaic plot, there is an association between perceived body image and gender. Males are about 4 times as likely as females to perceive that they are overweight. Males are also less likely to perceive that they are underweight. However, the overwhelming majority of both genders perceives that their body image is about right.

1.32 (a) 0.581 (b) 72.1% (c) Based on the mosaic plot, there is an association between water temperature and whether the eggs hatched or not. The python eggs were least likely to hatch in cold water (59.3% hatched) and most likely to hatch in hot water (72.1% hatched), so the chance of hatching increased as the water temperature increased.

1.33 Answers may vary. Regardless of whether a student went to a private or public college, most students chose a school that was at least 11 miles from home. Those who went to a public university were most likely to choose a school that was 11 to 50 miles from home (about 30%), while those who went to a private university were most likely to choose a school that was 101 to 500 miles from home (about 29%).

1.34 Answers may vary. Asians overwhelmingly preferred white cars (about 41%). North Americans also preferred white cars, but not as strongly (about 26%). Also noteworthy is that Asians preferred gray and red cars about 7% and 6% of the time, respectively, while North Americans preferred those colors with approximately twice those frequencies.

1.35 (a) The graph reveals that as age increases, the percent that use smartphones for navigation decreases. **(b)** It would not be appropriate to make a pie chart of these data because the category percentages are not parts of the same whole.

1.36 (a) The graph reveals that as age increases, the percentage of people who attended a movie decreases. **(b)** It would not be appropriate to make a pie chart of these data because the category percentages are not parts of the same whole.

1.37 Answers will vary. Two possible tables are given below.

10	40	30	20
50	0	30	20

1.38 (a) The distributions of survival status for males and females within each class of travel are:

First class

Female survived: $140/144 = 97.2\%$

Female died: $4/144 = 2.8\%$

Male survived: $57/175 = 32.6\%$

Male died: $118/175 = 67.4\%$

Second class

Female survived: $80/93 = 86\%$

Female died: $13/93 = 14\%$

Male survived: $14/168 = 8.3\%$

Male died: $154/168 = 91.7\%$

Third class

Female survived: $76/165 = 46.1\%$

Female died: $89/165 = 53.9\%$

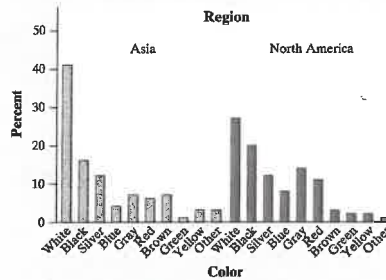
Male survived: $75/462 = 16.2\%$

Male died: $387/462 = 83.8\%$

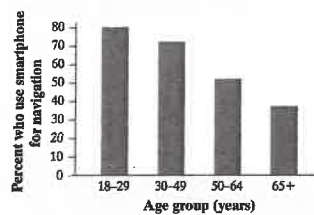
Regardless of class of travel, women survived the disaster at higher rates than men. Of those who were traveling first class, women were about 3 times more likely to survive than men. Of those who were traveling second class, women were about 10 times more likely to survive

than men. Of those who were traveling third class, women were about 3 times more likely to survive than men. **(b)** For women, there is a clear association between class of travel and survival status. Women in first class were more likely to survive than women in second class, who were more likely to survive than women in third class. For men, the same relationship does not hold true. Although men traveling first class were more likely to survive than men traveling second class, the men traveling third class were more likely to survive than men traveling second class as well.

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

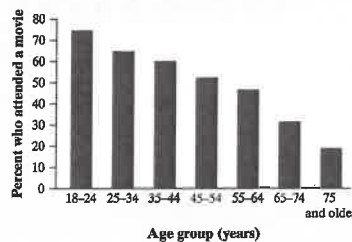


- 35. Phone navigation** The bar graph displays data on the percent of smartphone owners in several age groups who say that they use their phone for turn-by-turn navigation.²⁰



- (a)** Describe what the graph reveals about the relationship between age group and use of smartphones for navigation.
(b) Would it be appropriate to make a pie chart of the data? Explain.

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



- (a)** Describe what the graph reveals about the relationship between age group and movie attendance.
(b) Would it be appropriate to make a pie chart of the data? Explain.
- 37. Marginal totals aren't the whole story** Here are the row and column totals for a two-way table with two rows and two columns:

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

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

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

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

- (a)** Find the distributions of survival status for males and for females within each class of travel. Did women survive the disaster at higher rates than men? Explain.
(b) In an earlier example, we noted that survival status is associated with class of travel. First-class passengers had the highest survival rate, while third-class passengers had the lowest survival rate. Does this same relationship hold for both males and females in all three classes of travel? Explain.
- 39. Simpson's paradox** Accident victims are sometimes taken by helicopter from the accident scene to a hospital. Helicopters save time. Do they also save lives? The two-way table summarizes data from a sample of patients who were transported to the hospital by helicopter or by ambulance.²²

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

TRM Simpson's Paradox Optional Materials

Optional material, including exercises, on Simpson's paradox can be found in the TE-Book, on the TRFD, or in the Teacher's Resources on the book's digital platform. Although this is an interesting topic, it is not on the AP® Statistics Topic Outline. Exercise 39 can be used to introduce Simpson's paradox to students.

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

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

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

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

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

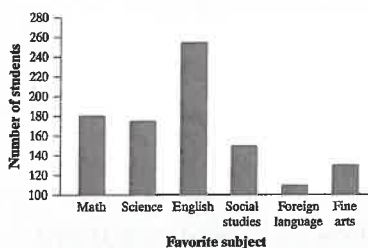
- (c) See if you can explain how the result in part (a) is possible given the result in part (b).

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

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

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

- (a) The distribution of car colors for vehicles purchased in the last month
 (b) The distribution of unemployment percentages for each of the 50 states
 (c) The distribution of favorite sport for a sample of 30 middle school students
 (d) The distribution of shoe type worn by shoppers at a local mall
 (e) The distribution of presidential candidate preference for voters in a state
 41. The following bar graph shows the distribution of favorite subject for a sample of 1000 students. What is the most serious problem with the graph?



- (a) The subjects are not listed in the correct order.
 (b) This distribution should be displayed with a pie chart.
 (c) The vertical axis should show the percent of students.
 (d) The vertical axis should start at 0 rather than 100.
 (e) The foreign language bar should be broken up by language.
 42. The Dallas Mavericks won the NBA championship in the 2010–2011 season. The two-way table displays the relationship between the outcome of each game in the regular season and whether the Mavericks scored at least 100 points.

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

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

- (a) The Mavericks won 57 games and lost only 25 games.
 (b) The Mavericks scored at least 100 points in 47 games and fewer than 100 points in only 35 games.
 (c) The Mavericks won 43 games when scoring at least 100 points and only 14 games when scoring fewer than 100 points.
 (d) The Mavericks won a higher proportion of games when scoring at least 100 points ($43/47$) than when they scored fewer than 100 points ($14/35$).
 (e) The combination of scoring 100 or more points and winning the game occurred more often (43 times) than any other combination of outcomes.

1.39 (a) 64 of the 200 patients transported by helicopter, or 32% of patients died. 260 of the 1100 patients transported by ambulance, or 23.6% of patients died. **(b)** Of the patients who were in serious accidents, 48% who were transported by helicopter died and 60% who were transported by ambulance died. Of the patients who were in less serious accidents, 16% who were transported by helicopter died and 20% who were transported by ambulance died. Whether the patients were in a serious accident or less serious accident, the percentage who died was greater for those who were transported by ambulance. **(c)** Overall, a greater percentage of patients who were transported by helicopter died, but when broken down by seriousness of the accident, in both instances, a greater percentage of patients who were transported by ambulance died. This is because people in more serious accidents were also more likely to be transported by helicopter and were more likely to die. People who were involved in less serious accidents were less likely to be transported by helicopter and were less likely to die. Overall, this makes it appear that those who are transported by helicopters are more likely to die; but when the patients are broken out by seriousness of the accident, we can see that the driving factor for the overall death rates is the seriousness of the accident, not the method of transportation.

1.40 b

1.41 d

1.42 d

1.43 c

1.44 The individuals are the hotels. The variables in the data set are whether there is a pool (categorical), whether there is an exercise room (categorical), Internet cost (\$/day) (quantitative), number of restaurants (quantitative), distance to site (mi) (quantitative), room service availability (categorical), and room rate (\$/day) (quantitative).

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

Dominant hand	Gender		
	Male	Female	Total
	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 (d) 45
 (b) 30 (e) Impossible to determine without more information
 (c) 36

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

PD Section 1.2 Overview

To watch the video overview of Section 1.2, click on the link in the TE-Book, look on the TRFD, or download from the Teacher's Resources on the book's digital platform.

Teaching Tip

Emphasize the switch from categorical (Section 1.1) to quantitative variables (Section 1.2). It is important that students begin to use this distinction to classify the different methods of data analysis.

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.

Preparing for Inference

Be aware of the power of the dotplot. Instructionally, the dotplot can be created very quickly to display class data. More importantly, we will use the dotplot later to display the distribution of a statistic calculated from many samples, which will lead us to an understanding of sampling distributions.