

UNIT 1

Exploring One-Variable Data

PART I Data Analysis

Introduction

1A Statistics: The Language of Variation

1B Displaying and Describing Categorical Data

1C Displaying Quantitative Data with Graphs

1D Describing Quantitative Data with Numbers

UNIT 1, PART I Wrap-Up

FRAPPY!

Review

Review Exercises

AP® Statistics Practice Test



Introduction

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 gasoline prices to elections to weather to the latest technology. These data are trying to tell us a story. To understand what the data are saying, we use **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 informed decisions based on data in your daily life. The following activity illustrates one of the many uses of statistics in the real world.

ACTIVITY

Smelling Parkinson's



Joy Milne, a retired nurse from Scotland, noticed a “subtle musky odor” on her husband Les that she had never encountered before. At first, Joy thought the smell might be from Les’s sweat after long hours of work. But when Les was diagnosed with Parkinson’s disease 6 years later, Joy suspected the odor might be a result of the disease.

Scientists were intrigued by Joy’s claim and designed an experiment to test her ability to “smell Parkinson’s.” Joy was presented with 12 different shirts, each worn by a different person, some of whom had Parkinson’s disease and some of whom did not. The shirts were given to Joy in a random order, and she had to decide whether or not each shirt was worn by a patient with Parkinson’s disease. Joy identified 11 of the 12 shirts correctly.

Although the researchers wanted to believe that Joy could detect Parkinson’s disease by smell, it is possible that she was just a lucky guesser. You and your classmates will perform a simulation with cards or technology to determine which explanation is more believable.



© Chris Watt/Telegraph Media Group Limited 2017

1. Your teacher will hand each pair of students a set of 12 cards (shirts). On the back of some cards is “Parkinson’s” and on the back of others is “No Parkinson’s.” Shuffle the cards thoroughly.
2. Decide who will guess first and have your partner act as the researcher. For each card, guess “Parkinson’s” or “No Parkinson’s.” The researcher will not reveal whether each guess is right or wrong, but will record the number of correct guesses. Now switch roles and repeat the process.
3. Your teacher will draw and label a number line for a class dotplot. Plot the number of correct guesses you made in Step 2 on the graph.
4. Repeat Steps 1 and 2 until you have a total of at least 50 trials of the simulation for your class.
5. How often were 11 or more shirts correctly identified by chance alone? Based on this result, which seems more believable: Joy is just a lucky guesser, or Joy really can smell Parkinson’s disease? Explain your reasoning.

The Smelling Parkinson's activity outlined the steps in the *statistical problem-solving process*: (1) Ask questions, (2) collect data, (3) analyze the data, and (4) interpret the results.¹ Researchers began by identifying the question to be answered: "Can Joy Milne smell Parkinson's disease?" To answer this question, the researchers designed and carried out an appropriate plan to collect data. The resulting data consisted of Joy's 11 correct and 1 incorrect shirt identifications. When analyzing the data, researchers had to consider the possibility that Joy was just guessing and correctly identified 11 of the 12 shirts by chance alone. After careful analysis, the researchers concluded that Joy Milne could actually smell Parkinson's disease. To make the case even stronger, the researchers later discovered that Joy had correctly identified all 12 shirts. Her one "mistake" was a person who was diagnosed with Parkinson's disease a few months later. That's pretty amazing!

In AP[®] Statistics, you will be asked to solve a variety of statistical problems. Your success will depend on steadily developing *skills* in four categories: selecting statistical methods, analyzing data, using probability and simulation, and making statistical arguments.² We'll emphasize both content and skills throughout the book.

SECTION 1A

Statistics: The Language of Variation

LEARNING TARGETS

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

- Identify the individuals and variables in a set of data, and classify the variables as categorical or quantitative.
- Make and interpret a frequency table or a relative frequency table for a distribution of data.

People's opinions vary on all kinds of issues. Data on physical characteristics of people, animals, and plants vary from one individual to another. Chance outcomes — like tosses of a coin or guesses about which shirts were worn by patients with Parkinson's disease — vary. Statistics is the language of variation.

This section starts by showing you how to organize and classify data. Then, you will learn how to represent data in tabular form.

Individuals and Variables

Every year, the U.S. Census Bureau collects data from more than 3.5 million randomly selected U.S. households as part of the American Community Survey (ACS). The table displays some data from 10 households included in the ACS in a recent year.³

Household	Region	People in household	Time in dwelling (years)	Response mode	Household income (\$)	Internet access?
1	South	3	20–29	Phone	272,000	Yes
2	Midwest	3	30+	Internet	54,600	Yes
5	South	1	10–19	Mail	49,900	Yes
6	South	2	2–4	Phone	297,000	Yes
7	Northeast	4	5–9	Internet	130,000	Yes
11	Midwest	3	2–4	Internet	82,000	Yes
14	Midwest	3	10–19	Internet	57,000	Yes
16	West	3	2–4	Mail	36,800	Yes
17	South	2	2–4	Mail	133,000	Yes
25	Midwest	1	20–29	Mail	80,000	No



Bruce Lighty/Photodisc/Getty Images

Most data tables follow this format: each row describes an **individual**, and each column contains 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 a characteristic that can take different values for different individuals.

For the ACS data set, the *individuals* are households. It is important to note that household is *not* a variable in this data set — the numbers in that column of the data table are just labels for the individuals. The *variables* recorded for each household are region, people in household, time in their 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**. People in household and household income (in dollars) are **quantitative variables**.

DEFINITION Categorical variable, Quantitative variable

A **categorical variable** takes values that are labels, which 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 nor measurements of anything. Instead, they are simply labels for a regional location, making zip code a categorical variable. 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.

The variable “year” is often treated as categorical. But it depends on how the data are being used. Consider a data set about cars, in which one of the variables recorded is model year. If we want to know what percentage of cars on the road are 2022 models, we treat year as categorical. If we want to know the average age of cars on the road, we would convert model year to age (in years) and treat this variable as quantitative.

EXAMPLE

Census at school Individuals and variables

Skill 2.A



PROBLEM: Census at School is an international project that collects data about primary and secondary school students using surveys. Since its launch in 2000, students from Australia, Canada, Ireland, Japan, New Zealand, South Africa, the United Kingdom, and the United States have taken part in the project. We selected a random sample of 50 U.S. high school students who completed the survey in a recent year. The table displays data on some of the survey questions for the first 10 students in the sample.



EF Volart/Moment/Getty Images

State	Birth month	Age (years)	Handed	Height (cm)	Home occupants	Allergies	Preferred communication
WI	11	17	Right	175	4	Yes	Internet chat/IM
IN	6	16	Right	175.5	5	No	In person
NY	6	17	Right	157	5	Yes	In person
NC	6	17	Right	169	3	No	Internet chat/IM
MA	6	18	Right	169	3	Yes	Phone call
MO	10	18	Right	170	5	No	Text messaging
PA	5	14	Right	170	6	No	Text messaging
IA	1	17	Left	176	2	No	Text messaging
NC	5	17	Right	175	5	No	Social media
CA	2	17	Right	158	8	Yes	Social media
:	:	:	:	:	:	:	:

- (a) Identify the individuals and variables in this data set.
 (b) Classify each variable as categorical or quantitative.

SOLUTION:

- (a) **Individuals:** 50 randomly selected U.S. high school students who completed the Census at School survey. **Variables:** State, birth month, age (years), handedness, height (cm), number of home occupants, whether the student has allergies, preferred communication method.
- (b) **Categorical:** State, birth month, handedness, whether the student has allergies, preferred communication method. **Quantitative:** Age (years), height (cm), and number of home occupants.

We'll see in Unit 3 why choosing at random, as we did in this example, is a good idea.

Note that birth month is categorical, even though the values listed are numbers.

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 data appropriately on the AP® Statistics 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. Be sure to include any units of measurement for a quantitative variable (like centimeters for height). To make life simpler, we sometimes refer to *categorical data* or *quantitative data* instead of identifying the variable as categorical or quantitative.

Summarizing Data with Tables

A variable generally takes values that vary from one individual to another. That's why we call it a variable! The **distribution** of a variable describes the pattern of variation of the values.

DEFINITION Distribution

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

We can summarize a variable's distribution with a **frequency table** or a **relative frequency table**. To make either kind of table, start by tallying the number of times that the variable takes each value.

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 percentage of individuals having each value.

We can use a frequency table or a relative frequency table to help describe the distribution of a variable, as the following example illustrates.

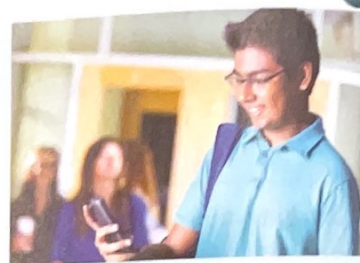
Skills 2.A, 2.B

EXAMPLE

Call me, maybe? Summarizing data with tables

PROBLEM: Here are the data on preferred method of communicating with friends for all 50 students in the Census at School sample:

Internet chat/IM	In person	In person	Internet chat/IM	Phone call
Text messaging	Text messaging	Text messaging	Social media	Social media
Text messaging	Text messaging	Text messaging	In person	In person
Text messaging	Text messaging	Text messaging	In person	In person
Text messaging	Text messaging	Internet chat/IM	Social media	Text messaging
Text messaging	Text messaging	In person	Text messaging	Internet chat/IM
Text messaging	Social media	Text messaging	Social media	Text messaging
Text messaging	In person	Social media	In person	Text messaging
Text messaging	Text messaging	Text messaging	Internet chat/IM	In person
Text messaging	In person	Internet chat/IM	Text messaging	In person



SDI Productions/E+/Getty Images

- (a) Make a frequency table and a relative frequency table to summarize the distribution of preferred communication method.
- (b) Do these data support the claim that a majority of students prefer to communicate with their friends using text messaging? Justify your answer.

SOLUTION:**(a) Frequency table**

Preferred method	Frequency
In person	12
Internet chat/IM	6
Phone call	1
Social media	6
Text messaging	25
Total	50

The frequency table shows the *number* of students who prefer each communication method. To create the frequency table, count how many students said "In person," how many said "Internet chat or instant messenger," and so on.

Relative frequency table

Preferred method	Relative frequency
In person	$12/50 = 0.24$ or 24%
Internet chat/IM	$6/50 = 0.12$ or 12%
Phone call	$1/50 = 0.02$ or 2%
Social media	$6/50 = 0.12$ or 12%
Text messaging	$25/50 = 0.50$ or 50%
Total	$50/50 = 1.00$ or 100%

The relative frequency table shows the *proportion* or *percentage* of students who prefer each communication method. Note that in statistics, a proportion is a value between 0 and 1 that is equivalent to a percentage.

- (b) No. Exactly 50% of the students in the sample said that they prefer to communicate with their friends via text messaging, but a majority is more than half.

FOR PRACTICE, TRY EXERCISE 5



Note that the frequencies and relative frequencies listed in these tables are not data. The frequency and relative frequency tables summarize the data by telling us how many, or what proportion or percentage of, students in the Census at School sample prefer each method of communicating with friends.

The same process can be used to summarize the distribution of a quantitative variable. However, it would be hard to make a frequency table or a relative frequency table for quantitative data that take many different values, like height (cm) in the Census at School data set. We'll look at a better option for summarizing quantitative variables in Section 1C.

FROM DATA ANALYSIS TO INFERENCE

Sometimes we're interested in drawing conclusions that go beyond the data at hand. That's the idea of *statistical inference*. In the Census at School survey, 12 of the 50 randomly selected students prefer communicating in person with their friends. That's 24% of the *sample*. Can we conclude that exactly 24% of the *population* of students who completed the online survey prefer communicating in person with their friends? No.

If another random sample of 50 students who completed the survey were selected, the percentage who prefer communicating in person with their friends would probably not be exactly 24%. Can we at least say that the actual population value is “close” to 24%? As you will learn in future units, that depends on what we mean by “close.”

Our ability to do statistical inference is determined by how the data are produced. Unit 3 discusses the two main methods of collecting data — sampling and experiments — and the types of conclusions that can be drawn from each. As the Smelling Parkinson's activity at the beginning of this unit illustrates, the logic of inference rests on asking, “What are the chances?” *Probability*, the study of chance behavior, is the focus of Units 4 and 5. We'll introduce the most common methods of statistical inference in Units 6–9.



CHECK YOUR UNDERSTANDING



Malena is a car buff who wants to find out more about the cars that high school students drive. The principal of a local high school gives Malena permission to go to the student parking lot and record some data. Later, Malena does some internet research on each model of car in the parking lot and makes a spreadsheet that includes each car's license plate, model, year, number of stickers on the car, color, weight (in kilograms), whether it has a navigation system, and highway gas mileage.

1. Identify the individuals and variables in Malena's study.
2. Classify each variable as categorical or quantitative.

A professor suspects that most students in an 8 A.M. introductory statistics class did not complete the pre-class reading assignment. To find out, the professor gave the students a 10-question pop quiz about the assigned reading at the beginning of class. Here are the number of correct answers on the pop quiz for the 50 students in the class:

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

3. Make a frequency table and a relative frequency table to summarize the distribution of number of correct answers on the pop quiz.
4. What proportion of students got fewer than 7 correct answers on the pop quiz? Does this result support the professor's belief that most students did not complete the pre-class reading assignment? Justify your answer.

SECTION 1A

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 is a characteristic that can take different values for different individuals.
- A **categorical variable** takes values that are labels, which place each individual into a particular group, called a category. A **quantitative variable** takes numerical values that count or measure some characteristic of each individual.
- The **distribution** of a variable describes what values the variable takes and how often it takes each value.
- To summarize the distribution of a variable, you can use a **frequency table** that shows the number of individuals having each value or a **relative frequency table** that shows the proportion or percentage of individuals having each value. You can use a frequency table or a relative frequency table to help describe the distribution of a variable.

AP® EXAM TIP
AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topics 1.1–1.3, which are available in AP® Classroom.

SECTION 1A

Exercises

The solutions to all exercises numbered in red can be found in the Solutions Appendix.

Individuals and Variables

1. A class survey Here is a small part of a data set that describes the students in a math class. The data come from anonymous responses to a questionnaire filled out on the second day of class.

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

- (a) Identify the individuals and variables in this data set.
 (b) Classify each variable 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 from around the world.⁴

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (sec)
Copperhead Strike	Steel	82.0	Sit down	50.0	144
Eurostar	Steel	98.9	Inverted	50.2	140
Jungle Trailblazer	Wood	108.3	Sit down	54.1	150
Falcon	Steel	197.5	Wing	73.3	156
Olympia Looping	Steel	106.7	Sit down	49.7	105
Time Traveler	Steel	100.0	Sit down	50.3	117

- (a) Identify the individuals and variables in this data set.
 (b) Classify each variable as categorical or quantitative.
3. **Hit movies** *Avatar* was the top-earning movie based on box-office receipts worldwide as of December 2022. The following table displays data on several popular movies.⁵ Identify the individuals and variables in this data set. Then classify each variable as categorical or quantitative.

Movie	Year	Rating	Time (min)	Genre	Box office (\$)
<i>Avatar</i>	2009	PG-13	162	Action	2,910,284,102
<i>Avengers: Endgame</i>	2019	PG-13	181	Action	2,797,732,053
<i>Titanic</i>	1997	PG-13	194	Drama	2,207,986,545
<i>Star Wars: The Force Awakens</i>	2015	PG-13	136	Adventure	2,064,615,817
<i>Avengers: Infinity War</i>	2018	PG-13	156	Action	2,048,359,754
<i>Spider Man: No Way Home</i>	2021	PG-13	148	Action	1,910,041,582
<i>Jurassic World</i>	2015	PG-13	124	Action	1,669,963,641
<i>The Lion King</i>	2019	PG	118	Adventure	1,647,733,638
<i>The Avengers</i>	2012	PG-13	143	Action	1,515,100,211
<i>Furious 7</i>	2015	PG-13	137	Action	1,514,553,486
<i>Top Gun: Maverick</i>	2022	PG-13	131	Action	1,488,187,968
<i>Frozen II</i>	2019	PG	103	Adventure	1,444,735,491

4. **Skyscrapers** Here is some information about the tallest buildings in the world as of 2023.⁶ Identify the individuals and variables in this data set. Then classify each variable as categorical or quantitative.

Building	Country	Height (m)	Floors	Use	Year completed
Burj Khalifa	United Arab Emirates	828.0	163	Mixed	2010
Merdeka 118	Malaysia	678.9	118	Mixed	2023
Shanghai Tower	China	632.0	128	Mixed	2015
Makkah Royal Clock Tower Hotel	Saudi Arabia	601.0	120	Hotel	2012
Ping An Finance Center	China	599.1	115	Mixed	2017
Lotte World Tower	South Korea	554.5	123	Mixed	2017
One World Trade Center	United States	541.3	94	Office	2014
Guangzhou CTF Finance Center	China	530.0	111	Mixed	2016
Tianjin CTF Finance Center	China	530.0	97	Mixed	2019
China Zun Tower	China	527.7	101	Mixed	2019

Summarizing Data with Tables

5.  **New York squirrels** How do squirrels in New York City's Central Park respond to humans? The Squirrel

Census project enlisted volunteers to record data on all squirrels observed in Central Park in a recent year. One of the variables measures how the squirrels responded to humans. We selected a random sample of 75 of the more than 3000 squirrels observed. Here are the data on their reactions to humans:⁷

Approach	Run from	Indifferent	Indifferent	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Approach	Indifferent	Approach	Run from	Run from
Indifferent	Indifferent	Indifferent	Run from	Indifferent
Indifferent	Indifferent	Run from	Indifferent	Indifferent
Indifferent	Run from	Run from	Indifferent	Indifferent
Run from	Run from	Approach	Approach	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Indifferent	Run from	Approach	Approach	Indifferent
Indifferent	Indifferent	Run from	Approach	Indifferent
Indifferent	Approach	Indifferent	Run from	Approach
Run from	Approach	Run from	Indifferent	Indifferent
Indifferent	Approach	Indifferent	Run from	Indifferent
Indifferent	Indifferent	Approach	Approach	Indifferent

- (a) Make a frequency table and a relative frequency table to summarize the distribution of response to humans.
- (b) Were these Central Park squirrels more likely to approach humans, run from humans, or be indifferent to humans? Justify your answer.
6. **Disc dogs** Here is a list of the breeds of dogs that won the World Canine Disc Championships over a 45-year period.⁸

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

- (a) Make a frequency table and a relative frequency table to summarize the distribution of dog breed among World Canine Disc Championship winners.
- (b) Which breeds of dog were the most frequent winners? Justify your answer.

7. **CubeSat missions** A CubeSat is a miniature satellite used for space research that can be easily deployed from a launch vehicle or from the International Space Station. More than 1000 CubeSats have been launched since specifications for these satellites were jointly developed by Cal Poly San Luis Obispo and Stanford University in 1999. The relative frequency table summarizes data on the type of mission for each CubeSat launched in November and December 2019.⁹

Mission type	Proportion
Science	0.18
Technology	0.28
Imaging	0.34
Communications	0.12
Education	0.08

- (a) There were a total of 50 CubeSats launched during these two months. How many had a communications mission type?
- (b) A report states that a majority of these CubeSat launches had an imaging mission type. Explain why this statement is incorrect.

Multiple Choice Select the best answer for each question.

Exercises 8 and 9 refer to the following setting. A realtor in Hilton Head, South Carolina, collects data on all of the

homes sold in the town during 2023. The resulting data set lists each home's address, along with the following information: zip code, number of bedrooms, primary building material (wood, brick, etc.), square footage, whether it has a pool, age of home, and sales price.

8. The individuals in this data set are
- (A) the Hilton Head realtor who collected the data.
 - (B) all Hilton Head homeowners who sold their homes during 2023.
 - (C) the addresses of all homes sold in Hilton Head during 2023.
 - (D) all homes sold in Hilton Head during 2023.
 - (E) all homes in Hilton Head during 2023.
9. This data set contains
- (A) 7 variables, 2 of which are categorical.
 - (B) 7 variables, 3 of which are categorical.
 - (C) 7 variables, 4 of which are categorical.
 - (D) 8 variables, 3 of which are categorical.
 - (E) 8 variables, 4 of which are categorical.

SECTION 1B

Displaying and Describing Categorical Data

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

- Make and interpret bar graphs of categorical data.
- Compare distributions of categorical data.
- Identify what makes some graphs of categorical data misleading.

In this section, you will learn how to display and describe the distribution of a categorical variable. Sections 1C, 1D, 1E, and 1F focus on displaying and describing the distribution of a quantitative variable. Unit 2 examines relationships between two categorical or two quantitative variables. This process of exploratory data analysis is known as *descriptive statistics*.

Displaying Categorical Data: Bar Graphs

A frequency table or relative frequency table summarizes a variable's distribution with numbers. To display the distribution more clearly, use a graph. The most

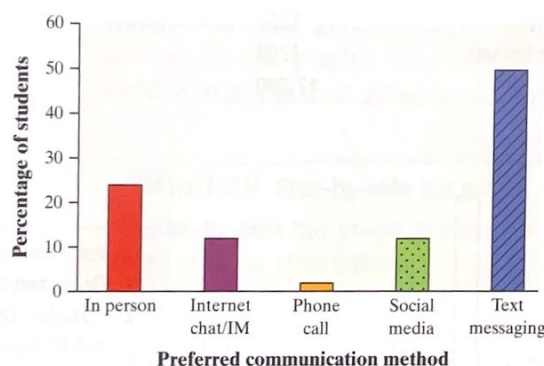
common way to display categorical data is with a **bar graph** (also known as a *bar chart*).

DEFINITION Bar graph

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

Figure 1.1 shows a bar graph of the data on preferred communication method for the random sample of 50 U.S. high school students who completed the Census at School survey. Note that the percentages for each category come from the relative frequency table.

FIGURE 1.1 Bar graph of the distribution of preferred communication method for a random sample of 50 U.S. high school students.



Relative frequency table

Preferred method	Relative frequency
In person	$12/50 = 0.24$ or 24%
Internet chat/IM	$6/50 = 0.12$ or 12%
Phone call	$1/50 = 0.02$ or 2%
Social media	$6/50 = 0.12$ or 12%
Text messaging	$25/50 = 0.50$ or 50%

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

HOW TO MAKE A BAR GRAPH

1. **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 (percentage or proportion) of individuals in each category.
2. **"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 equal or exceed the largest frequency or relative frequency in any category.
3. **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 real purpose of a graph is to help us interpret the data. When you look at a graph, always ask, "What do I see?" In Figure 1.1, the bar graph reveals that the most preferred method of communicating with friends for these high school students is text messaging (50%). The next most preferred method is talking with their friends in person (24%). Social media and internet chat/instant messenger are somewhat less popular (12% each) methods of communicating with friends. By far the least preferred method of communication for these students is a phone call (2%).

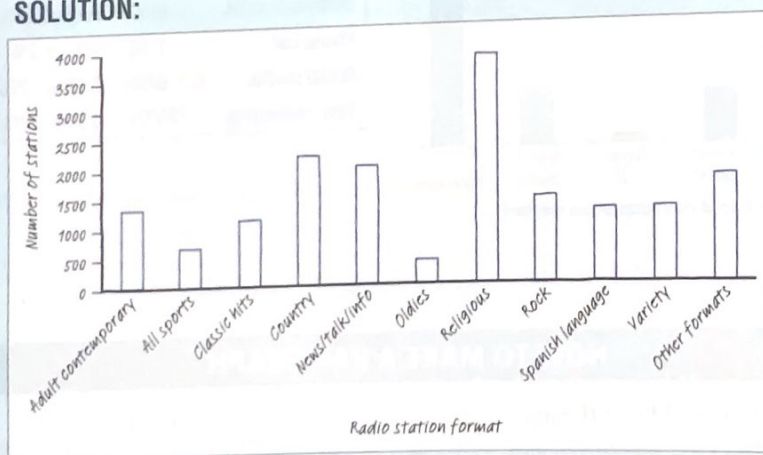
EXAMPLE**What's on the radio?****Displaying categorical data: Bar graphs**

PROBLEM: Nielsen Audio, the rating service for radio audiences, places U.S. radio stations into categories that describe the kinds of programs they broadcast. The frequency table summarizes the distribution of station format in a recent year.¹⁰ Make a frequency bar graph to display the data. Describe what you see.

Format	Number of stations	Format	Number of stations
Adult contemporary	1357	Religious	3837
All sports	669	Rock	1466
Classic hits	1140	Spanish language	1228
Country	2200	Variety	1257
News/talk/information	2002	Other formats	1769
Oldies	405	Total	17,330



ACORN 1/Alamy Stock Photo

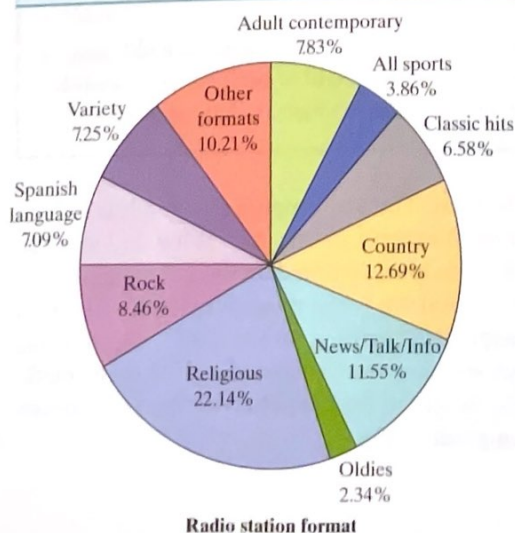
SOLUTION:

To make the bar graph:

1. **Draw and label the axes.**
2. **"Scale" the axes.** The largest frequency is 3837, so we choose a vertical scale from 0 to 4000, with tick marks 500 units apart.
3. **Draw bars above the category names.**

On U.S. radio stations, the most common formats are Religious (3837), Country (2200), and News/Talk/Information (2002), while the least common formats are Oldies (405) and All Sports (669). Moderately common formats offered by a similar number of stations include Classic Hits (1140), Spanish Language (1228), Variety (1257), Adult Contemporary (1357), and Rock (1466). There are 1769 stations with Other formats.

FOR PRACTICE, TRY EXERCISE 1

**PIE CHARTS**

Here is a **pie chart** of the radio station format data from the preceding example. Notice that this graph shows the *relative frequency* for each response category. For instance, the "Spanish language" slice makes up 7.09% of the graph because the relative frequency for this category is $1228/17,330 = 0.0709 = 7.09\%$.

DEFINITION Pie chart

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.

Use a pie chart when you want to emphasize each category's relation to the whole. Each slice of the pie shows the count or percentage of individuals in that category. Pie charts are challenging to make by hand, but technology will do the job for you.



A bar graph or pie chart that displays a distribution of categorical data must include **all individuals in the data set**. This might require including an “Other” category, as in the radio station example. Bar graphs are more flexible than pie charts. Both types of graphs can display the distribution of a categorical variable, but a bar graph can also compare any set of quantities that are measured in the same units.

Comparing Distributions of Categorical Data

You can use a bar graph to display the distribution of a categorical variable. A **side-by-side bar graph** can be used to compare the distribution of a categorical variable in two or more groups.

DEFINITION Side-by-side bar graph

A **side-by-side bar graph** displays the distribution of one categorical variable in each of two or more groups.

It's a good idea to use relative frequencies when comparing data for multiple groups, especially if the groups have different sizes.

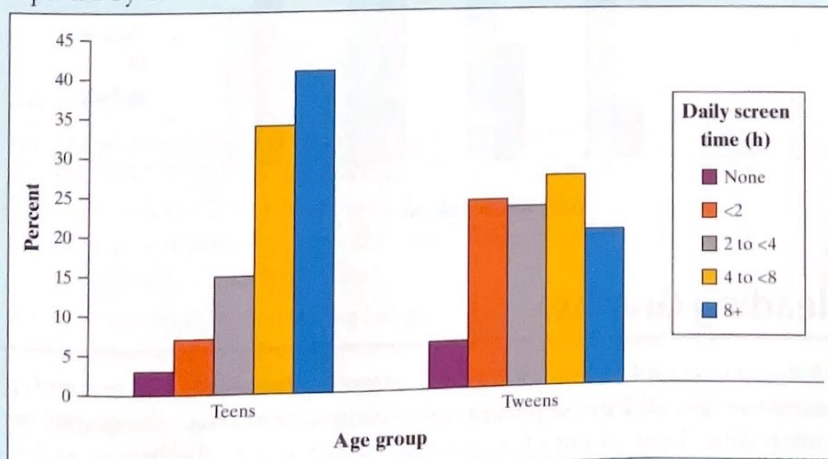
EXAMPLE

Screen time for teens and tweens Comparing distributions of categorical data

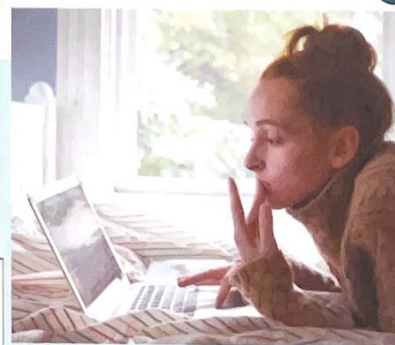
Skill 2.D



PROBLEM: How much time do tweens (ages 8–12) and teens (ages 13–18) spend using digital devices each day? Researchers surveyed a random sample of more than 1300 U.S. 8- to 18-year-olds to find out. The side-by-side bar graph summarizes the data on daily screen time reported by the 560 tweens and 746 teens in the sample.¹¹



Compare the distributions of daily screen time for teens and tweens.



Melanie Acevedo/DigitalVision/Getty Images

SOLUTION:

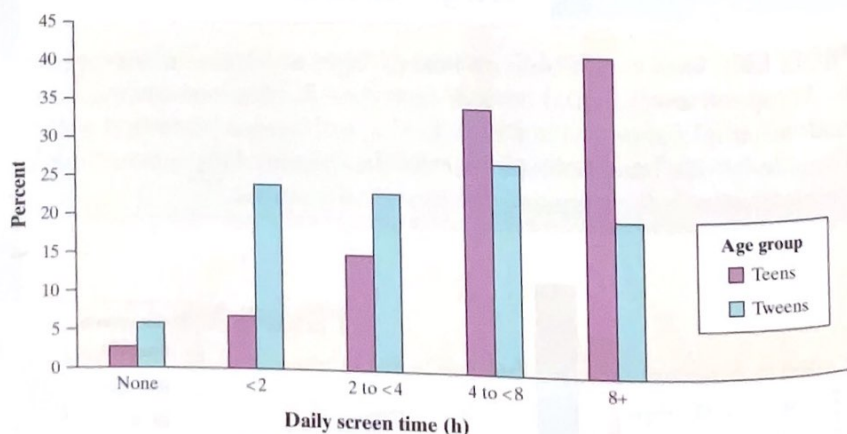
Tweens tend to report less daily screen time than teens. Tweens are more likely than teens to spend no time (6% versus 3%), less than 2 hours (24% versus 7%), and from 2 to less than 4 hours (23% versus 15%) on digital devices per day. Teens are more likely than tweens to spend from 4 to less than 8 hours (34% versus 27%) and 8 or more hours (41% versus 20%) on digital devices per day. More than half (53%) of tweens report less than 4 hours of screen time per day, while 75% of teens report 4 or more hours per day of screen time.

FOR PRACTICE, TRY EXERCISE 7

The following relative frequency table summarizes the data on daily screen time from the example by age group. As a result, the percentages in each row add to 100%. Note that we could have also compared the distributions with this relative frequency table.

	Daily time spent on digital devices				
	None	< 2 hours	2 to < 4 hours	4 to < 8 hours	8+ hours
Teens	3%	7%	15%	34%	41%
Tweens	6%	24%	23%	27%	20%

In the example, we grouped the bars by age group in the side-by-side bar graph. This arrangement clearly shows the two distributions of daily screen time — one for the teens and one for the tweens in the sample. It is also possible to group the bars by daily screen time, as in the following graph. This arrangement makes category-by-category comparisons for the daily time spent on digital devices easier — but now it's harder to see the individual distributions of daily screen time for the teens and for the tweens. Whichever way you organize the bars, be sure to include a key that describes what each color or type of shading on the side-by-side bar graph represents.



Misleading Graphs

Bar graphs can be a bit dull to look at. It is tempting to replace the bars with pictures or to use special three-dimensional (3D) 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? Misleading graphs

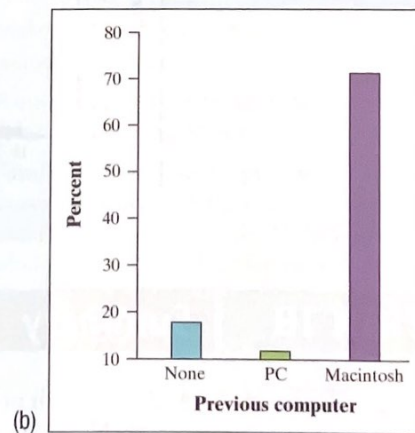
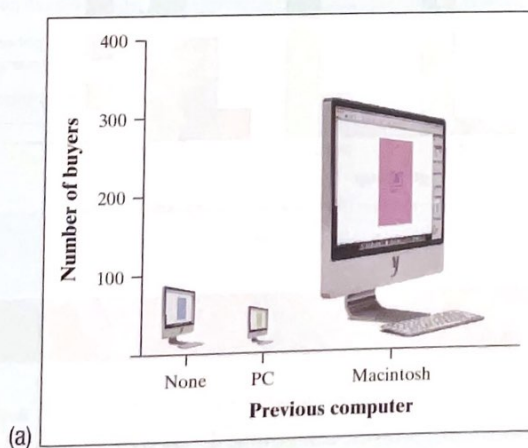
Skill 2.A



PROBLEM: When Apple, Inc., first introduced the iMac in 1998, the company wanted to know whether this new computer was expanding its market share. (The iMac has enjoyed great success ever since!) 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 randomly selected iMac customers. The firm categorized each customer as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey data.¹²

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

- (a) A graph of the data that uses pictures instead of the more traditional bars is shown in figure (a). How is this pictograph misleading?
- (b) A bar graph of the data is shown in figure (b). Explain why this graph could be considered deceptive.



SOLUTION:

- (a) The pictograph is misleading because the areas of the computers make it look like the number of iMac buyers who are former Mac owners (355) is at least 10 times as large as the number of buyers in either of the other two categories (None: 85, PC owner: 60), which isn't true.
- (b) The bar graph is misleading because starting the vertical scale at 10 instead of 0 makes it look like the percentage of iMac buyers who previously owned a PC (12.0%) is less than half the percentage who are first-time computer buyers (17.0%), 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 11



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



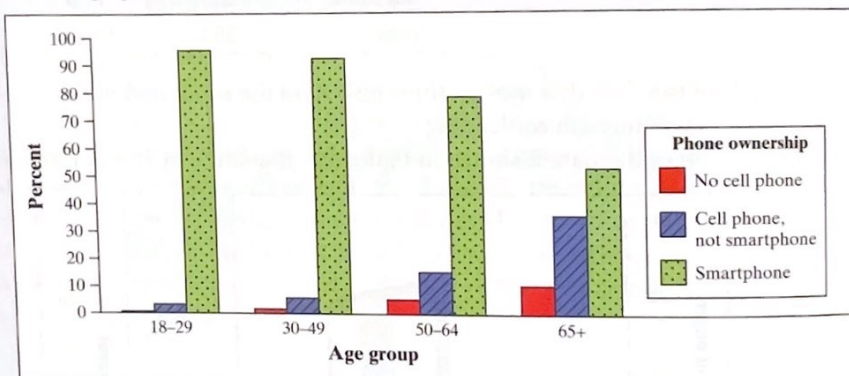
CHECK YOUR UNDERSTANDING



The Pew Research Center asked a random sample of 1502 U.S. adults about their cell phone ownership. The frequency table summarizes their responses.¹³

Type of cell phone	Frequency
None	73
Cell phone, not smartphone	212
Smartphone	1217
Total	1502

1. Make a relative frequency bar graph to display the distribution of cell phone ownership for the 1502 people in the sample. Describe what you see.
2. The following graph displays the distribution of cell phone ownership for each of four age groups in the sample. Compare the distributions.



SECTION 1B

Summary

AP® EXAM TIP AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topic 1.4, which are available in AP® Classroom.

- A bar graph or pie chart can be used to display the distribution of a categorical variable. When examining any graph, ask yourself, “What do I see?”
- You can use a **side-by-side bar graph** to compare the distribution of one categorical variable for two or more groups. Be sure to use relative frequencies when comparing groups of different sizes.
- 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.

SECTION 1B

Exercises

Displaying Categorical Data: Bar Graphs

1. **Birth days** The frequency table summarizes data on the numbers of babies born on each day of a single week in the United States.¹⁴

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

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

2. **Going up?** The website fivethirtyeight.com published an interesting article about the more than 75,000 elevators in New York City. The frequency table summarizes data on the number of elevators of each type at that time.¹⁵

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

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

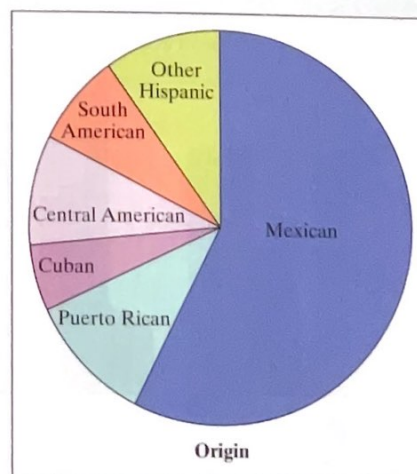
3. **Cool car colors** The popularity of colors for cars and light trucks changes over time. Silver passed green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. The relative frequency table summarizes data on the colors of vehicles sold worldwide in 2020.¹⁶

Color	Percentage of vehicles
Black	19
Blue	7
Brown	3
Gray	15
Green	1
Red	5
Silver	9
White	38
Yellow	2
Other	??

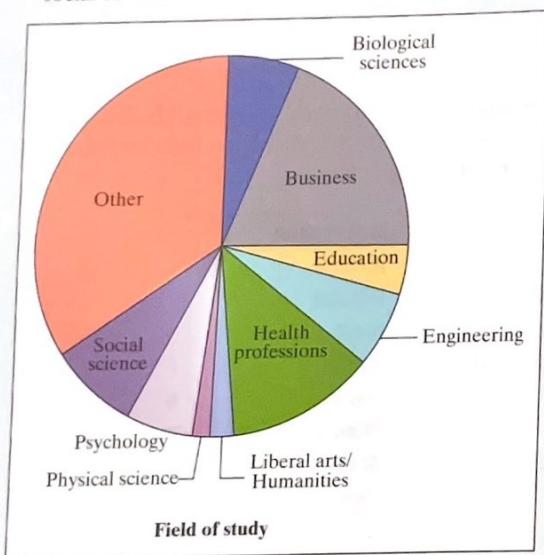
- (a) What percentage of vehicles would fall in the “Other” category?
- (b) Make a bar graph to display the data. Describe what you see.
- (c) Would it be appropriate to make a pie chart of these data? Explain your answer.
4. **Spam** Email spam is very annoying. The relative frequency table summarizes data on the most common types of spam.¹⁷

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

- (a) What percentage 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 your answer.
5. **Family origins** Here is a pie chart of U.S. Census Bureau data showing the origin of more than 62 million Hispanic people in the United States.¹⁸ About what percentage of Hispanic people are of Mexican origin? Puerto Rican origin?

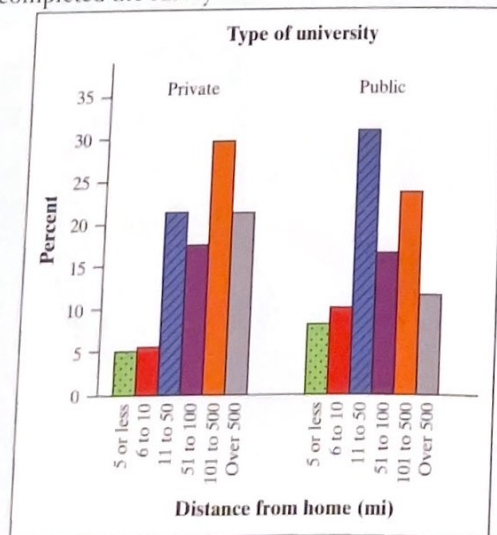


6. **Which major?** More than 2 million students earn bachelor's degrees in U.S. colleges and universities each year. The pie chart displays data on bachelor's degrees earned by field of study in a recent year.¹⁹ About what percentage of students earned their bachelor's degrees in business? In health professions? In social science?

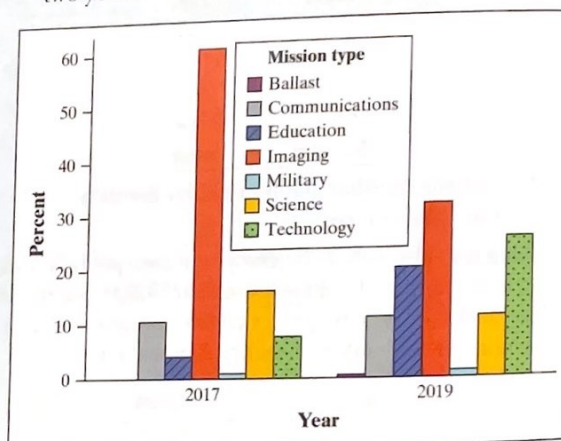


Comparing Distributions of Categorical Data

7. **Far from home** A survey asked first-year college students, "How many miles is this college from your permanent home?" Students selected 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.²⁰ Compare the distributions of distance from home for students from private and public 4-year colleges who completed the survey.



8. **More CubeSat missions** A CubeSat is a miniature satellite used for space research that can be easily deployed from a launch vehicle or from the International Space Station. More than 1000 CubeSats have been launched since specifications for these satellites were jointly developed by Cal Poly San Luis Obispo and Stanford University in 1999. The side-by-side bar graph displays data on the types of missions undertaken by CubeSats launched in 2017 and 2019. Compare the distributions of mission type for these two years.²¹



9. **Reading for pleasure** Researchers surveyed a random sample of more than 1300 U.S. tweens (ages 8–12) and teens (ages 13–18) about how often they read for pleasure. The following table summarizes the responses in each age group.²²

	Age group	
	Teens	Tweens
Every day	21%	34%
At least once a week	27%	29%
At least once a month	14%	12%
Less than once a month	20%	12%
Never	18%	13%

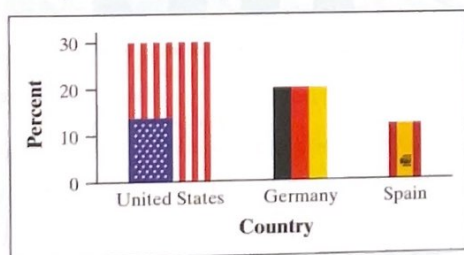
- (a) Make a side-by-side bar graph to compare the distributions of how often teens and tweens read for pleasure.
- (b) Describe similarities and differences between the two distributions.
10. **Popular car colors** Favorite car colors may differ among countries in different parts of the world. The following table summarizes data on the most popular car colors in a recent year for the countries in North America and Asia.²³

	Continent	
	North America	Asia
White	27%	45%
Gray	21%	13%
Black	20%	18%
Blue	10%	5%
Silver	10%	7%
Red	8%	4%
Brown	2%	3%
Green	1%	2%
Yellow	0.5%	2%
Other	0.5%	1%

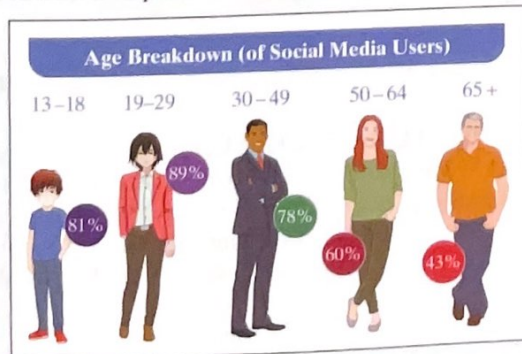
- (a) Make a side-by-side bar graph to compare the distributions of most popular car color in North America and Asia.
- (b) Describe similarities and differences between the two distributions.

Misleading Graphs

11. **Who dislikes shopping?** Harris Interactive asked adults from several countries if they like or dislike shopping for clothes. The following pictograph displays the percentage of adults in the United States, Germany, and Spain who said that they dislike shopping for clothes. Explain how this graph is misleading.²⁴

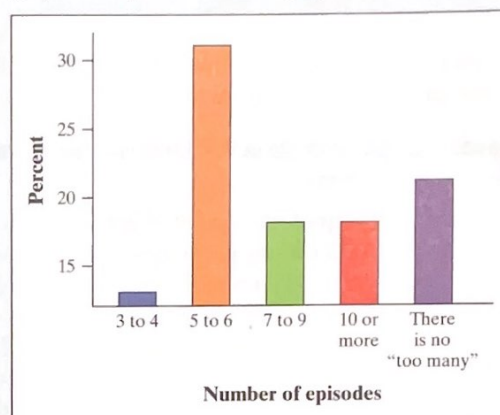


12. **Social media** The Pew Research Center surveyed a random sample of U.S. teens and adults about their use of social media. The following pictograph displays the percentage of people in various age groups who report using social media. Explain how this graph is misleading.

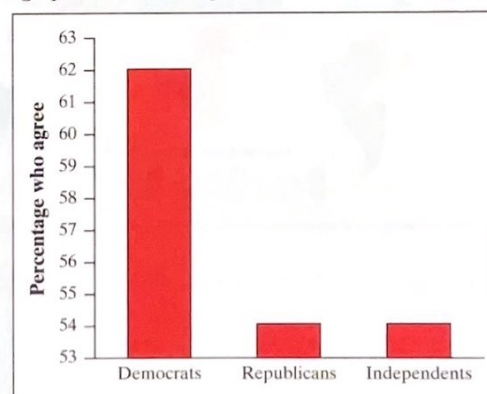


13. **Binge-watching** Do you “binge-watch” television series by viewing multiple episodes of a series at one sitting? A survey of 800 people who binge-watch were asked how many episodes is too many to watch in one

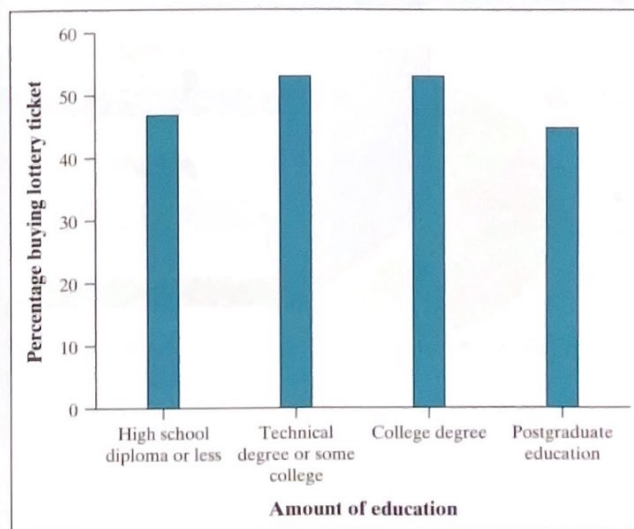
viewing session. The results are displayed in the bar graph.²⁵ Explain how this graph is misleading.



14. **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 shown here. Explain how this graph is misleading. (Note: When notified about the misleading nature of its graph, the network posted a corrected version.)



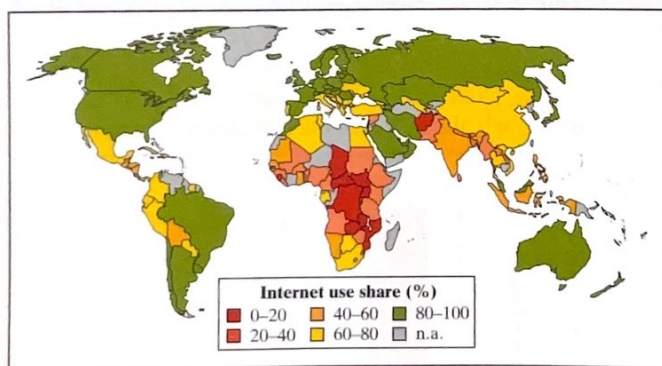
15. **Lotteries and education** A Gallup Poll asked respondents about their highest level of education and whether they had bought a state lottery ticket in the last 12 months.²⁶ Here is a bar graph of the data:



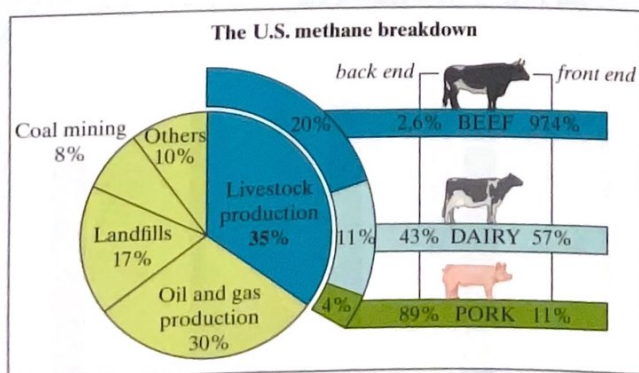
- (a) Describe what this graph reveals about lottery ticket buying habits among the different education groups.
- (b) Would it be appropriate to make a pie chart for this data set? Explain your answer.

For Investigation Apply the skills from the section in a new context or nonroutine way.

16. **Choropleth maps** A *choropleth map* is a graphical representation of data by geographic region in which values are depicted by color. For instance, the choropleth map presented here shows the percentage of people who were internet users in each country in 2020.²⁷ Write a thorough analysis of what the graph shows. Be sure to include a comparison of internet use in different regions of the world.

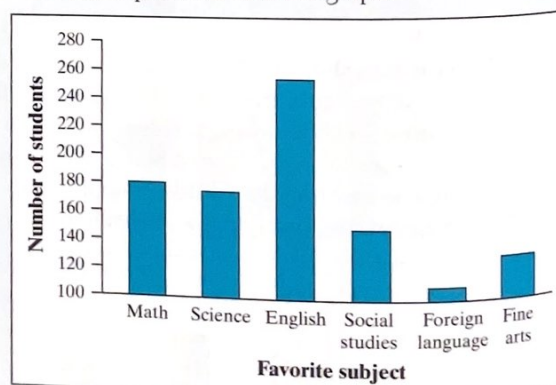


17. **Cow pie chart?** Methane is a greenhouse gas that may contribute to climate change. The modified pie chart shown here displays data on the sources of methane emissions in the United States.²⁸ Write a thorough analysis of what the graph shows. Be sure to include a comparison of methane emissions by beef cows, dairy cows, and pigs.



Multiple Choice Select the best answer for each question.

18. 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
19. 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 percentage of students.
- (D) The vertical axis should start at 0 rather than 100.
- (E) The foreign language bar should be broken up by language.

Recycle and Review Practice what you learned in previous sections.

20. **Hotels (1A)** 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.

Table for Exercise #20

Hotel	Pool	Exercise room?	Internet cost (\$/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 only	114
Hilton	In	Y	10.00	2	0.1	Y	156
Marriott	In	Y	9.95	2	0.0	Y	145

SECTION 1C

Displaying Quantitative Data with Graphs

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

- Make and interpret dotplots of quantitative data.
- Describe the shape of a distribution of quantitative data.
- Describe the distribution of a quantitative variable.
- Compare distributions of quantitative data.
- Make and interpret stemplots of quantitative data.
- Make and interpret histograms of quantitative data.

As you learned in the previous section, bar graphs are typically used to display the distribution of a categorical variable. In this section, you will learn to make and interpret several types of graphs that can be used to display the distribution of a quantitative variable: *dotplots*, *stemplots*, and *histograms*.

There are two types of quantitative variables: *discrete* and *continuous*. Most **discrete variables** result from counting something, like the number of people in a household or the number of lottery tickets a person buys until they win the jackpot. Note that the number of possible values of a discrete variable can be finite or infinite. **Continuous variables** typically result from measuring something, like height (in inches) or time to run a 100-meter dash (in seconds). Note that there are infinitely many possible values for a continuous variable. Age is technically a continuous variable — a high school student's age might be 17.30162... years. But it is often treated as a discrete variable — for example, age = 17 years.

DEFINITION Discrete variable, Continuous variable

A quantitative variable that takes a countable set of possible values with gaps between them on the number line is a **discrete variable**.

A quantitative variable that can take any value in an interval on the number line is a **continuous variable**.