AP Statistics Summer Assignment

I will be teaching using a TI-84 Plus Graphing Calculator. You must have your own TI-84+ or equivalent in class with you every day (make sure that I approve your calculator). It is our software and you cannot do the work without it.

A. Practice using your calculator with the Tech Corner on pages 40-41 in the attached reading.

B. Read the notes and examples in each of the 3 sections. Complete the following problems:

PG 10-13: #1-21

PG 22-27: #1-26

PG 43-48: #1-31

This is due on the first Day in September. If Absent, email me your answers on or prior to the first day of school.

tsanchez@nutleyschools.org

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

Applet

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.[1]

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 using the *Can You Smell Parkinson's?* applet at **stapplet.com**) to determine which explanation is more believable. Here are the directions for the simulation with cards:

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 the process 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 outlines the steps in the **statistical problem-solving process.**[2]

DEFINITION Statistical problem-solving process

- **Formulate questions:** Clarify the research problem and ask one or more valid statistical investigative questions.
- **Collect data:** Design and carry out a plan to collect appropriate data.
- **Analyze data:** Use appropriate graphical and numerical methods to analyze the data.
- **Interpret results:** Draw conclusions based on the data analysis. Be sure to answer the investigative question(s)!

Researchers began by identifying the *investigative question* to be answered: "Can Joy Milne correctly identify Parkinson's disease status by smell for more than 50% of all shirts like the ones in this experiment?" To answer this question, the researchers designed and conducted a study to gather appropriate 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 mastering the course content and on developing *skills* related to the four *statistical practices* highlighted in the statistical problem-solving process: formulate questions, collect data, analyze data, and interpret results. See the table near the inside front cover for a listing of the statistical practices and skills in AP® Statistics.[3] We'll emphasize content as well as the statistical practices and skills throughout this book.

SECTION 1A Statistics: Learning from Data

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

- Determine a valid investigative question in a statistical study.
- Identify the population and sample in a statistical study.
- Identify the observational units and variables in a statistical study or data set, and classify the variables as categorical or quantitative.

AP® EXAM TIP: AP® Classroom
Preview the content of this section with the resources in AP® Classroom for Topics 1.1 and 1.2.

This section begins with a closer look at the first step in the statistical problem-solving process: formulate questions. Next, you will learn some basics about gathering and considering data. Unit 1, Part II provides additional details about the second step in the statistical problem-solving process: collect data.

Determining Investigative Questions

A statistical study starts with an investigative question. But not just any question will do. Unlike most mathematical questions, *a valid statistical investigative question is based on data that vary*. For instance, "Can Joy Milne correctly identify Parkinson's disease status by smell for more than 50% of all shirts like the ones in this experiment?" is a valid investigative question because Joy may identify some shirts correctly and other shirts incorrectly. However, "How much screen time did you have yesterday?" is not a valid investigative question because the question can be answered with a single data value.

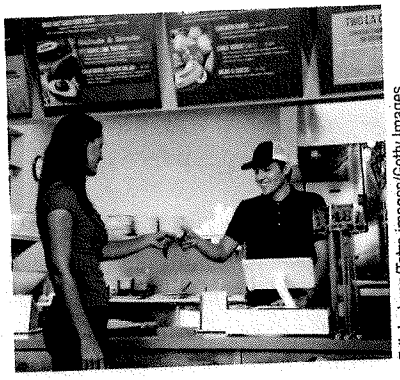
We'll examine the components of statistical investigative questions in more detail in Section 1F. For now, you should be able to determine a valid investigative question in a statistical study.

Skill 1.A

EXAMPLE

Out to lunch
Determining investigative questions

PROBLEM: Two friends go off campus every day to eat lunch. Because the lunch period is short, they wonder whether it would be faster to order inside their favorite fast-food restaurant or to use the drive-thru. Each day, they flip a coin to determine which method (inside or drive-thru) to use and record the total length of time it takes from the moment they enter the parking lot to the moment they receive their food. After several weeks of collecting data, they analyze their results and determine that ordering inside took about 2 minutes and 34 seconds less than using the drive-thru, on average. Their conclusion: it's faster to go inside.



Erik Isakson/Tetra images/Getty Images

Determine the investigative question in this statistical study.

SOLUTION:

During all possible lunch visits to this restaurant, is the average time to receive food when ordering inside different than the average time to receive food when ordering at the drive-thru?

This is a valid investigative question because the length of time it takes these friends to enter the parking lot, place their order, and receive their food will vary from day to day.

FOR PRACTICE, TRY EXERCISE 3

Every statistical study or data set involves a real-world context. On the AP® Statistics exam, you will be expected to answer questions "in context."

Populations and Samples

Suppose we want to find out what percentage of young drivers in the United States text while driving. To answer this question, we will survey 16- to 20-year-old drivers who live in the United States. Ideally, we would ask them all by conducting a **census**. Of course, contacting every driver in this age group wouldn't be practical—it would take too much time and cost too much money. Instead, we pose the question to a **sample** chosen to represent the entire **population** of young drivers.

DEFINITION Population, Census, Sample

The **population** in a statistical study is the entire group of items or individuals we want information about.

A **census** collects data from every item or individual in the population.

A **sample** is a subset of items or individuals in the population from which we collect data.

Skill 2.A

EXAMPLE Sampling monitors and voters
Populations and samples

Vladimir Vladimirov/E+/Getty Images

PROBLEM: Identify the population and the sample in each of the following settings.

(a) The quality control manager at a factory selects 10 computer monitors from the 50 monitors produced during a particular hour and inspects each monitor for defects in construction and performance.

(b) Prior to an election, a news organization surveys 1000 registered voters to predict the percentage of voters who prefer candidate A for president.

SOLUTION:

(a) The population is all 50 computer monitors produced in this factory during that hour. The sample is the 10 monitors selected and inspected for defects.

(b) The population is all registered voters. The sample is the 1000 registered voters surveyed.

To identify the population, consider which items or individuals could have been selected for the sample. In part (a), the sample was selected from just computer monitors produced during that hour, so the population is limited to all monitors produced in this factory during that hour.

FOR PRACTICE, TRY EXERCISE 7

In a statistical study, the sample size is represented by n and the population size is represented by N. In part (a) of the example, the quality control inspector selected a sample of $n = 10$ computer monitors from the population of $N = 50$ monitors produced that hour.

Most statistical studies collect data from samples to answer investigative questions about larger populations. In part (b) of the example, the news organization wants to answer the investigative question "What are the plausible (believable) values for the percentage of all registered voters who prefer candidate A for president?" We refer to this unknown percentage as the population **parameter**. Suppose that 528 of the 1000 registered voters surveyed prefer candidate A for president. That's 52.8%, which we refer to as the sample **statistic**. Remember p and s: parameters come from populations and statistics come from samples.

DEFINITION Parameter, Statistic

A **parameter** is a number that describes some characteristic of a population.

A **statistic** is a number that describes some characteristic of a sample.

We use sample statistics to estimate population parameters. In the news organization survey, 52.8% of the sample of registered voters prefer candidate A for president. That's our best guess for the unknown population percentage. Should we conclude that exactly 52.8% of the population of registered voters prefer candidate A for president? No. If another sample of 1000 registered voters was surveyed, the percentage who prefer candidate A for president would probably yield a different sample statistic. Can we at least say that the actual population parameter is "close" to 52.8%? As you will learn in future units, that depends on what we mean by "close" and how the sample was selected.

Observational Units 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.[4]

Household	Region	Number of 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 Leighty/Photodisc/Getty Images

Most data tables follow this format: each row describes an **observational unit**, and each column contains the values of a **variable**. The entry in a single cell of the data table, representing the value of one variable for a specific observational unit, is called a *datum*—the singular form of the word *data*. Sometimes the observational units in a data set are called *individuals* or *cases*. Note that observational units can be people, animals, or things, such as photographs, sounds, or videos.

DEFINITION Observational unit, Variable

An **observational unit** is an item or individual described in a data set or statistical study.

A **variable** is a characteristic that can take different values for different observational units.

For the ACS data set, the *observational units* are households. The *variables* recorded for each household are region, number of people in the household,

time in their current dwelling, survey response mode, household income (in dollars), and whether the dwelling has internet access. 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 observational units. Region, time in dwelling, response mode, and internet access status are **categorical variables**. Number of people in the 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 item or individual into a particular group, called a category.

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

Sometimes categorical variables are called *qualitative variables* and quantitative variables are called *numerical variables*. It can also be helpful to think of a quantitative variable as one that takes number values for which it makes sense to find the average.

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. 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.

In many data sets, the variable "year" is 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 2026 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.

Skill 2.A

EXAMPLE Census at School Observational units and variables

PROBLEM: Census at School is an international project that collects data about primary and secondary school students using online 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.[5] We selected a random sample of 50 U.S. high school students who completed the survey in a recent year. The table displays data from some of the survey questions for the first 10 students in the sample.

EF Volart/Moment/Getty Images

State	Birth month	Age (years)	Handedness	Height (cm)	Number of home occupants	Allergies	Preferred communication method
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 observational units and variables in this data set.

(b) Classify each variable as categorical or quantitative.

SOLUTION:

(a) Observational units: 50 randomly selected U.S. high school students who completed the Census at School survey. Variables: State where student lives, birth month, age (years), handedness, height (cm), number of home occupants, whether the student has allergies, preferred communication method.

We'll see in Unit 1, Part II why selecting at random, as we did in this example, is a good idea.

(b) Categorical: State where student lives, birth month, handedness, whether the student has allergies, preferred communication method. Quantitative: Age (years), height (cm), and number of home occupants.

Note that birth month is categorical, even though the values listed are numbers. It wouldn't make sense to find the average for this variable!

FOR PRACTICE, TRY EXERCISE 13

As you will learn, the proper method of data analysis depends on whether a variable is categorical or quantitative. For that reason, it is important to distinguish between these two types of variables. Be sure to include any units of measurement for a quantitative variable (such as 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.

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.

There are two types of quantitative variables: *discrete* and *continuous*. Most **discrete quantitative variables** result from counting something, such as the number of people in a household or the number of lottery tickets a person buys until they win the jackpot. **Continuous quantitative variables** typically result

from measuring something, such as height (in inches) or time to run a 100-meter dash (in seconds). Age is technically a continuous quantitative variable—a high school student's age might be 17.30162... years. Even so, it is often treated as a discrete quantitative variable—for example, age = 17 years.

DEFINITION Discrete quantitative variable, Continuous quantitative variable

A **discrete quantitative variable** can take a set of possible values with gaps between them on the number line.

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

Note that the number of possible values of a discrete quantitative variable can be finite or infinite, whereas a continuous quantitative variable always has an infinite number of possible values. Some people say that the values of discrete quantitative variables are *countable*, which means that we can number the possible values of the variable using whole numbers. As an example, the number of lottery tickets a person buys until they win the jackpot can take any of the values 1, 2, 3, and so on. The values of a continuous quantitative variable, however, are not countable. For instance, the height of a fully grown giraffe (in feet) could be *any* value between about 14 feet and 18 feet.[6] There is no way to number all of these possible heights using whole numbers.

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 large local high school gives Malena permission to go to the student parking lot and collect some data. Malena selects a random sample of 50 cars from the lot and creates a spreadsheet with 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 (in miles per gallon).

1. Identify the sample and the population in this statistical study.
2. Identify the observational units and variables. Classify each variable as categorical or quantitative.
3. Formulate a valid investigative question that Malena could ask about the number of stickers on students' cars.

SECTION 1A Summary

- **Statistics** is the science and art of collecting, analyzing, and drawing conclusions from data.
- The **statistical problem-solving process** involves four steps: formulate questions, collect data, analyze data, and interpret results.
- Most statistical studies collect data from samples to answer investigative questions about larger populations. A valid investigative question is based on data that vary.

- The **population** in a statistical study is the entire group of items or individuals we want information about. A **census** collects data from every item or individual in the population.
- A **sample** is a subset of items or individuals in the population from which we collect data.
- A **statistic** is a number that describes some characteristic of a sample. A **parameter** is a number that describes some characteristic of a population.
- An **observational unit** is an item or individual described in a data set or statistical study. Observational units may be people, animals, or things.
- A **variable** is a characteristic that can take different values for different observational units.
- 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. Sometimes categorical variables are called *qualitative variables* and quantitative variables are called *numerical variables*.
- There are two types of quantitative variables: discrete and continuous. A **discrete quantitative variable** can take a set of possible values with gaps between them on the number line. A **continuous quantitative variable** can take any value in an interval on the number line. Discrete quantitative variables usually result from counting something; continuous quantitative variables usually result from measuring something.

AP® EXAM TIP
AP® Classroom
Review the content of this section with the resources in AP® Classroom for Topics 1.1 and 1.2.

SECTION 1A Exercises

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

Determining Investigative Questions

1. **Regular exercise** Determine whether each of the following is a valid statistical investigative question. Explain your answer.

(a) What are the plausible values for the proportion of all U.S. adults who engage in vigorous exercise at least once per week?

(b) On how many days in the past week did you engage in vigorous exercise?

2. **Carrying cash** Determine whether each of the following is a valid statistical investigative question. Explain your answer.

(a) How much money are you carrying right now?

(b) What are the plausible values for the average amount of money that all high school students are carrying on the first day of school?

3. **Boarding time** Airlines are interested in finding ways for passengers to board their flights more quickly and efficiently. Researchers tested different boarding methods using a group of 72 volunteer passengers of varying ages. The researchers compared the "back-to-front" boarding method used by many airlines at the time with a modified version of this method proposed by astrophysicist Jason Steffen. In the modified method, passengers lined up in advance at the gate in a predetermined order, so that all passengers with an even-numbered window seat would board first from back to front, followed by those with an odd-numbered window seat, then those with an even-numbered middle seat, and so on. On average, the Steffen method was twice as fast as the traditional boarding method.[7] Determine the investigative question in this statistical study.

4. **Paid to quit** In an effort to reduce health care costs, General Motors sponsored a study to help its employees stop smoking. In the study, 439 volunteer subjects were assigned at random to receive up to $750 for quitting for a year, while the other 439 volunteer subjects were simply encouraged to use traditional methods to stop smoking. After one year, people who had the financial incentive were 3 times more likely to have quit smoking.[8] Determine the investigative question in this statistical study.

5. **Online bullying** The Pew Research Center surveyed a random sample of 1316 U.S. teens aged 13 to 17. The survey asked about whether the person had ever experienced online bullying and also recorded data about the respondents' age, race/ethnicity, area where they live (urban/suburban/rural), and household income.[9] Formulate a valid investigative question that researchers could ask using the data from this study.

6. **Beach living?** Researchers surveyed more than 15,000 people in Europe and Australia about their health and where they live. One question on the survey was: "In general, would you say your health is: 1 – Excellent, 2 – Very good, 3 – Good, 4 – Fair, or 5 – Poor?" Another question asked, "Approximately how far do you live from the coast in kilometers (km): <1 km, 1–<2 km, 2–<5 km, 5–<10 km, 10–<20 km, 20–<50 km, 50–<100 km, or 100+ km?"[10] Formulate a valid investigative question that researchers could ask using the data from this study.

Populations and Samples

7. **Sampling stuffed envelopes** A large retailer prepares its customers' monthly credit card bills using an automatic machine that folds the bills, stuffs them into envelopes, and seals the envelopes for mailing. To ensure that the envelopes are completely sealed, inspectors choose 40 envelopes at random from the 1000 envelopes stuffed in one hour for visual inspection. Identify the population and the sample. pg 5

8. **Student archaeologists** An archaeological dig turns up large numbers of pottery shards, broken stone tools, and other artifacts. Students working on the project classify each artifact and assign a number to it. The counts in different categories are important for understanding the site, so the project director chooses 2% of the artifacts at random and checks the students' work. Identify the population and the sample.

9. **Students as customers** A high school's student newspaper plans to survey businesses in their large city about the importance of students as customers. From an alphabetical list of all local businesses, the newspaper staff chooses 150 businesses at random. Of these, 73 return the questionnaire mailed by the staff. Identify the population and the sample.

10. **Customer satisfaction** A department store mails a customer satisfaction survey to people who make credit card purchases at the store. This month, 45,000 people made credit card purchases. Surveys are mailed to 1000 of these people, chosen at random, and 137 people return the survey. Identify the population and the sample.

11. **Dairy cows** A farmer has 1000 dairy cows and wants to estimate the minimum daily milk production for dairy cows in the herd. For 20 randomly chosen dairy cows, the minimum production was 6.3 gallons. Identify the population, the parameter, the sample, and the statistic in this setting.

12. **On sale now?** Advertisements for a local clothing store claim that 60% of the items are on sale this week. A consumer group randomly selected 20 items and found that only 9 were on sale. Identify the population, the parameter, the sample, and the statistic in this setting.

Observational Units and Variables

13. **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. pg 7

Grade level	Dominant hand	GPA	Number of children in family	Homework time last night (min)	Type of phone
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 observational units and variables in this data set.

(b) Classify each variable as categorical or quantitative.

14. **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.[11]

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 observational units and variables in this data set.

(b) Classify each variable as categorical or quantitative.

15. **Hit movies** *Avatar* was the top-earning movie based on box-office receipts worldwide as of February 2025. The following table displays data on several popular movies.[12] Identify the observational units and variables in this data set. Then classify each variable as categorical or quantitative.

Movie	Year	Rating	Time (min)	Genre	Box office ($)
Avatar	2009	PG-13	162	Action	2,923,706,026
Avengers: Endgame	2019	PG-13	181	Action	2,748,242,781
Avatar: The Way of Water	2022	PG-13	190	Action	2,313,161,020
Titanic	1997	PG-13	194	Drama	2,223,048,786
Star Wars: The Force Awakens	2015	PG-13	136	Adventure	2,056,046,835
Avengers: Infinity War	2018	PG-13	156	Action	2,048,359,754
Spider Man: No Way Home	2021	PG-13	148	Action	1,921,206,586
Inside Out 2	2024	PG	100	Adventure	1,698,863,816
Jurassic World	2015	PG-13	124	Action	1,671,063,641
The Lion King	2019	PG	118	Adventure	1,661,454,403

16. **Skyscrapers** Here is some information about the tallest buildings in the world as of 2025.[13] Identify the observational units 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 PNB 118	Malaysia	678.9	118	Mixed	2023
Shanghai Tower	China	632.0	128	Mixed	2015
Abraj Al Bait	Saudi Arabia	601.0	120	Hotel	2012
Ping An Finance Center	China	599.1	115	Office	2017
Lotte World Tower	South Korea	554.5	123	Mixed	2017
One World Trade Center	United States	541.3	94	Office	2014
Tianjin CTF Finance Center	China	530.4	97	Mixed	2019
Guangzhou CTF Finance Centre	China	530.0	111	Mixed	2016
CITIC Tower	China	527.7	109	Mixed	2018

17. **Protecting wood** How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study that attempted to answer this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables the researchers recorded: type of wood, paint thickness, paint color, weathering time (1, 2, or 3 months), and number of blemishes.

(a) Identify the observational units.

(b) Classify each variable as categorical, quantitative (discrete), or quantitative (continuous).

18. **Social media** A social media company records data on each of its users for several variables: internet provider, age, how many times they visited the site, total time spent on the site, country where they live, and how long since they created a member profile.

(a) Identify the observational units.

(b) Classify each variable as categorical, quantitative (discrete), or quantitative (continuous).

Multiple Choice *Select the best answer for each question.*

Exercises 19 and 20 refer to the following setting. A realtor in Hilton Head, South Carolina, collected data on all homes sold in the town during 2025. 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 (in dollars).

19. The observational units in this data set are

(A) dollars.

(B) all Hilton Head homeowners who sold their homes during 2025.

(C) all homes sold in Hilton Head during 2025.

(D) all homes in Hilton Head during 2025.

20. This data set contains

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

(B) 7 variables, 3 of which are categorical.

(C) 8 variables, 3 of which are categorical.

(D) 8 variables, 4 of which are categorical.

21. The manager of a town's athletic complex is trying to decide whether to convert some of the existing tennis courts to pickleball courts. To help with the decision, the manager will survey a random sample of 200 adult town residents about whether they are in favor of this conversion. If the data provide convincing evidence that more than half of all adult town residents favor the conversion, the manager will submit a budget proposal to the town council. Which of the following is the investigative question for this statistical study?

(A) Is the proportion of adult town residents in the sample who use the town's athletic complex greater than 0.5?

(B) Is the proportion of all adult town residents who use the town's athletic complex greater than 0.5?

(C) Is the proportion of adult town residents in the sample who favor converting some existing tennis courts to pickleball courts greater than 0.5?

(D) Is the proportion of all adult town residents who favor converting some existing tennis courts to pickleball courts greater than 0.5?

SECTION 1B Displaying and Describing Categorical Data

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

- Make and interpret a frequency table or a relative frequency table for a distribution of categorical data.
- Make and interpret bar charts and pie charts of categorical data.
- Compare distributions of categorical data.
- Identify what makes some graphs of categorical data misleading.

AP® EXAM TIP: AP® Classroom
Preview the content of this section with the resources in AP® Classroom for Topics 1.3 and 1.4.

A variable generally takes values that vary from one observational unit 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.

In this section, you will learn how to display and describe the distribution of one categorical variable. You will also learn how to compare distributions of a categorical variable in two or more groups. Sections 1C, 1D, and 1E focus on displaying and describing the distribution of one quantitative variable, and on comparing distributions of a quantitative variable in multiple groups. Section 2A examines relationships between two categorical variables, and Unit 5 examines relationships between two quantitative variables. This process of exploratory data analysis is known as *descriptive statistics*.

Summarizing Categorical Data with Tables

We can summarize the distribution of a categorical variable 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 observational units having each value of a variable.

A **relative frequency table** shows the proportion or percentage of observational units having each value of a variable.

We can use a frequency table or a relative frequency table to describe the distribution of a categorical variable or to help justify a claim about the variable in context.

Skills 3.A, 4.B

EXAMPLE Call me, maybe? Summarizing categorical data with tables

SDI Productions/E+/Getty Images

PROBLEM: Here are the data on preferred method of communicating with friends for all 50 students in the Census at School sample from Section 1A:

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

(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 *ratio* (as a fraction), *proportion*, and *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 1

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 could 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, such as height (cm) in the Census at School data set. We'll look at a better option for summarizing the distribution of a quantitative variable in Section 1C.

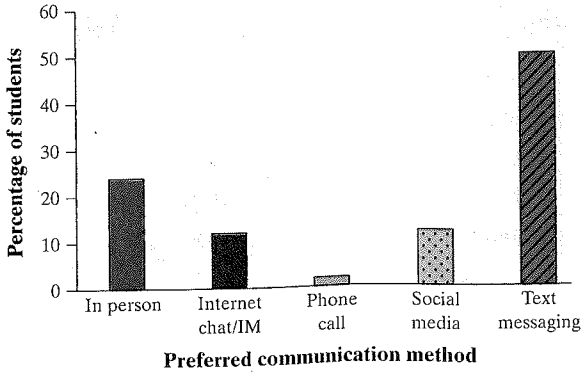
Displaying Categorical Data: Bar Charts and Pie Charts

A frequency table or relative frequency table summarizes a categorical 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 chart** (also known as a *bar graph*).

DEFINITION Bar chart

A **bar chart** is a graph of data for one categorical variable that displays each category as a bar. The height of each bar shows the category frequency or relative frequency.

Figure 1.1 shows a bar chart 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.



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%

FIGURE 1.1 Bar chart and relative frequency table of the distribution of preferred communication method for a random sample of 50 U.S. high school students.

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

HOW TO MAKE A BAR CHART

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 observational units in each category.
2. **"Scale" the axes.** Write the names of the categories in a logical order 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 observational units in that category.

Note: You can also make a bar chart with the axes reversed, so that values of the categorical variable are on the vertical axis and frequencies or relative frequencies are on the horizontal axis.

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?" We can use a graph to describe the distribution of a categorical variable or to help justify a claim about the variable in context.

In Figure 1.1, the bar chart 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%).

Skills 3.A, 4.A

EXAMPLE What's on the radio? Bar charts

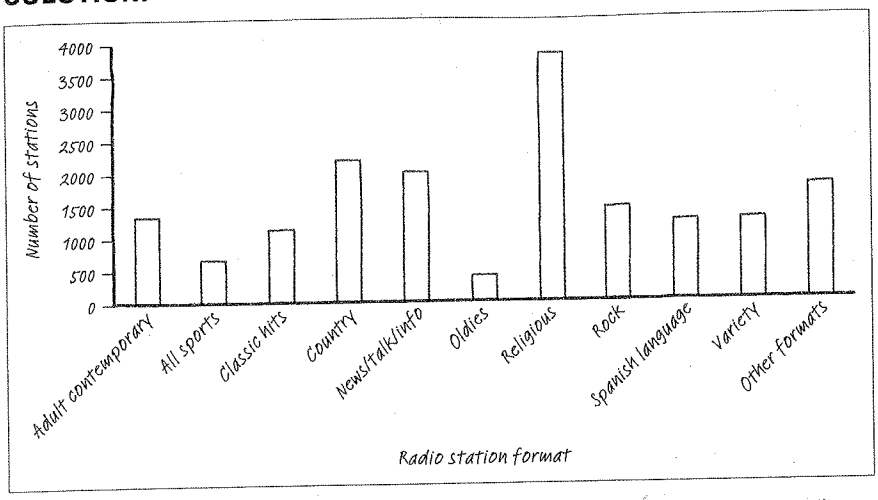
PROBLEM: Nielsen Audio, the rating service for radio audiences, places U.S. radio stations into categories that describe the kinds of programs they broadcast. The frequency table summarizes the distribution of station format in a recent year.[14]

ACORN 1/Alamy Stock Photo

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**

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

SOLUTION:



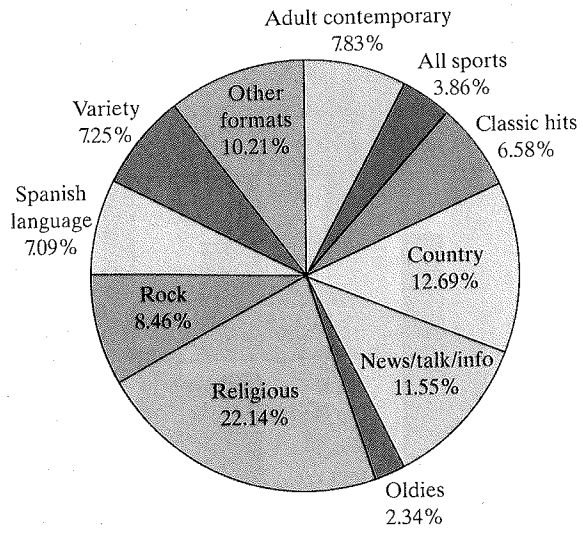
To make the bar chart:
1. **Draw and label the axes.**
2. **"Scale" the axes.** The largest frequency is 3837, so we choose a vertical scale from 0 to 4000, with tick marks 500 units apart.
3. **Draw bars** above the category names.

For 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.

Note that the observational units in this statistical study are radio stations. The variable recorded for each observational unit is station format, which is categorical.

FOR PRACTICE, TRY EXERCISE 5

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** is a graph of data for one categorical variable that displays each category as a slice of the "pie." The area of each slice is proportional to the category frequency or relative frequency.

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 observational units in that category. Pie charts are challenging to make by hand, but technology can easily do the job for you.

A bar chart or pie chart that displays a distribution of categorical data must include *all* observational units in the data set. This might require including an "Other" category, as in the radio station example.

Comparing Distributions of Categorical Data

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

DEFINITION Side-by-side bar chart

A **side-by-side bar chart** is a graph of data for one categorical variable in each of two or more groups that displays a separate bar corresponding to each group for every category. The height of each bar shows the category frequency or relative frequency within that group.

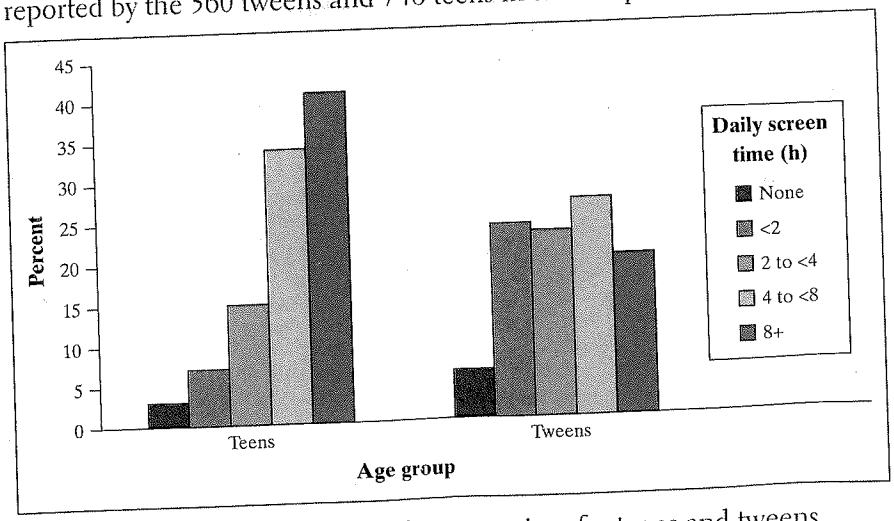
It's a good idea to use relative frequencies when comparing data for multiple groups, especially if the groups have different sizes.

Skill 4.A

EXAMPLE Screen time for teens and tweens
Comparing distributions of categorical data


Melanie Acevedo/DigitalVision/Getty Images

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 chart summarizes the data on daily screen time reported by the 560 tweens and 746 teens in the sample.[15]



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

SOLUTION:

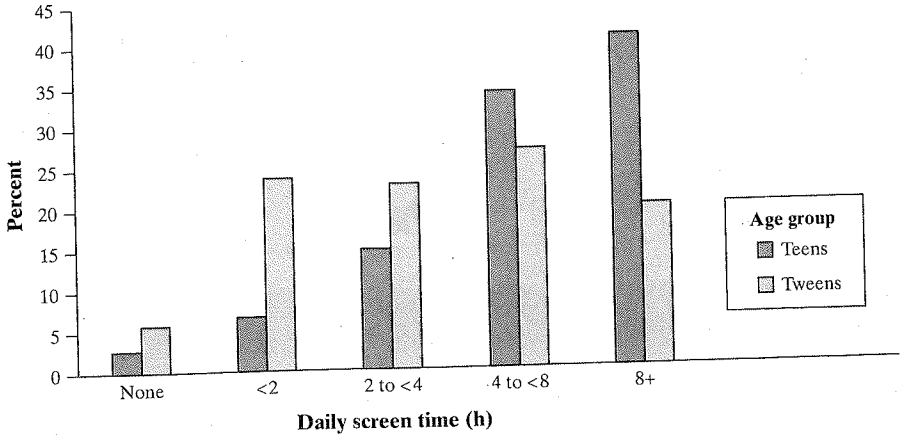
Tweens tend to report less daily screen time than teens. Tweens are more likely than teens to spend no time (6% versus 3%), less than 2 hours (24% versus 7%), and from 2 to less than 4 hours (23% versus 15%) on digital devices per day. Teens are more likely than tweens to spend from 4 to less than 8 hours (34% versus 27%) and 8 or more hours (41% versus 20%) on digital devices per day. Fewer than half (47%) of tweens report 4 or more hours of screen time per day, while 75% of teens report 4 or more hours of screen time per day.

FOR PRACTICE, TRY EXERCISE 11

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 using only this relative frequency table.

		Daily time spent on digital devices				
		None	< 2 hours	2 to <4 hours	4 to <8 hours	8+ hours
Age group	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 chart. 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 chart represents.



It is possible, but more challenging, to compare distributions of categorical data with pie charts.

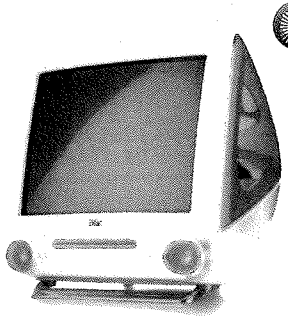
Misleading Graphs

Bar charts 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 × height) varies in proportion to the height, and our eyes receive the right impression about the quantities being compared.

Skill 3.A

EXAMPLE Who buys iMacs? Misleading graphs



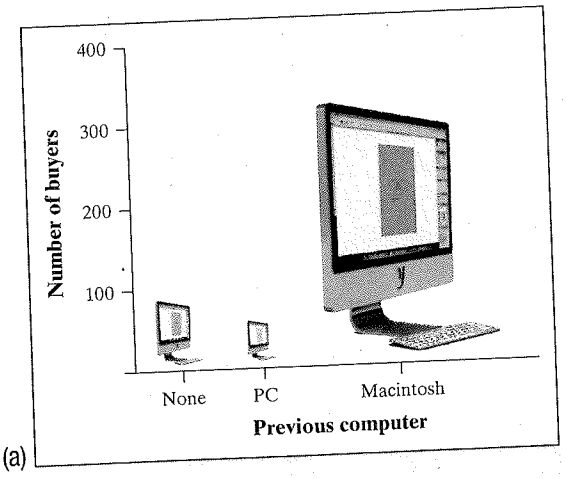
Getty Images/Getty Images

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.[16]

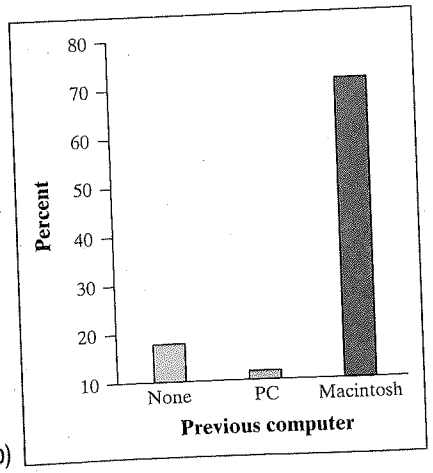
Previous ownership	Count	Percentage (%)
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	**500**	**100.0**

(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.



(a)



(b)

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.

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!

(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.

FOR PRACTICE, TRY EXERCISE 15

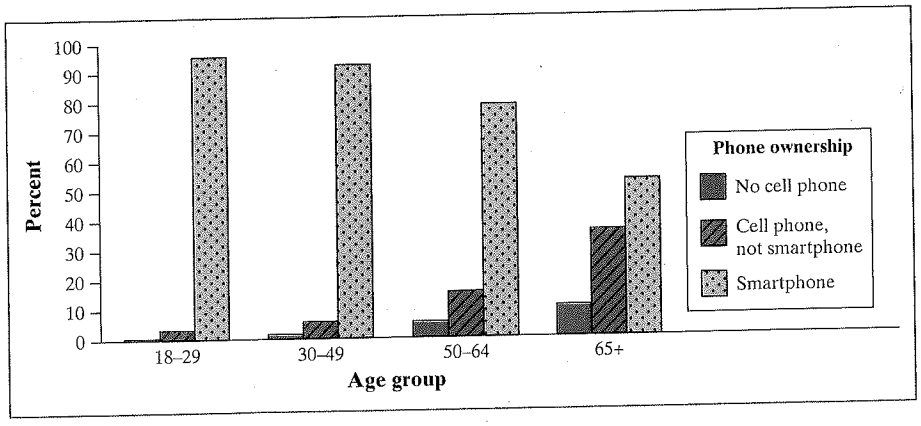
There are two important lessons about misleading graphs to be learned from this example: (1) **beware the pictograph and (2) watch those scales.**

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.[17]

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

1. Make a relative frequency table for these data.
2. Make a relative frequency bar chart to display the distribution of cell phone ownership for the 1502 people in the sample. Describe what you see.
3. The side-by-side bar chart displays the distribution of cell phone ownership for each of four age groups in the sample. Use the graph to justify the claim that cell phone ownership differs across age groups.



SECTION 1B Summary

- 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 observational units having each value or a **relative frequency table** that shows the proportion or percentage of observational units having each value.
- A **bar chart** or **pie chart** can be used to display the distribution of a categorical variable.
- You can use a table or graph of categorical data to describe the distribution of a categorical variable or to justify a claim about the variable in context.

AP® EXAM TIP
AP® Classroom

Review the content of this section with the resources in AP® Classroom for Topics 1.3 and 1.4.

- You can use a **side-by-side bar chart** 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 chart with pictures whose height and width vary.

SECTION 1B Exercises

Summarizing Categorical Data with Tables

1. **New York squirrels** How do squirrels in New York pg 14 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:[18]

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.

2. **Disc dogs** Here is a list of the breeds of dogs that won the World Canine Disc Championships over a 45-year period.[19]

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.

3. **Don't worry, be happy!** College admissions departments are continually honing their message to appeal to high school students. Knowing about potential students' long-term goals and priorities helps colleges tailor information about the benefits of attending their institutions. Several colleges use an online survey to collect data about secondary school students. Here is one of the survey questions:

Which would you prefer to be? Select one.
_____ Rich _____ Happy _____ Famous _____ Healthy

The relative frequency table summarizes data on preferred status for a random sample of students who completed the survey in a recent year.

Preferred status	Proportion
Famous	0.040
Happy	0.520
Healthy	0.133
Rich	0.307

(a) There were a total of 75 students in the sample. How many said that their preferred status was rich?

(b) A college admissions officer suggests a recruiting campaign for prospective students that focuses on future happiness. Explain why this suggestion makes sense based on the data.

4. **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 2500 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.[20]

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.

Displaying Categorical Data: Bar Charts and Pie Charts

5. **Birth days** The frequency table summarizes data on the numbers of babies born on each day of a single week in the United States.[21]

pg 16

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 chart to display the data. Describe what you see.

6. **Going up?** Oliver Roeder wrote 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.[22]

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

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

7. **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.[23]

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

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

(b) Make a bar chart to display the data. Describe the distribution.

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

8. **Spam** Email spam is very annoying. The relative frequency table summarizes data on the most common types of spam.[24]

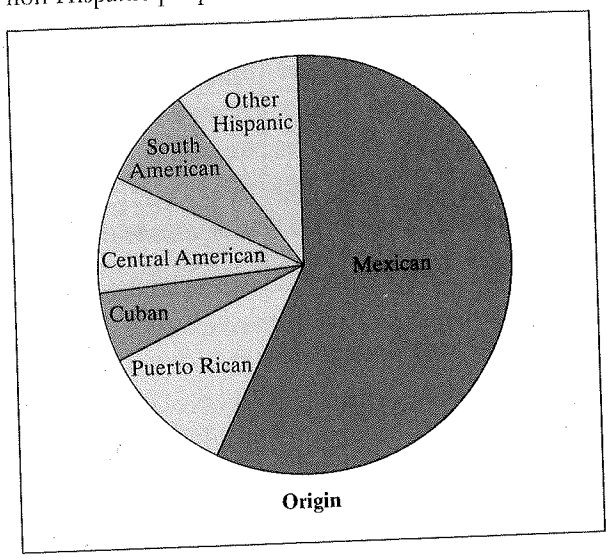
Type of spam	Percentage
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

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

(b) Make a bar chart to display the data. Describe the distribution.

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

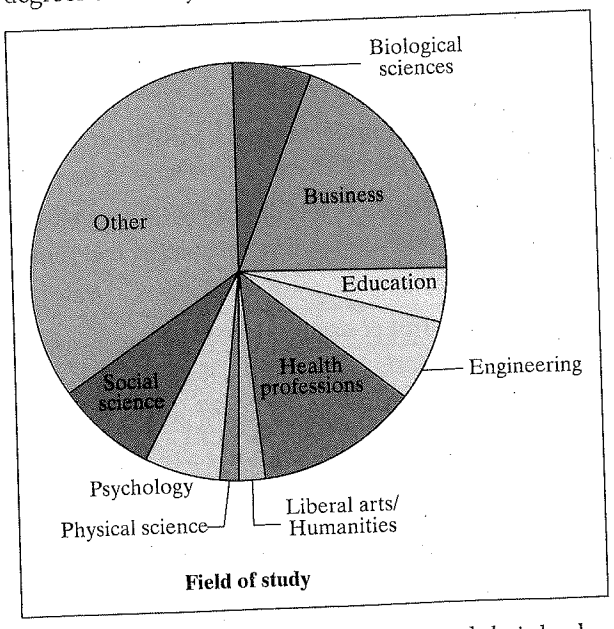
9. **Family origins** Here is a pie chart of U.S. Census Bureau data showing the origin of more than 62 million Hispanic people in the United States.[25]



(a) About what percentage of Hispanic people in the United States are of Mexican origin? Puerto Rican origin?

(b) Can we use this graph to justify the claim that people of South American origin make up the smallest proportion of Hispanic people in the United States? Explain your answer.

10. **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.[26]



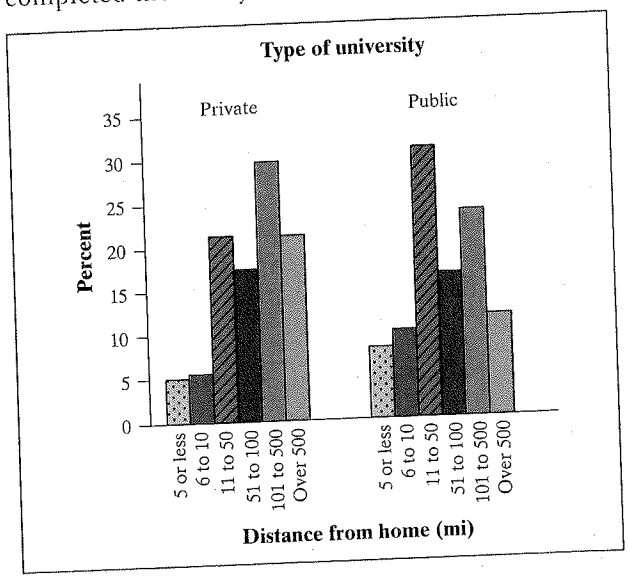
(a) About what percentage of students earned their bachelor's degrees in business? In health professions? In social science?

(b) Can we use this graph to justify the claim that the proportions of bachelor's degrees earned in psychology and engineering were about the same in this recent year in the United States? Explain your answer.

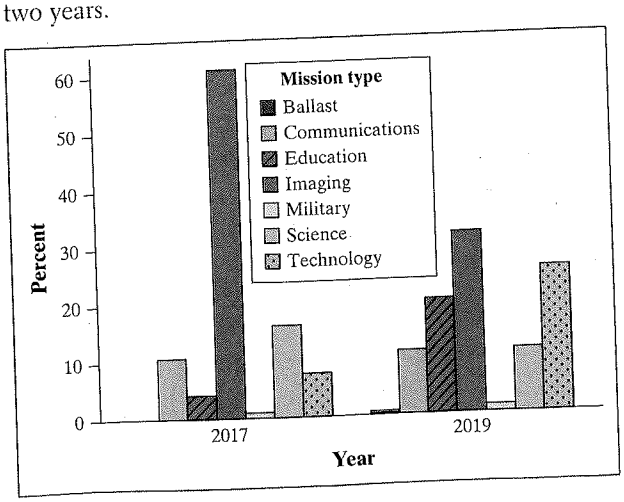
Comparing Distributions of Categorical Data

11. **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 chart shows the percentage of students at public and private 4-year colleges who chose each option.[27] Compare the distributions of distance from home for students from private and public 4-year colleges who completed the survey.

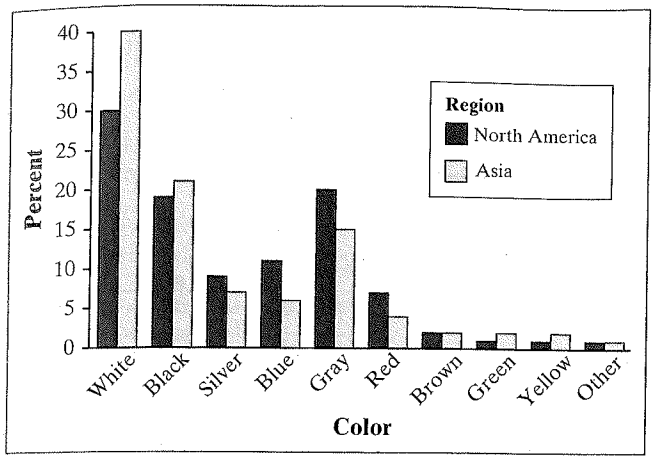
pg 18



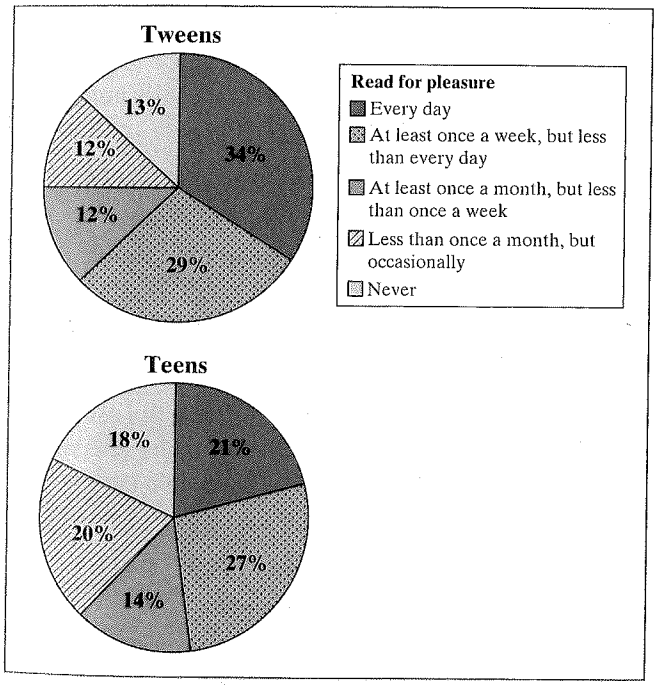
12. **More CubeSat missions** Refer to Exercise 4. The side-by-side bar chart displays data on the types of missions undertaken by CubeSats launched in 2017 and 2019. Compare the distributions of mission type for these two years.



13. **Popular car colors** Favorite car colors may differ among regions of the world. The side-by-side bar chart displays data on the most popular car colors in a recent year for North America and Asia.[28] Describe similarities and differences in the distributions for these two regions.



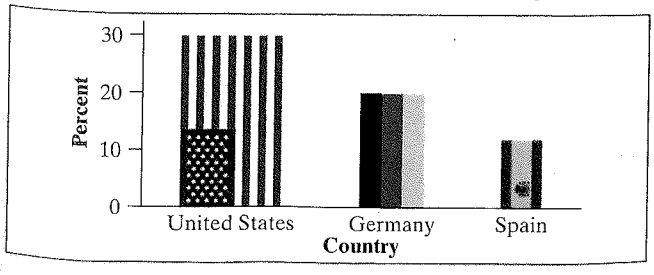
14. **Reading for pleasure** Researchers surveyed separate random samples of U.S. tweens (ages 8–12) and teens (ages 13–18) about how often they read for pleasure. The pie charts display the data.[29] Describe similarities and differences in the distributions for tweens and teens.



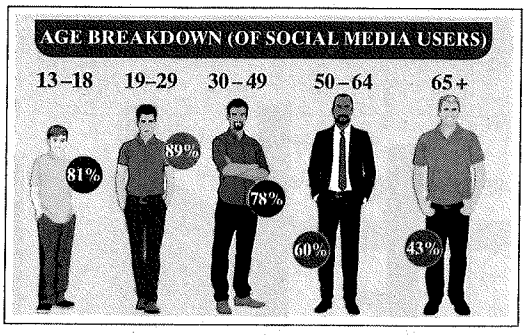
Misleading Graphs

15. **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.[30]

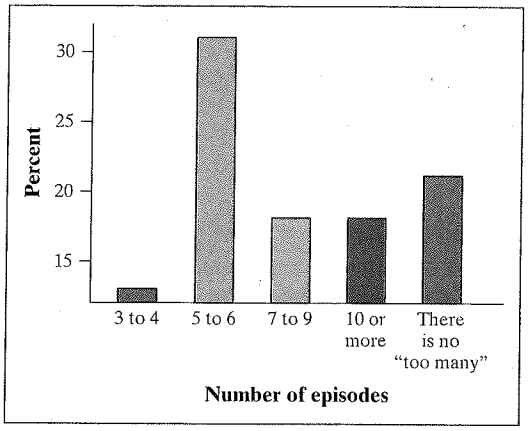
pg 20



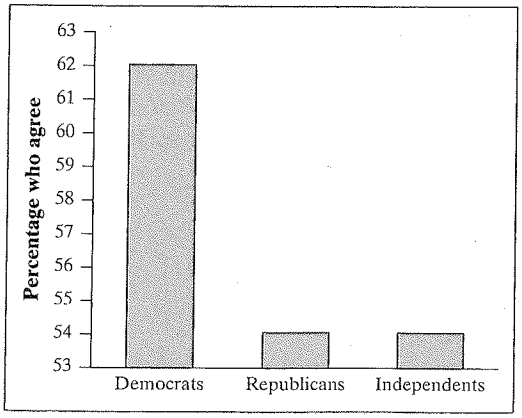
16. **Social media** The Pew Research Center surveyed a random sample of U.S. teens and adults about their use of social media. The following pictograph (not created by the Pew Research Center!) displays the percentage of people in various age groups who report using social media. Explain how this graph is misleading.



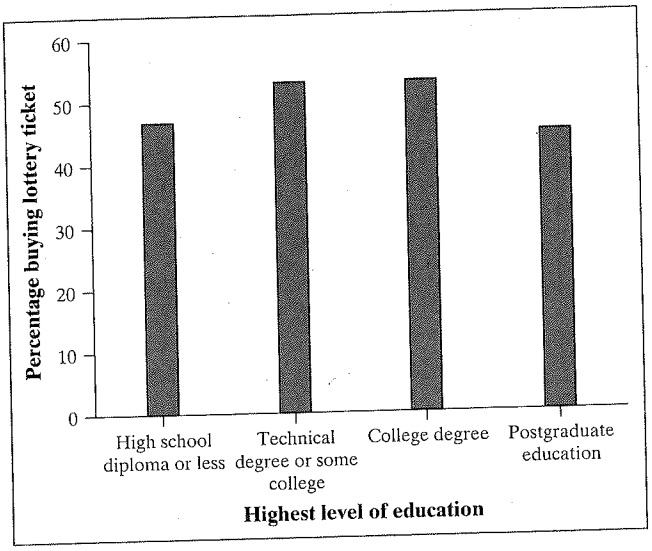
17. **Binge-watching** Do you "binge-watch" television series by viewing multiple episodes of a series at one sitting? A survey of 800 people who binge-watch were asked how many episodes is too many to watch in one viewing session. The results are displayed in the bar chart.[31] Explain how this graph is misleading.



18. **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 chart 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.)



19. **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.[32] Here is a modified bar chart of the data that displays only the percentage of respondents who bought a lottery ticket in each category of highest educational level.



(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.

Connecting the Statistical Practices *Multi-focus exercises that connect two or more statistical practices.*

20. **Teen versus tween reading** Refer to Exercise 14.

(a) Identify the populations and the samples.

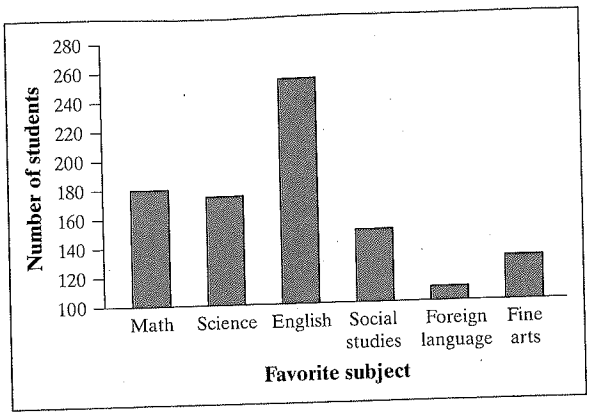
(b) Identify the variable of interest. Classify the variable as categorical or quantitative.

(c) Make a relative frequency table that compares the survey responses for teens and tweens.

(d) Formulate a possible investigative question for this statistical study.

Multiple Choice *Select the best answer for each question.*

21. The following bar chart 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.

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

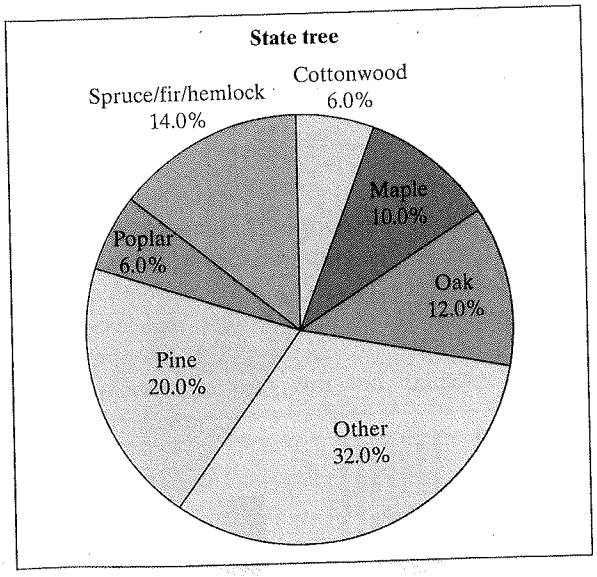
(A) The distribution of car color for vehicles purchased in the last month

(B) The distribution of unemployment percentage for each of the 50 states

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

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

23. The pie chart displays the distribution of the designated state tree by type for the 50 U.S. states. The category "Other" includes all trees that are the state tree for two or fewer states.



Which of the following conclusions can we justify from this graph?

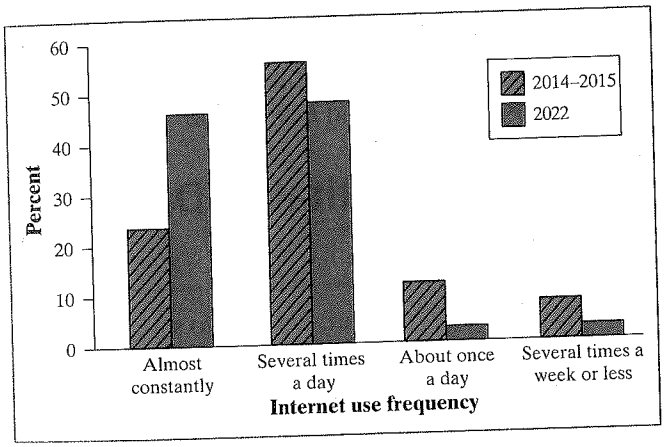
(A) The cottonwood is the state tree for 12 states.

(B) Taken together, oak, pine, and maple are the state trees for more than half the states.

(C) There are 10 states that have designated a pine as their state tree.

(D) There is no state that has designated the Eastern Red Cedar as its state tree.

24. How has teen internet use changed over time? The Pew Research Center surveyed separate random samples of U.S. teens aged 13 to 17 in 2014–2015 and in 2022. The side-by-side bar chart summarizes the teens' responses to the question, "About how often do you use the internet, either on a computer or a cellphone?"[33]



Which of the following is a correct statement?

(A) A majority of teens said "almost constantly" in 2022.

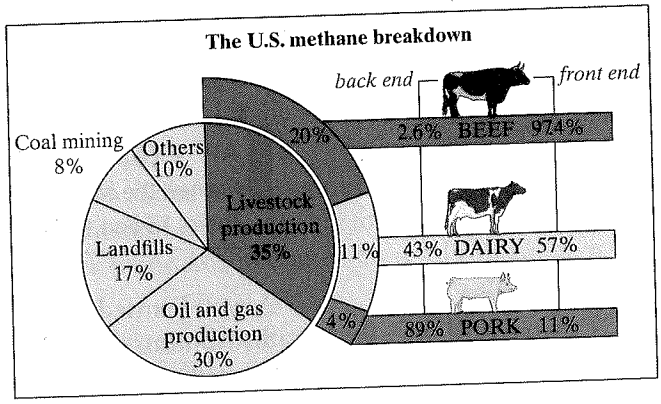
(B) The percentage of teens who said "almost constantly" approximately doubled from 2014–2015 to 2022.

(C) A majority of teens said "several times a day" in both 2014–2015 and 2022.

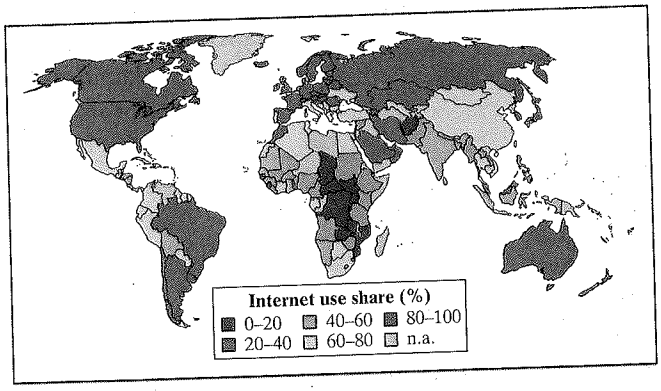
(D) The percentage of teens who said "about once a day" doubled from 2014–2015 to 2022.

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

25. **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.[34] 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.



26. **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.[35] Write a thorough analysis of what the graph shows. Be sure to include a comparison of internet use in different regions of the world.



SECTION 1C Displaying and Describing Quantitative Data with Graphs

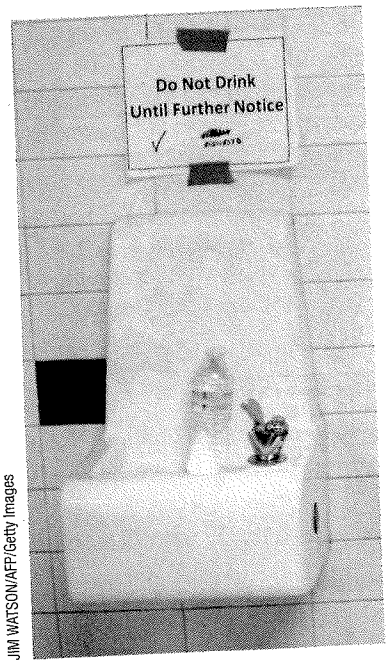
LEARNING TARGETS *By the end of the section, you should be able to:*

- Make and interpret dotplots of quantitative data.
- Describe the shape of a graph of quantitative data.
- Describe the distribution of a quantitative variable.
- Make and interpret stem-and-leaf plots of quantitative data.
- Make and interpret histograms of quantitative data.

AP® EXAM TIP: AP® Classroom
Preview the content of this section with the resources in AP® Classroom for Topics 1.5 and 1.6.

As you learned in Section 1B, you can use a bar chart or pie chart 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*, *stem-and-leaf plots*, and *histograms*.

Displaying and Describing Quantitative Data: Dotplots



JIM WATSON/AFP/Getty Images

In April 2014, managers for the city of Flint, Michigan, decided to save money by using water from the Flint River rather than continuing to buy water sourced from Lake Huron. Soon after, Flint residents noticed that the water coming out of their taps looked, tasted, and smelled bad. Some residents developed rashes, hair loss, and itchy skin. Authorities insisted that drinking water from the Flint River was safe.

As part of its regular testing program, city officials measured lead levels in water samples collected from 71 randomly selected Flint dwellings between January and June 2015. Here are the data (in parts per billion, ppb).[36] The U.S. Environmental Protection Agency (EPA) requires action if a water system's lead level exceeds 15 parts per billion.

0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	2	2	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3	3	3
3	3	3	4	4	5	5	5	5	5	5	5
5	6	6	6	6	7	7	7	8	8	9	10
10	11	13	18	20	21	22	29	42	42	104	

Figure 1.2 is a **dotplot** of these data.

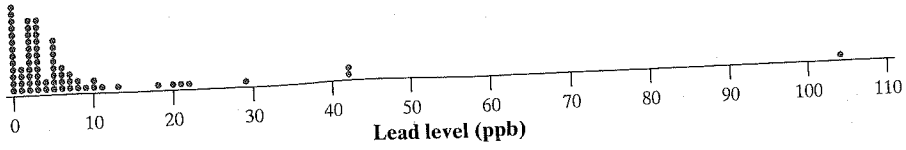


FIGURE 1.2 Dotplot of lead levels in water samples from 71 randomly selected Flint, Michigan, dwellings in January to June, 2015.

DEFINITION Dotplot

A **dotplot** is a graph of data for one quantitative variable that displays each data value as a dot above its location on a number line.

A dotplot is the simplest graph for displaying the distribution of a quantitative variable. It is fairly easy to make a dotplot by hand for small sets of quantitative data.

HOW TO MAKE A DOTPLOT

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

Note: You can also make a dotplot that is oriented vertically, by placing each dot to the right of the location on the vertical axis corresponding to its data value.

Remember what we said in Section 1B: making a graph is not an end in itself. When you look at a graph, always ask, "What do I see?" The dotplot in Figure 1.2 shows that $8/71 = 0.113 = 11.3\%$ of the Flint water samples had lead levels that exceeded 15 parts per billion. Does that mean Flint's water system required action based on the EPA guideline? We'll answer that question—and tell you the rest of the story about the Flint water crisis—in Section 1D.

EXAMPLE Eating healthy Dotplots

Skills 3.A, 4.B

PROBLEM: How healthy is plant-based yogurt? The *Nutrition Action Healthletter* provided data on calories, saturated fat, protein, calcium, and total sugars for many popular brands of yogurt. Here are the data on the amount of added sugar, in teaspoons (tsp), in single-serving containers for all of the plant-based yogurt brands:[37]

0.0	1.0	1.0	2.5	0.0	0.0	1.0	1.5	2.0	2.0	3.0	3.5	2.5
3.0	5.0	3.0	3.5	3.5	3.5	2.0	4.0	3.5	2.0	3.5	1.5	2.0

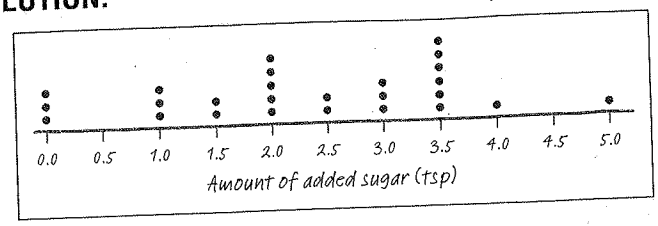
Clare Gainey/Alamy Stock Photo

(a) Make a dotplot of these data.

(b) The American Heart Association (AHA) recommends a maximum of 6 teaspoons of added sugar per day for women.[38] Can we use the graph in part (a) to justify the claim that a majority of these plant-based yogurt brands have less than half of the AHA's daily recommendation for added sugar? Explain your answer.

SOLUTION:

(a)



To make the dotplot:

1. **Draw and label the axis.** Be sure to include units along with the variable name.
2. **Scale the axis.** The smallest value is 0.0 and the largest value is 5.0. So, we choose a scale from 0 to 5 with tick marks placed 0.5 unit apart.
3. **Plot the values.**

(b) Half of the AHA's recommended daily maximum would be 3 teaspoons of added sugar. From the graph $15/26 = 0.577 = 57.7\%$ of these yogurt brands have less than 3 teaspoons of added sugar. Because $57.7\% > 50\%$, the dotplot can be used to justify the claim that a majority of these plant-based yogurt brands have less than half of the AHA's daily recommendation for added sugar.

FOR PRACTICE, TRY EXERCISE 1

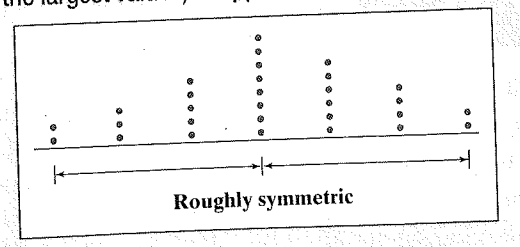
We can use a dotplot or other graph of quantitative data to help justify a claim about a quantitative variable in context, as in the preceding example, or to describe its distribution.

Describing the Shape of a Graph of Quantitative Data

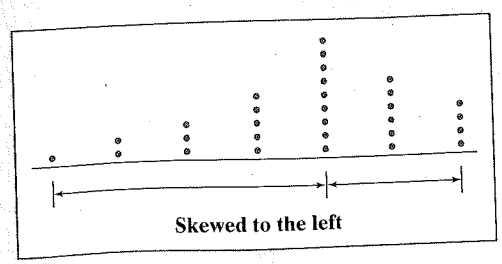
When you describe the shape of a dotplot or another graph of quantitative data, focus on the main features. Look for clear *peaks*, not for minor ups and downs in the graph. Look for distinct *clusters* of values and obvious *gaps*. Decide whether the distribution is **roughly symmetric**, **skewed to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

DEFINITION Roughly symmetric, Skewed to the left, Skewed to the right

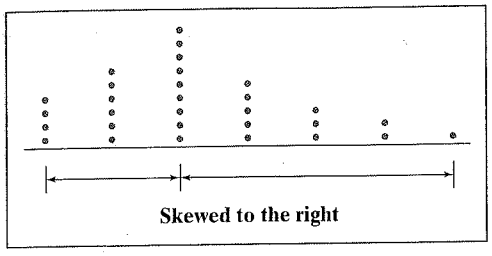
A distribution is **roughly symmetric** if the right side of the graph (containing the half of the observations with the largest values) is approximately a mirror image of the left side.

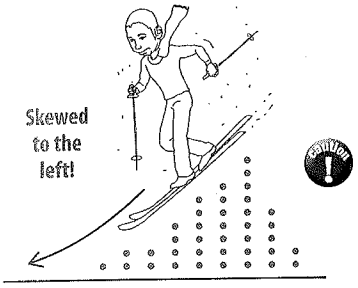


A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.



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

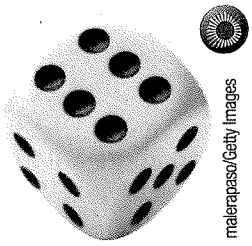




We could also describe a distribution with a long tail to the left as "skewed toward negative values" or "negatively skewed," and a distribution with a long right tail as "positively skewed." For ease, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed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

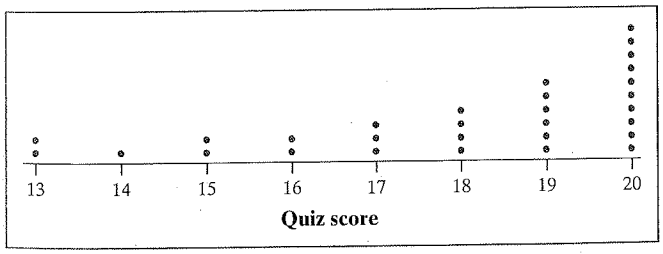
EXAMPLE Quiz scores and die rolls
Describing shape

Skill 4.A

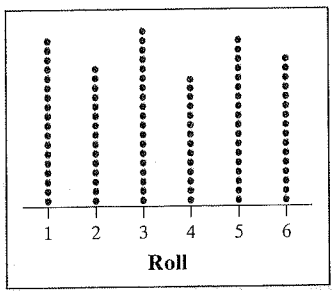


PROBLEM: The dotplots display two different sets of quantitative data. Graph (a) shows the scores on a 20-point quiz for each of the 30 students in a statistics class. Graph (b) shows the results of 100 rolls of a six-sided die. Describe the shape of each distribution.

(a)



(b)



SOLUTION:

(a) The distribution of statistics quiz scores is skewed to the left, with a single peak at 20 (a perfect score).

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

FOR PRACTICE, TRY EXERCISE 5

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

DEFINITION Approximately uniform

A distribution in which the frequency or relative frequency of each possible value is about the same is **approximately uniform.**

Graphs with a single peak, like the dotplot of quiz scores in part (a) of the example, are **unimodal**. Figure 1.3 is a dotplot of the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995, when the Starnes family made its first trip to Yellowstone National Park. We describe this graph as **bimodal** because it has two clear peaks: one at about 2 minutes and one at about 4.5 minutes.

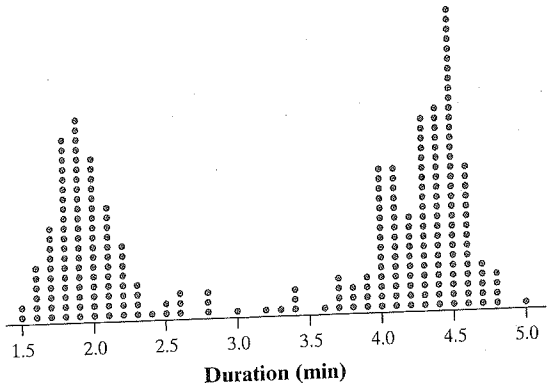


FIGURE 1.3 Dotplot displaying the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995. This distribution has two main clusters of data and two clear peaks—one near 2 minutes and the other near 4.5 minutes.

DEFINITION Unimodal, Bimodal

A distribution of quantitative data with one clear peak is **unimodal.**

A distribution of quantitative data with two clear peaks is **bimodal.**

Although we could continue the pattern with "trimodal" for three peaks, and so on, it's more common to refer to distributions with more than two clear peaks as "multimodal". When you examine a graph of quantitative data, describe any pattern you see as clearly as you can.

Describing Distributions of Quantitative Data

Here is a general strategy for describing the distribution of a quantitative variable.

HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

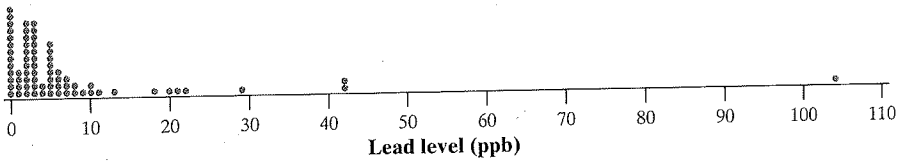
In any graph, look for the *overall pattern* and for clear *departures* from that pattern.

- You can describe the overall pattern of a distribution of quantitative data by its **shape**, **center**, and **variability**.
- An important kind of departure is an **outlier**, a value that is unusually small or unusually large relative to the rest of the data.

Variability is sometimes referred to as *spread*. We prefer variability because students often think that spread refers only to the distance between the maximum and minimum values of a quantitative data set (the *range*).

We will discuss more formal ways to measure center and variability and to identify outliers in Section 1D. For now, just use the middle value in the ordered data set (what you may have learned as the *median* in previous math classes) when describing center and the *minimum* and *maximum* values when describing variability. Visually identify any data values in the graph that are unusually small or unusually large relative to the rest of the data as possible outliers.

Let's practice with the dotplot of lead levels in water samples from 71 randomly selected Flint, Michigan, dwellings that you saw in Figure 1.2.



Shape: The distribution of lead level in these Flint, Michigan dwellings is skewed to the right, with a single peak at 0 ppb. There are noticeable gaps from 13 to 18 ppb, 22 to 29 ppb, 29 to 42 ppb, and 42 to 104 ppb.

Outliers: The lead level of 104 ppb appears to be an outlier. The two lead levels of 42 ppb are also possible outliers.

Center: The middle value (median) is a lead level of 3 ppb.

Variability: The lead levels vary from 0 to 104 ppb.

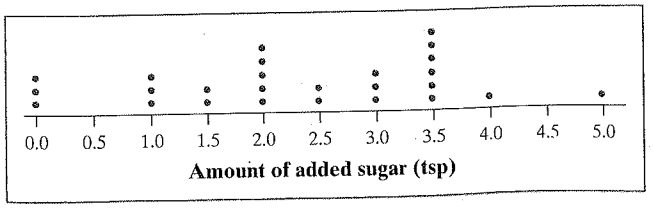
When describing a distribution of quantitative data, don't forget: **S**tatistical **O**pinions **C**an **V**ary (Shape, Outliers, Center, Variability). And always answer in context! For this example, include the variable name (lead level) and the observational units (dwellings in Flint, Michigan).

Skill 4.A

EXAMPLE Eating healthy
Describing distributions of quantitative data

PROBLEM: How healthy is plant-based yogurt? Here is a dotplot of data on the amount of added sugar, in teaspoons (tsp), in single-serving containers of several plant-based yogurt brands:[39]

Sara Stathas/Alamy Stock Photo



Describe the distribution.

SOLUTION:

Shape: The distribution of amount of added sugar for these plant-based yogurt brands is roughly symmetric, with two clear peaks at 2 tsp and 3.5 tsp of added sugar. There are small gaps between 0 and 1 tsp and between 4 and 5 tsp.

Outliers: There are no obvious outliers in the dotplot.

Center: The middle value is between 2 and 2.5 tsp of added sugar (median amount of added sugar = 2.25 tsp).

Variability: The amount of added sugar varies from 0 to 5 tsp.

AP® EXAM TIP

Always be sure to include context when you are asked to describe a distribution. This means including the observational units (plant-based yogurt brands) and the variable name (amount of added sugar), not just the units the variable is measured in (tsp).

FOR PRACTICE, TRY EXERCISE 7

CHECK YOUR UNDERSTANDING

Knoebels Amusement Park in Elysburg, Pennsylvania, has earned acclaim for being an affordable, family-friendly entertainment venue. Knoebels does not charge for general admission or parking, but it does charge customers for each ride they take. How much do the rides cost at Knoebels? The table shows the cost for each ride in a sample of 22 rides in a recent year.[40]

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

1. Make a dotplot of the data.
2. Describe the distribution.
3. A park spokesperson states that a family of four with a $100 budget can enjoy an affordable day at the park that includes at least 12 different rides together. Does the dotplot you made in Question 1 support this claim? Justify your answer.

Displaying and Describing Quantitative Data: Stem-and-Leaf Plots

Another simple type of graph for displaying quantitative data is a **stem-and-leaf plot**, also called a *stemplot*.

DEFINITION Stem-and-leaf plot

A **stem-and-leaf plot** is a graph of data for one quantitative variable that displays each data value separated into two parts: a *stem*, which consists of the leftmost digits, and a *leaf*, consisting of the final digit. The stems are ordered from least to greatest and arranged in a vertical column. The leaves are arranged in increasing order out from the appropriate stems.

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

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

Figure 1.4 shows a stem-and-leaf plot of these data.

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

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

FIGURE 1.4 Stem-and-leaf plot of the resting pulse rates of 19 middle school students.

According to the American Heart Association, a resting pulse rate greater than 100 bpm is considered high for this age group. We can see that $2/19 = 0.105 = 10.5\%$ of these students have high resting pulse rates by this standard. Also, the distribution of pulse rate for these 19 students is skewed to the right. Because stem-and-leaf plots are oriented vertically, you need to check if the skewness is toward the larger values (skewed to the right) or toward the smaller values (skewed to the left).

Stem-and-leaf plots give us a quick picture of a distribution that includes the individual data values in the graph. It is fairly easy to make a stem-and-leaf plot by hand for small sets of quantitative data.

HOW TO MAKE A STEM-AND-LEAF PLOT

1. **Make stems.** Separate each data value into a stem (all but the final digit) and a leaf (the final digit). Write the stems in a vertical column from smallest to largest. Draw a vertical line at the right of this column. Do not skip any stems, even if there is no data value for a particular stem.
2. **Add leaves.** Write each leaf in the row to the right of its stem.
3. **Order the leaves.** Arrange the leaves in increasing order out from the stem.
4. **Add a key.** Provide a key that explains in context what the stems and leaves represent.

Skills 3.A, 4.A

EXAMPLE Preventing concussions Stem-and-leaf plots

PROBLEM: Many athletes (and their parents) worry about the risk of concussions when playing sports. A youth football coach plans to obtain specially made helmets for the players that are designed to reduce their chance of getting a concussion. Here are the measurements of head circumference (in inches) for the 30 players on the team:[41]

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

Pete Saloutos/AGE Fotostock

(a) Make a stem-and-leaf plot of these data.

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

SOLUTION:

(a)

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

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

(b) The distribution of head circumference for the 30 players on the youth football team is roughly symmetric, with a single peak on the 22-inch stem. There don't appear to be any potential outliers.

To make the stem-and-leaf plot:

1. **Make stems.** The smallest head circumference is 20.8 inches and the largest is 25.6 inches. We use the first two digits as the stem and the final digit as the leaf. So, we need stems from 20 to 25.
2. **Add leaves.** For the player with a head circumference of 23.0 inches, place a 0 on the 23 stem. For the player with a head circumference of 22.2 inches, place a 2 on the 22 stem. Continue in this way until you have added the data for all the players.
3. **Order the leaves.**
4. **Add a key.**

FOR PRACTICE, TRY EXERCISE 11

We can get a better picture of the head circumference data by *splitting stems*. In Figure 1.5(a), the leaves from 0 to 9 are placed on the same stem, as in the example. Figure 1.5(b) shows another stem-and-leaf plot of the same data. This time, leaves 0 through 4 are placed on one stem, while leaves 5 through 9 are placed on another stem. Now we can see the shape of the distribution more clearly—including the possible outlier at 25.6 inches.

FIGURE 1.5 Two stem-and-leaf plots showing the head circumference data. The graph in (b) improves on the graph in (a) by splitting stems.

(a)

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

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

(b)

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

Be sure to include these stems even though they include no data.

Here are a few tips to consider before making a stem-and-leaf plot:

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

Displaying and Describing Quantitative Data: Histograms

You can use a dotplot or a stem-and-leaf plot to display quantitative data. Both of these types of graphs show every individual data value. However, for large data sets, showing each value can make it difficult to see the overall pattern in the graph. We often get a cleaner picture of the distribution by grouping nearby values together. Doing so allows us to make a new type of graph: a **histogram**.

DEFINITION Histogram

A **histogram** is a graph of data for one quantitative variable that displays each interval of values on the horizontal axis as a bar. The height of each bar shows the frequency or relative frequency of data values in that interval.

Note: You can also make a histogram with the intervals of values (also called *bins*) on the vertical axis and the bars going horizontally.

Figure 1.6 shows a dotplot and a histogram of the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995. Notice how the histogram groups nearby values together into equal-width intervals.

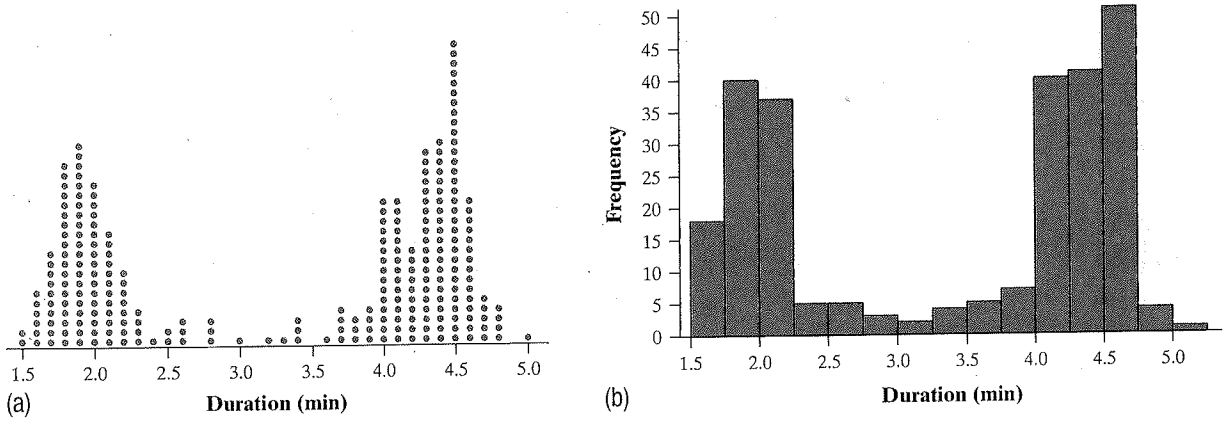


FIGURE 1.6 (a) Dotplot and (b) histogram of the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995.

You can make a histogram by hand, even for fairly large sets of quantitative data. For details on making histograms with technology, see the Tech Corner later in this section.

HOW TO MAKE A HISTOGRAM

1. **Choose equal-width intervals** that span the data. Five intervals is a good minimum.
2. **Make a table** that shows the frequency (count) or relative frequency (percentage or proportion) of data values in each interval.
3. **Draw and label the axes.** Draw horizontal and vertical axes. Put the name of the quantitative variable under the horizontal axis. To the left of the vertical axis, indicate whether the graph shows the frequency (count) or relative frequency (percentage or proportion) of data values in each interval.

4. **Scale the axes.** Place equally spaced tick marks at the smallest value in each interval along the horizontal axis until you equal or exceed the largest data value. On the vertical axis, start at 0 and place equally spaced tick marks until you equal or exceed the largest frequency or relative frequency in any interval.
5. **Draw bars** above the intervals. Make the bars equal in width and leave no gaps between them. Be sure that the height of each bar corresponds to the frequency or relative frequency of data values in that interval. An interval with no data values will appear as a bar of height 0 on the graph.

You can also choose intervals (bins) of unequal widths when making a histogram, but such graphs are beyond the scope of this book.

Skills 3.A, 4.B

EXAMPLE How much tax? Histograms

Yoon S. Byun/Portland Press Herald/Getty Images

PROBLEM: How much do sales tax rates vary in the 50 U.S. states? The table presents the data on sales tax rate (as a percentage) for each state in a recent year.[42]

State	Percentage	State	Percentage	State	Percentage
Alabama	4.00	Louisiana	4.00	Ohio	5.75
Alaska	0.00	Maine	5.50	Oklahoma	4.50
Arizona	5.60	Maryland	6.00	Oregon	0.00
Arkansas	6.50	Massachusetts	6.25	Pennsylvania	6.00
California	7.50	Michigan	6.00	Rhode Island	7.00
Colorado	2.90	Minnesota	6.88	South Carolina	6.00
Connecticut	6.35	Mississippi	7.00	South Dakota	4.00
Delaware	0.00	Missouri	4.23	Tennessee	7.00
Florida	6.00	Montana	0.00	Texas	6.25
Georgia	4.00	Nebraska	5.50	Utah	5.95
Hawaii	4.00	Nevada	6.85	Vermont	6.00
Idaho	6.00	New Hampshire	0.00	Virginia	5.30
Illinois	6.25	New Jersey	7.00	Washington	6.50
Indiana	7.00	New Mexico	5.13	West Virginia	6.00
Iowa	6.00	New York	4.00	Wisconsin	5.00
Kansas	6.50	North Carolina	4.75	Wyoming	4.00
Kentucky	6.00	North Dakota	5.00		

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

(b) Can you use the graph in part (a) to justify the claim that a majority of states have a sales tax rate of at least 5%? Explain your reasoning.

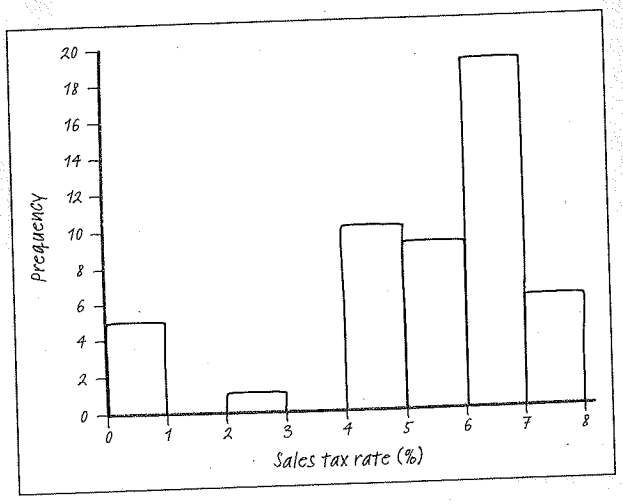
SOLUTION:

(a)

Interval	Frequency
0 to < 1	5
1 to < 2	0
2 to < 3	1
3 to < 4	0
4 to < 5	10
5 to < 6	9
6 to < 7	19
7 to < 8	6

To make the histogram:

1. **Choose equal-width intervals** that span the data. The data vary from 0.00% to 7.50%. We choose intervals of width 1, beginning at 0: 0 to < 1, 1 to < 2, and so on. This choice results in more than the minimum of five intervals.
2. **Make a table.** Record the number of observational units (states) with data values in each interval when making a frequency histogram.



3. **Draw and label the axes.**
4. **Scale the axes.** The scale on the horizontal axis matches the intervals we chose in Step 1. The highest frequency in an interval is 19, so we scale the vertical axis from 0 to 20, placing tick marks every 2 units.
5. **Draw bars.**

(b) In the histogram, $(9+19+6)/50 = 34/50 = 0.68$ of states have a sales tax rate of at least 5%. Because $0.68 > 0.50$, we can use the graph to justify the claim that a majority of states have a sales tax rate of at least 5%.

FOR PRACTICE, TRY EXERCISE 17

From the histogram, we can see that the distribution of sales tax rate in the 50 U.S. states is skewed to the left and unimodal, with a single peak in the 6% to < 7% interval. The states with a sales tax rate of 0.00% appear to be outliers. There are two gaps in the graph, indicating no states with a sales tax rate from 1% to < 2% or from 3% to < 4%. How should we describe the center and variability of the distribution? Because a histogram does not show individual data values, we can only give *estimates* of the center and variability using the intervals on the graph. With 50 data values, the middle value (median) falls between the 25th and 26th values in the ordered data set, so the median sales tax rate is about 6%. The state sales tax rates vary from at least 0% to at most 7.99%. Using the raw data from the example, we can confirm that the median is 5.975% and that the data vary from 0.00% to 7.50%.

Note that the convention in this book is to include the left endpoint of an interval and exclude the right endpoint when making histograms. For instance, Alabama's sales tax rate of 4.0% was placed in the 4 to < 5 bar, not the 3 to < 4 bar, in the preceding example.

Figure 1.7 shows two different histograms of the state sales tax rate data. The one on the left (a) uses the intervals of width 1 from the example. The distribution has a single peak in the 6 to <7 interval. The one on the right (b) uses intervals that are half as wide: 0 to <0.5, 0.5 to <1, and so on. Now we see a distribution with more than one clear peak. **The choice of intervals in a histogram can affect the appearance of a distribution.** Histograms with more intervals show more detail but may have a less clear overall pattern.

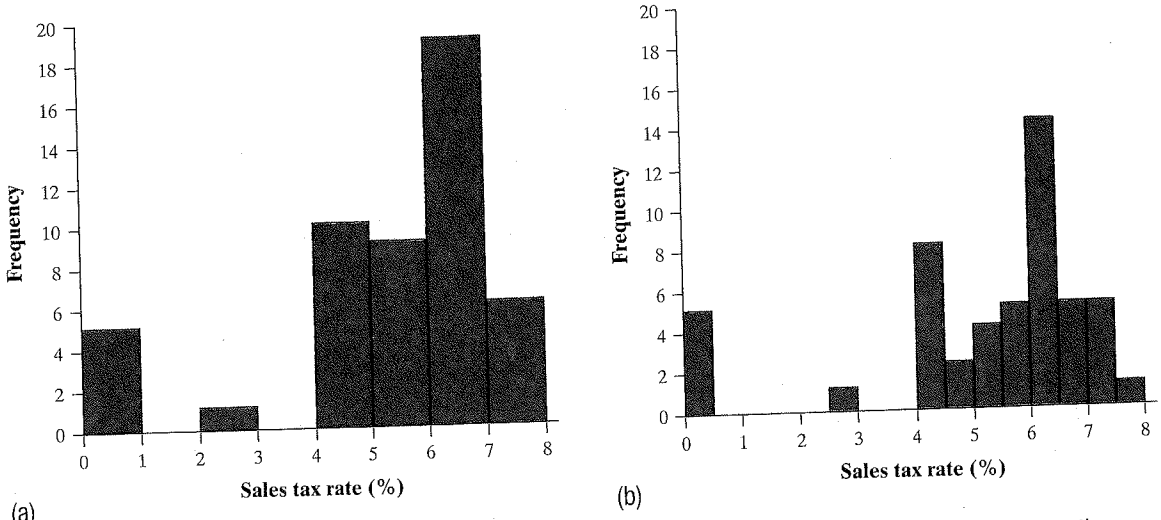
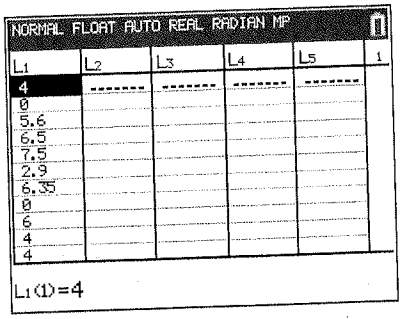


FIGURE 1.7 (a) Frequency histogram of the sales tax rate (%) in the 50 states with intervals of width 1, from the previous example. (b) Frequency histogram of the data with intervals of width 0.5.

1. Tech Corner MAKING HISTOGRAMS

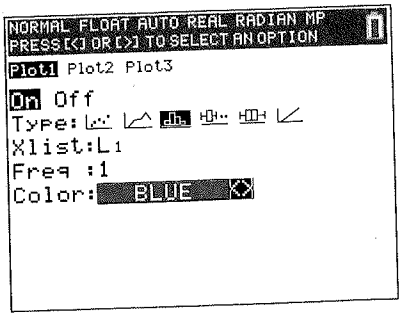
Instructions for other technology, including Desmos, NumWorks, and the latest TI-84 calculator, can be found on Achieve or your teacher can share this content with you.

You can use technology to make a histogram. The technology's default choice of intervals is a good starting point, but you should adjust the intervals to fit with common sense. We'll illustrate using data on the sales tax rate in the 50 U.S. states from the preceding example.

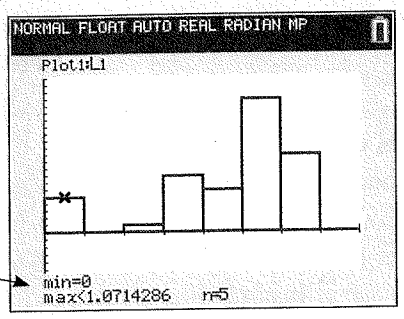


1. Press [STAT] and choose Edit..., then type the values into list L1.

2. Press [2nd] [Y=] (STAT PLOT), press [ENTER] or [1] to go into Plot1, and adjust the settings as shown.

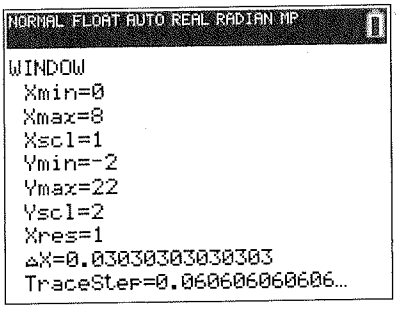


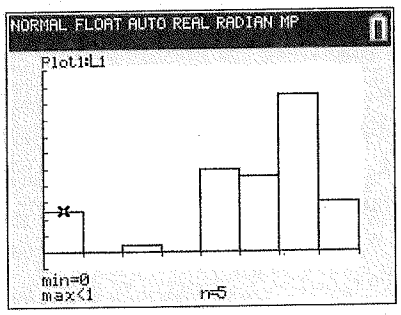
3. Press [ZOOM] and choose ZoomStat. Press [TRACE] and use the left and right arrow keys to examine the intervals.



Note the calculator's unusual choice of intervals.

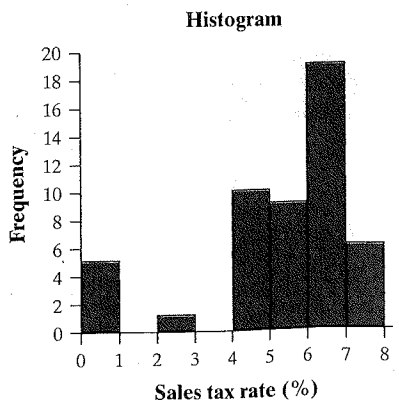
4. To adjust the intervals to match those in Figure 1.7(a), press [WINDOW] and enter the values shown for Xmin, Xmax, Xscl, Ymin, Ymax, and Yscl. Then press [GRAPH]. Press [TRACE] and use the left and right arrow keys to examine the intervals.

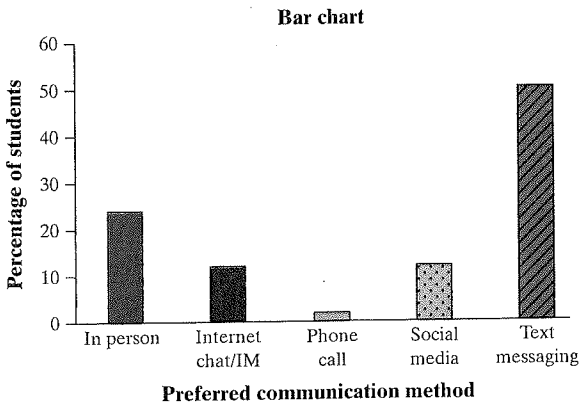




If you're asked to make a graph, be sure to label and scale your axes. Don't just transfer what you see on a calculator screen to your paper and expect to earn full credit.

Don't confuse histograms and bar charts. Although histograms resemble bar charts, their details and uses are different. A histogram displays the distribution of a *quantitative variable*. Its horizontal axis identifies intervals of values that the variable can take. A bar chart displays the distribution of a *categorical variable*. Its horizontal axis identifies the categories. Be sure to draw bar charts with blank space between the bars to separate the categories. There should only be space between the bars in a histogram if there are no data values in one or more intervals (bins). For comparison, here is one of each type of graph from earlier examples:





CHECK YOUR UNDERSTANDING

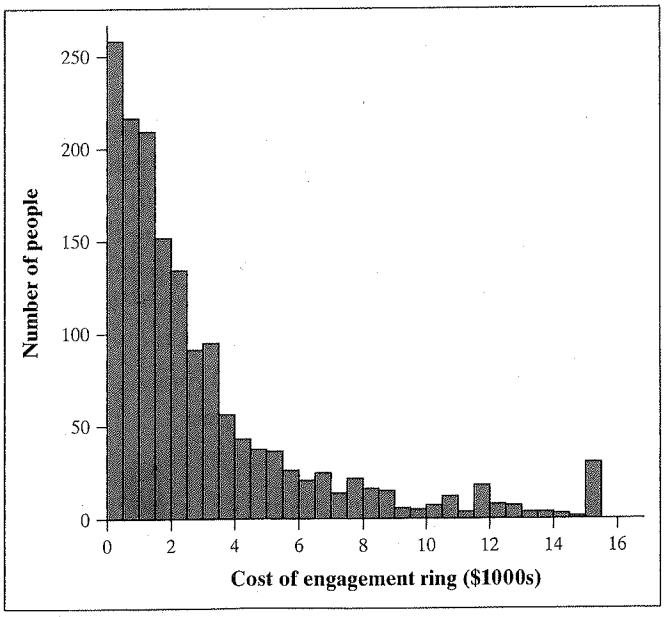
Questions 1–3 refer to the following setting. Here is a stem-and-leaf plot of the percentage of residents aged 65 and older in each of the 50 U.S. states:[43]

11	18
12	6
13	9
14	23
15	44445677889999
16	23444557999
17	0112225567
18	12477
19	39
20	56

Key: 19|3 means 19.3% of the state's residents are aged 65 and older.

1. The largest data value is for Maine. What percentage of Maine residents are aged 65 or older? (Note that Florida has the second largest data value.)
2. Describe the distribution.
3. Make another stem-and-leaf plot of the data by splitting stems. What does this new graph reveal that is not apparent from the original stem-and-leaf plot?

Questions 4–6 refer to the following scenario. When getting married, some people choose to buy an engagement ring for their partner. How much do people spend on engagement rings, on average? To find out, *The New York Times* and *Morning Consult* surveyed a random sample of 1640 U.S. adults who bought an engagement ring. Here is a histogram of the data:[44]



4. About what percentage of people reported spending at least $2000 on an engagement ring? Show your method clearly.
5. Describe the shape of the distribution.
6. A marketing campaign claims that U.S. adults typically spend at least $5000 on an engagement ring. Do these data support the claim? Justify your answer.

SECTION 1C Summary

- You can use a **dotplot**, **stem-and-leaf plot**, or **histogram** to display the distribution of a quantitative variable. A dotplot displays individual data values on a number line. Stem-and-leaf plots separate each data value into a stem and a one-digit leaf. Histograms plot the frequencies (counts) or relative frequencies (proportions or percentages) of data values in equal-width intervals.
- When examining any graph of quantitative data, look for an *overall pattern* and for clear *departures* from that pattern. **Shape**, **center**, and **variability** describe the overall pattern of the distribution of a quantitative variable. **Outliers** are observations that are unusually small or unusually large relative to the rest of the data values.
- Some distributions have simple shapes, such as **roughly symmetric**, **skewed to the left**, or **skewed to the right**. A distribution in which the frequency (relative frequency) of each possible value is about the same is **approximately uniform**.
- The number of peaks is another aspect of overall shape. So are distinct *clusters* and *gaps*. A graph of quantitative data with one clear peak is called **unimodal**, and a graph with two clear peaks is called **bimodal**.
- Histograms are for quantitative data; bar charts are for categorical data. In both types of graphs, be sure to use relative frequencies when comparing data sets of different sizes.

AP® EXAM TIP
AP® Classroom

Review the content of this section with the resources in AP® Classroom for Topics 1.5 and 1.6.

1C Tech Corner

Instructions for other technology, including Desmos, NumWorks, and the latest TI-84 calculator, can be found on Achieve or your teacher can share this content with you.

1. Making histograms Page 40

SECTION 1C Exercises

Displaying and Describing Quantitative Data: Dotplots

1. **Women's soccer** How good was the 2019 U.S. women's soccer team? With players like Carli Lloyd, Alex Morgan, and Megan Rapinoe, the team put on an impressive showing en route to winning the 2019 Women's World Cup. Here are data on the number of goals scored by the team in games played in the 2019 season:[45] pg 29

1	1	2	2	1	5	6	3	5	3	13	3
2	2	2	2	2	3	4	3	2	1	3	6

(a) Make a dotplot of these data.

(b) In what proportion of games did the team score 4 or more goals?

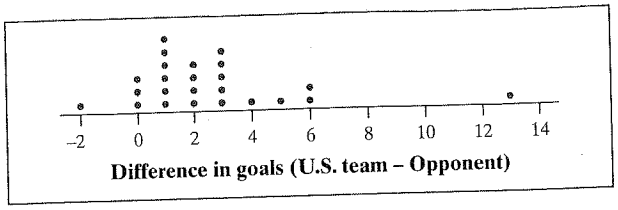
2. **Fuel efficiency** The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy ratings for cars. Here are the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 21 model-year 2022 midsize cars:[46]

25	30	27	31	38	26	28	40	25	28	30
31	30	30	34	30	31	31	32	48	31	

(a) Make a dotplot of these data.

(b) What percentage of the car models in the sample get more than 35 mpg on the highway?

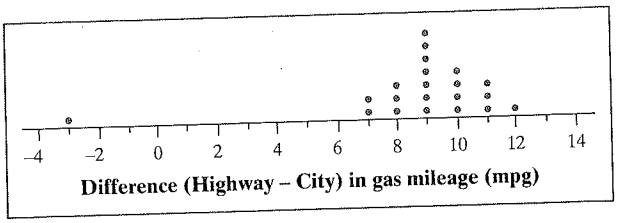
3. **More women's soccer** The following dotplot shows the difference in the number of goals scored (U.S. women's team – Opponent) in each game from Exercise 1.



(a) Explain what the dot above −2 represents.

(b) Can we use the graph in part (a) to justify the claim that the team did very well in 2019? Explain your answer.

4. **Better fuel efficiency** The following dotplot shows the difference in EPA mileage ratings (Highway – City) in miles per gallon (mpg) for each of the 21 model-year 2022 midsize cars from Exercise 2.



(a) The dot above −3 is for the Toyota Prius. Explain what this dot represents.

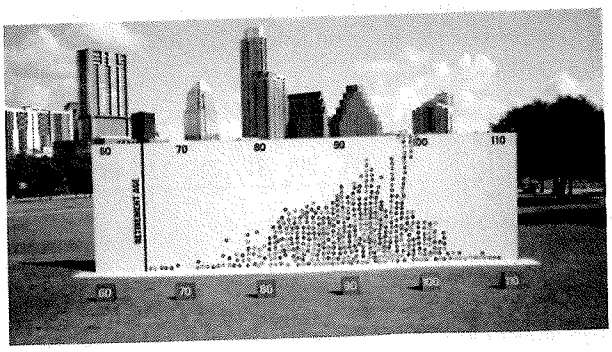
(b) Can we use the graph in part (a) to justify the claim that most of these model-year 2022 cars get better gas mileage on the highway than in the city? Explain your answer.

Describing the Shape of a Graph of Quantitative Data

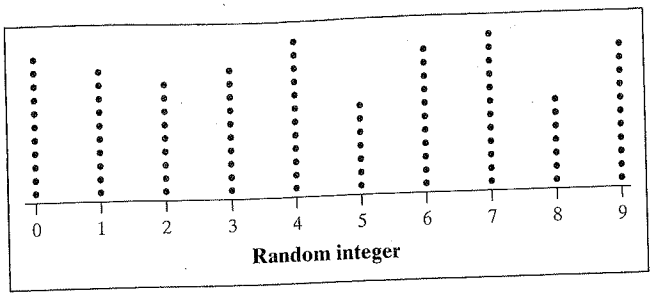
5. **Getting older and random digits**

pg 31

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

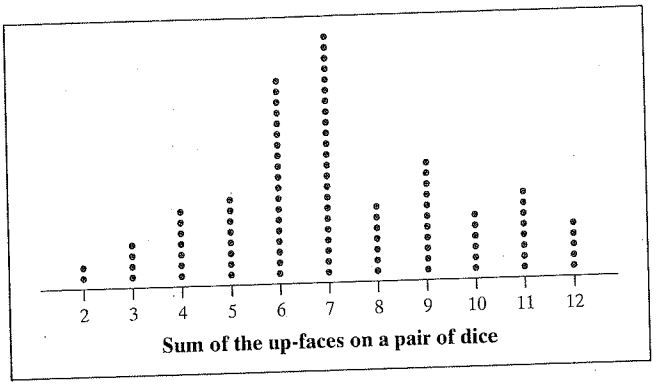


(b) The dotplot displays the results of using a random number generator to produce 100 digits between 0 and 9, inclusive. Describe the shape of the distribution.

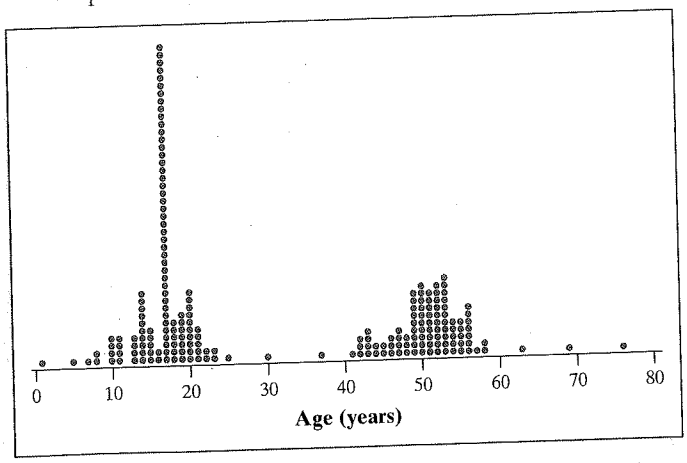


6. **Roll the dice and family ages**

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



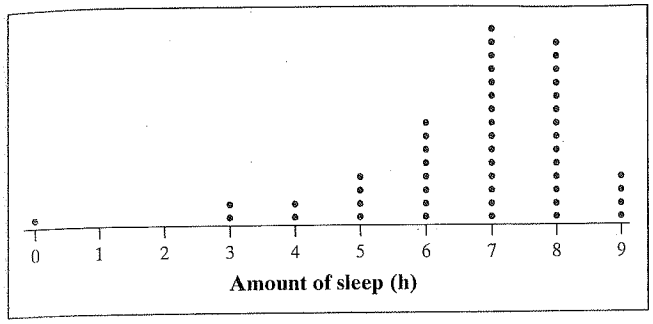
(b) Statistics instructor Paul Myers collected data on the ages (in years) of family members for each student in his class. The dotplot displays the data. Describe the shape of the distribution.



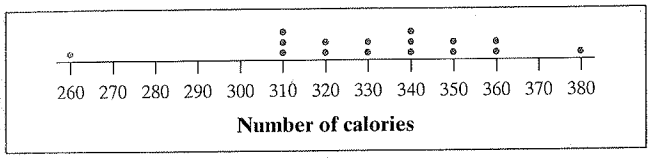
Describing Distributions of Quantitative Data

7. **Feeling sleepy?** A statistics professor asked how much sleep (in hours) students got on the night prior to their first exam. Here is a dotplot of the data from the 50 students in the class. Describe the distribution.

pg 33



8. **Frozen pizza** *Consumer Reports* collected data on the number of calories per serving for 16 brands of frozen cheese pizza as part of its product reviews. Here is a dotplot of the data.[47] Describe the distribution.



9. **Soccer distribution** Refer to Exercise 1.

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

(b) How many goals did the team score in a typical game that season? Explain your answer.

10. **Fuel distribution** Refer to Exercise 2.

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

(b) What is the typical fuel efficiency of the 21 cars in the sample? Explain your answer.

Displaying and Describing Quantitative Data: Stem-and-Leaf Plots

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

pg 36

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

(a) Make a stem-and-leaf plot of these data. What interesting feature does the graph reveal?

(b) The advertised weight of a Snickers Fun Size bar is 17 grams. Can you use the graph from part (a) to justify the claim that most of the candy bars in this sample weigh less than advertised? Explain your answer.

12. **Eat your beans!** Beans and other legumes are a great source of protein. The following data give the protein content of 30 different varieties of beans, in grams per 100 grams of cooked beans.[48]

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

(a) Make a stem-and-leaf plot of these data. What interesting feature does the graph reveal?

(b) Bean varieties with at least 9 grams of protein per 100 grams of cooked beans are classified as "high protein." Can you use the graph in part (a) to justify the claim that a majority of these varieties are "high protein"? Explain your answer.

13. **South Carolina counties** Here is a stem-and-leaf plot of the areas of the 46 counties in South Carolina.[49] Note that the data have been rounded to the nearest 10 square miles (mi^2).

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

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

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

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

14. **Shopping spree** The stem-and-leaf plot displays data on the amount spent by 50 shoppers at a grocery store. Note that the values have been rounded to the nearest dollar.

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

Key: 9|3 represents a shopper at this store who spent $92.50 to $93.49.

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

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

15. **Arizona heat** Here is a stem-and-leaf plot of the high temperature readings (in degrees Fahrenheit) for Phoenix, Arizona, for each day in July in a recent year:[50]

8	4
8	
9	3
9	799
10	011223444
10	556667788999
11	0113
11	5

(a) Why did we split the stems?

(b) Give an appropriate key for this stem-and-leaf plot.

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

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

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

(a) Why did we split the stems?

(b) Give an appropriate key for this stem-and-leaf plot.

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

Displaying and Describing Quantitative Data: Histograms

17. **Carbon dioxide emissions** Burning fuels in power plants and motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. The table displays CO_2 emissions in metric tons per person from 48 countries with populations of at least 20 million in a recent year.[51]

pg 38

Country	CO_2	Country	CO_2	Country	CO_2
Algeria	4.0	Italy	5.6	South Africa	8.2
Argentina	4.0	Japan	8.7	Spain	5.4
Australia	16.3	Kenya	0.3	Sudan	0.5
Bangladesh	0.6	Korea, North	1.5	Tanzania	0.2
Brazil	2.2	Korea, South	11.9	Thailand	4.1
Canada	15.4	Malaysia	7.8	Turkey	4.9
China	7.1	Mexico	3.4	Ukraine	5.1
Colombia	2.0	Morocco	2.0	United Kingdom	5.5
Congo	0.6	Myanmar	0.5	United States	16.1
Egypt	2.5	Nepal	0.5	Uzbekistan	3.3
Ethiopia	0.2	Nigeria	0.7	Venezuela	4.1
France	5.0	Pakistan	1.2	Vietnam	2.6
Germany	8.4	Peru	1.7		
Ghana	0.5	Philippines	1.3		
India	1.9	Poland	8.5		
Indonesia	2.3	Romania	3.9		
Iran	9.4	Russia	11.5		
Iraq	5.6	Saudi Arabia	17.0		

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

(b) What proportion of these countries had CO_2 emissions of at least 10 metric tons per person?

18. **Traveling to work** How long do people travel each day to get to work? The following table gives the average travel times to work (in minutes) in a recent year for workers in each state and the District of Columbia who are at least 16 years old and don't work at home.[52]

State	Travel time to work (min)	State	Travel time to work (min)	State	Travel time to work (min)
AL	24.9	KY	23.6	ND	17.3
AK	19.1	LA	25.7	OH	23.7
AZ	25.7	ME	24.2	OK	21.9
AR	21.7	MD	33.2	OR	23.9
CA	29.8	MA	30.2	PA	27.2
CO	25.8	MI	24.6	RI	25.2
CT	26.6	MN	23.7	SC	25.0
DE	26.3	MS	24.8	SD	17.2
DC	30.8	MO	23.9	TN	25.2
FL	27.8	MT	18.3	TX	26.6
GA	28.8	NE	18.8	UT	21.9
HI	27.5	NV	24.6	VT	23.3
ID	21.1	NH	27.5	VA	28.7
IL	29.2	NJ	32.2	WA	28.0
IN	23.8	NM	22.3	WV	25.9
IA	19.3	NY	33.6	WI	22.2
KS	19.4	NC	24.8	WY	17.9

(a) Make a histogram to display the travel time data using intervals of width 2 minutes, starting at 16 minutes.

(b) In what proportion of states plus the District of Columbia is the average travel time at least 20 minutes?

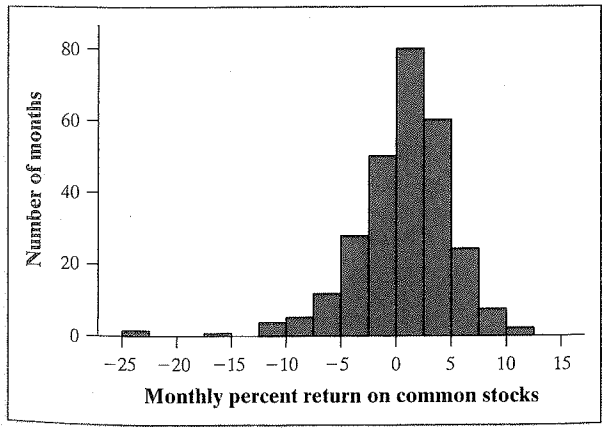
19. **Home runs** Here are the number of home runs hit by each of the 30 Major League Baseball teams in a recent season.[53] Make a histogram that effectively displays the distribution of number of home runs hit. Describe the distribution.

220	249	213	245	256	182	227	223	224	149
288	162	220	279	146	250	307	242	306	257
215	163	219	239	167	210	217	223	247	231

20. **Country music** Here are the lengths, in minutes, of 50 songs by country artist Dierks Bentley. Make a histogram that effectively displays the distribution of song length. Describe the distribution.

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

21. **Returns on common stocks** The return on a stock is the change in its market price plus any dividend payments made. Return is usually expressed as a percentage of the beginning price. The figure shows a histogram of the distribution of monthly percent return for the U.S. stock market (total return on all common stocks) in 273 consecutive months.[54]

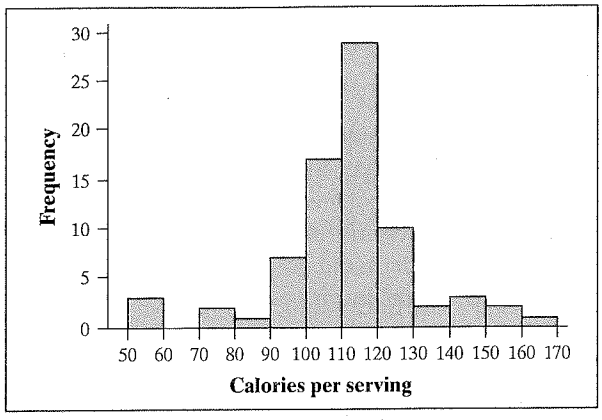


(a) A return less than zero means that stocks lost value in that month. An analyst claims that the U.S. stock market lost value in a majority of these 273 months. Use the graph to determine whether this claim is valid. Explain your reasoning.

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

(c) Estimate the center and variability of the distribution.

22. **Healthy cereal?** Researchers collected data on calories per serving for 77 brands of breakfast cereal at a local supermarket. The histogram displays the data.[55]



(a) A group of health-conscious people want to avoid breakfast cereals with 130 or more calories per serving, but claim that such cereals are difficult to find. Use the graph to determine whether this claim is valid for this supermarket. Explain your reasoning.

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

(c) Estimate the center and variability of the distribution.

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

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

Connecting the Statistical Practices *Multi-focus exercises that connect two or more statistical practices.*

25. **Do the reading!** A college professor usually assigns pre-class reading to the students in their introductory statistics class. The professor suspects that most students did not complete the reading assignment prior to a Monday morning class. To find out, the professor gives a 10-question pop quiz about the assigned reading to a random sample of 50 of the 200 students in the class. The number of questions each student answers correctly is recorded.

(a) Determine a valid investigative question for this statistical study.

(b) Identify the observational units and the variable of interest. Classify the variable as categorical or quantitative.

(continued on next page)

Here are the number of correct answers on the pop quiz for the 50 randomly selected students:

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

(c) Make an appropriate graph of the data.

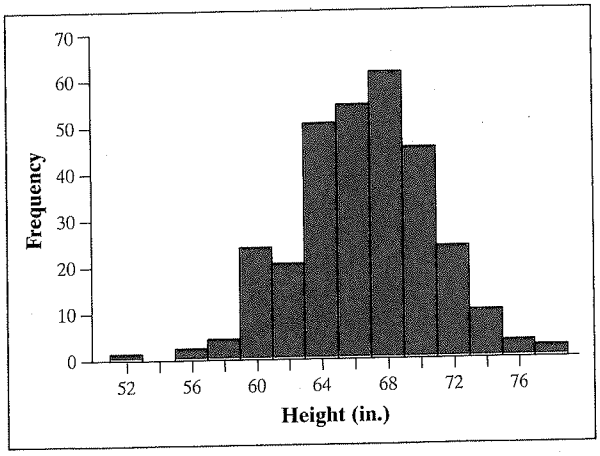
(d) Students who complete the pre-class reading generally get at least 7 correct answers on the quiz. What proportion of students got fewer than 7 correct answers on the pop quiz? Does this result support the professor's belief that most students did not complete the pre-class reading assignment? Justify your answer.

Multiple Choice *Select the best answer for each question.*

26. Here are the amounts of money (cents) in coins carried by 10 students in a statistics class: 50, 35, 0, 46, 86, 0, 5, 47, 23, 65. Which of the following sets of stems should be used to make a stem-and-leaf plot of these data?

(A) 0, 2, 3, 4, 5, 6, 8
(B) 0, 1, 2, 3, 4, 5, 6, 7, 8
(C) 0, 3, 5, 6, 7
(D) None of these

27. The histogram shows the heights (in inches) of 300 randomly selected high school students. Which of the following is the best description of the shape of the distribution of height?

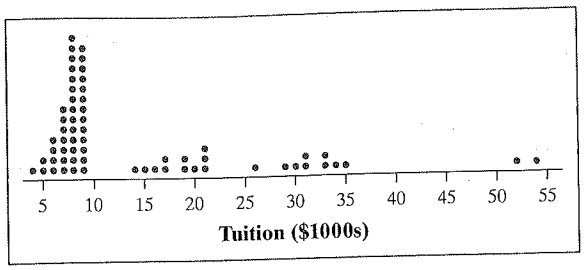


(A) Roughly symmetric and unimodal (single-peaked)
(B) Roughly symmetric and bimodal (double-peaked)
(C) Skewed to the left
(D) Skewed to the right

28. Which of the following is the best reason for choosing a stem-and-leaf plot rather than a histogram to display the distribution of a quantitative variable?

(A) Stem-and-leaf plots are better for displaying very large sets of data.
(B) Stem-and-leaf plots make it easier to determine the shape of a distribution.
(C) Stem-and-leaf plots allow you to split stems; histograms don't.
(D) Stem-and-leaf plots allow you to see individual data values.

29. The dotplot shows the tuition (to the nearest $1000) for the 63 largest colleges and universities in North Carolina in a recent year.[56]



Which of the following statements about the dotplot is not correct?

(A) There are more North Carolina colleges and universities with tuitions greater than $10,000 than with tuitions less than $10,000.
(B) The tuitions vary from about $4000 to about $54,000.
(C) There are two obvious potential outliers—institutions with tuitions greater than $50,000.
(D) The distribution of tuition is skewed to the right.

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

30. **Five-way stem splitting** Sometimes, the variability in a data set is so small that splitting stems in two doesn't produce a stem-and-leaf plot that shows the shape of the distribution well. We can often solve this problem by splitting the stem into five parts, each consisting of two leaf values: 0 and 1, 2 and 3, 4 and 5, and so on.

Here are the weights, in ounces, of 36 navel oranges selected from a large shipment to a grocery store.

5.7	5.4	5.8	5.3	4.6	4.9	5.6	5.3	5.5	5.5	5.4	5.8
5.3	5.5	5.5	5.4	5.8	5.9	5.4	5.1	5.0	5.5	5.7	4.9
5.0	5.3	5.1	5.2	5.7	5.6	5.8	4.5	5.2	5.4	5.7	5.6

Make a stem-and-leaf plot of the data by splitting stems into five parts. Describe the shape of the distribution.

31. **Changing bin widths** Refer to Exercise 17.

(a) Use technology to make a histogram of the data using intervals (bins) of width 4. How does this change the appearance of the distribution of CO_2 emissions from the graph you made in Exercise 17?

(b) Use technology to make a histogram of the data using intervals (bins) of width 1. How does this change the appearance of the distribution of CO_2 emissions from the graph you made in Exercise 17?