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

Introduction Summary

- A data set contains information about a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- Some variables are **categorical** and others are **quantitative.** A categorical variable assigns a label that places each individual into one of several groups, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or salary in dollars.
- The **distribution** of a variable describes what values the variable takes and how often it takes them.

Introduction Exercises

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

1. **Protecting wood** How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables recorded: type of wood (yellow poplar, pine, cedar); type of water repellent (solvent-based, water-based); paint thickness (millimeters); paint color (white, gray, light blue); weathering time (months). Identify each variable as categorical or quantitative.

2. **Medical study variables** Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Here are some of the variables recorded: gender (female or male); age (years); race (Asian, black, white, or other); smoker (yes or no); systolic blood pressure (millimeters of mercury); level of calcium in the blood (micrograms per milliliter). Identify each as categorical or quantitative.

3. **A class survey** Here is a small part of the data set that pg 3 describes the students in an AP® Statistics class. The data come from anonymous responses to a questionnaire filled out on the first day of class.

Gender	Hand	Height (in.)	Homework time (min)	Favorite music	Pocket change (cents)
F	L	65	200	Hip-hop	50
M	L	72	30	Country	35
M	R	62	95	Rock	35
F	L	64	120	Alternative	0
M	R	63	220	Hip-hop	0
F	R	58	60	Alternative	76
F	R	67	150	Rock	215

(a) What individuals does this data set describe?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

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

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (s)
Wild Mouse	Steel	49.3	Sit down	28	70
Terminator	Wood	95	Sit down	50.1	180
Manta	Steel	140	Flying	56	155
Prowler	Wood	102.3	Sit down	51.2	150
Diamondback	Steel	230	Sit down	80	180

(a) What individuals does this data set describe?

(b) What variables were measured? Identify each as categorical or quantitative.

(c) Describe the individual in the highlighted row.

5. **Ranking colleges** Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe two categorical variables and two quantitative variables that you might record for each institution.

6. **Students and TV** You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student.

Multiple choice: Select the best answer.

Exercises 7 and 8 refer to the following setting. At the Census Bureau Web site www.census.gov, you can view detailed data collected by the American Community Survey. The following table includes data for 10 people chosen at random from the more than 1 million people in households contacted by the survey. "School" gives the highest level of education completed.

Weight (lb)	Age (yr)	Travel to work (min)	School	Gender	Income last year ($)
187	66	0	Ninth grade	1	24,000
158	66	n/a	High school grad	2	0
176	54	10	Assoc. degree	2	11,900
339	37	10	Assoc. degree	1	6000
91	27	10	Some college	2	30,000
155	18	n/a	High school grad	2	0
213	38	15	Master's degree	2	125,000
194	40	0	High school grad	1	800
221	18	20	High school grad	1	2500
193	11	n/a	Fifth grade	1	0

7. The individuals in this data set are

(a) households.
(b) people.
(c) adults.
(d) 120 variables.
(e) columns.

8. This data set contains

(a) 7 variables, 2 of which are categorical.
(b) 7 variables, 1 of which is categorical.
(c) 6 variables, 2 of which are categorical.
(d) 6 variables, 1 of which is categorical.
(e) None of these.

1.1 Analyzing Categorical Data

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Display categorical data with a bar graph. Decide if it would be appropriate to make a pie chart.
- Identify what makes some graphs of categorical data deceptive.
- Calculate and display the marginal distribution of a categorical variable from a two-way table.
- Calculate and display the conditional distribution of a categorical variable for a particular value of the other categorical variable in a two-way table.
- Describe the association between two categorical variables by comparing appropriate conditional distributions.

The values of a categorical variable are labels for the categories, such as "male" and "female." The distribution of a categorical variable lists the categories and gives either the *count* or the *percent* of individuals who fall within each category. Here's an example.