# PART I

# Producing Data


Richard B. Levine/Newscom

You and your friends are not typical. What you listen to on the radio or on music streaming sites like Pandora is probably not what we listen to. Of course, we and our friends are also not typical. To get a true picture of the country as a whole (or even of college students), we must recognize that the picture may not resemble us or what we see around us. We need *data*. Data from Nielsen (a consumer research firm) for April 15, 2015, show that the most popular radio formats for millennials (ages 18–34) are pop contemporary hit radio (12.8% of listeners) and country (9.0% of listeners). If you like hot adult contemporary (7.5% of listeners) and we like classic rock (4.8% of listeners), we may have no clue about the tastes of radio audiences as a whole. If we are in the broadcasting business, or even if we are interested in pop culture, we must put our own tastes aside and look at the data.

You can find data in the library or on the Internet (that's where we found the radio format data). But how can we know whether data can be trusted? Good data are as much a human product as wool sweaters and tablet PCs. Sloppily produced data will frustrate you as much as a sloppily made sweater. You examine a sweater before you buy, and you don't buy if it is not well made. Neither should you use data that are not well made. The first part of this book shows how to tell if data are well made.

# Where Do Data Come From?

1


Sacramento Bee/Contributor/Getty Images

**CASE STUDY** You can read the newspaper and watch TV news for months without seeing an algebraic formula. No wonder algebra seems unconnected to life. You can't go a day, however, without meeting data and statistical studies. You hear that last month's unemployment rate was 5.4%. A news article says that 88% of AAAS (the American Association for the Advancement of Science) scientists agree that it is generally safe to eat genetically modified foods, while only 37% of all U.S. residents 18 years of age and over agree. A longer article says that low-income children who received high-quality day care did better on academic tests given years later and were more likely to go to college and hold good jobs than other similar children.

Where do these data come from? Why can we trust them? Or maybe we can't trust them. Good data are the fruit of intelligent human effort. Bad data result from laziness or lack of understanding, or even the desire to mislead others. "Where do the data come from?" is the first question you should ask when someone throws a number at you.

In 2012, Colorado voters legalized marijuana. Subsequently, voters in several other states have considered legalizing marijuana. One of these is Michigan. In February 2014, the Michigan online news site MLive ran the story "Take our online poll: Should Michigan legalize marijuana?" Of 9684 respondents, 7906 (81.64%) said Yes, 1190 (12.29%) said No, and 588 (6.07%) said Decriminalize but not legalize. These results would seem to indicate overwhelming support for legalizing marijuana in Michigan.

What can we say about data from this poll? By the end of this chapter you will have learned some basic questions to ask about the data from the MLive online poll. The answers to these questions will help us assess whether the data from the poll are good or bad, as we will explore further in Chapter 2.

## Talking about data: Individuals and variables

Statistics is the science of data. We could almost say "the art of data" because good judgment and even good taste, along with good math, make good statistics. A big part of good judgment lies in deciding what you must measure in order to produce data that will shed light on your concerns. We begin with some vocabulary to describe the raw materials that go into data.

### Individuals and variables

**Individuals** are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

For example, here are the first lines of a professor's data set at the end of a statistics course:

| NAME | MAJOR | POINTS | GRADE |
|---|---|---|---|
| ADVANI, SURA | COMM | 397 | B |
| BARTON, DAVID | HIST | 323 | C |
| BROWN, ANNETTE | LIT | 446 | A |
| CHIU, SUN | PSYC | 405 | B |
| CORTEZ, MARIA | PSYC | 461 | A |

The *individuals* are students enrolled in the course. In addition to each student's name, there are three *variables*. The first says what major a student has chosen. The second variable gives the student's total points out of 500 for the course, and the third records the grade received.

Statistics deals with numbers, but not all variables are numerical. Some are "categorical" and simply place an individual into one of several groups or categories. Of the three variables in the professor's data set, only total points has numbers as its values. Major and grade are categorical, and to do statistics with these variables, we use *counts* or *percentages*. We might give the percentage of students who got an A, for example, or the percentage who are psychology majors.

### Categorical and numerical variables

A **categorical variable** simply places an individual into one of several groups or categories.

A **numerical variable** takes numerical values for which arithmetic operations such as adding and averaging make sense. A numerical variable is sometimes referred to as a **quantitative variable**.

Bad judgment in choosing variables can lead to data that cost lots of time and money but don't shed light on the world. What constitutes good judgment can be controversial. Here are examples of the challenges in deciding what data to collect.

### EXAMPLE 1 Who recycles?

Who takes the trouble to recycle? Researchers spent lots of time and money weighing the stuff put out for recycling in two neighborhoods in a California city; call them Upper Crust and Lower Mid. The *individuals* here are households because trash and recycling pickup are done for residences, not for people one at a time. The *variable* measured was the weight in pounds of the curbside recycling basket each week.

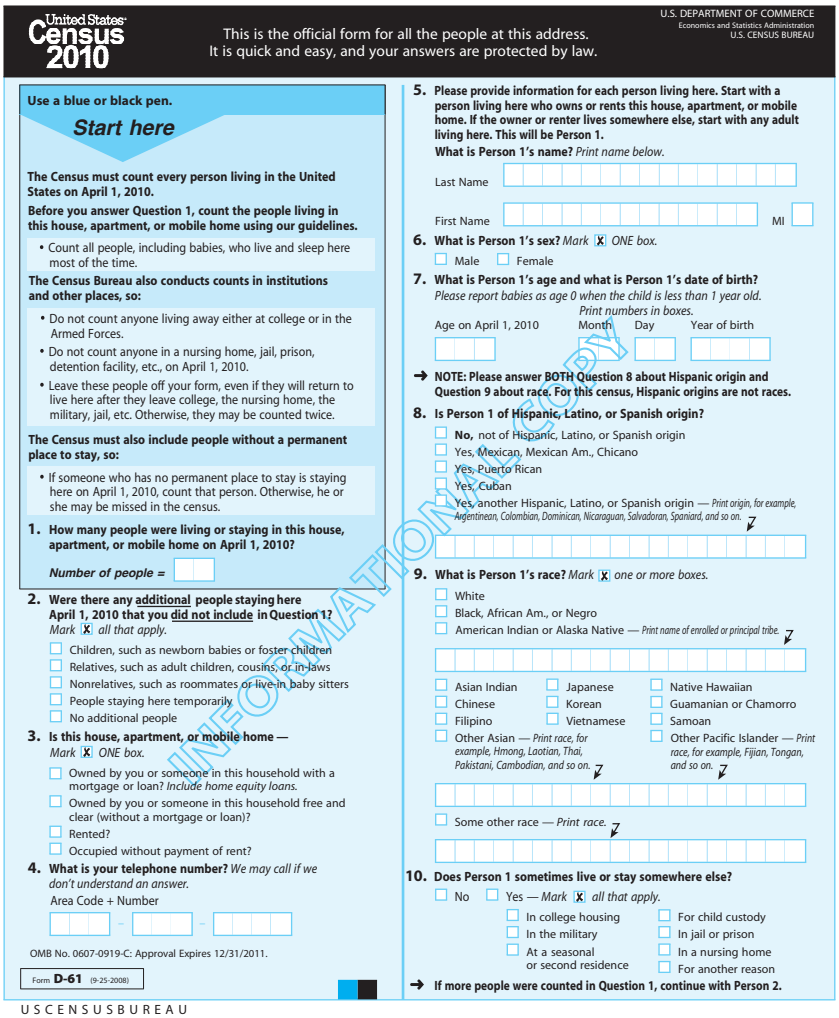
The Upper Crust households contributed more pounds per week on the average than did the folk in Lower Mid. Can we say that the rich are more serious about recycling? No. Someone noticed that Upper Crust recycling baskets contained lots of heavy glass wine bottles. In Lower Mid, they put out lots of light plastic soda bottles and light metal beer and soda cans. The conclusion: weight tells us little about commitment to recycling.

### EXAMPLE 2 What's your race?

The U.S. census asks, "What is this person's race?" for every person in every household. "Race" is a *variable*, and the Census Bureau must say exactly how to measure it. The census form does this by giving a list of races. Years of political squabbling lie behind this list.

How many races shall we list, and what names shall we use for them? Shall we have a category for people of mixed race? Asians wanted more national categories, such as Filipino and Vietnamese, for the growing Asian population. Pacific Islanders wanted to be separated from the larger Asian group. Black leaders did not want a mixed-race category, fearing that many blacks would choose it and so reduce the official count of the black population.

The 2010 census form (see Figure 1.1) ended up with six Asian groups (plus "Other Asian") and three Pacific Island groups (plus "Other Pacific Islander"). There is no "mixed-race" group, but you can mark more than one race. That is, people claiming mixed race can count as both so that the total of the racial group counts in 2010 is larger than the population count. Unable to decide what the proper term for blacks should be, the Census Bureau settled on "Black, African American, or Negro." What about Hispanics? That's a separate question because



United States Census 2010

This is the official form for all the people at this address.
It is quick and easy, and your answers are protected by law.

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

Use a blue or black pen.

Start here

The Census must count every person living in the United States on April 1, 2010.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2010.
- Leave these people off your form, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

The Census must also include people without a permanent place to stay, so:

- If someone who has no permanent place to stay is staying here on April 1, 2010, count that person. Otherwise, he or she may be missed in the census.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?
Number of people =

2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1?
Mark [X] all that apply.
☐ Children, such as newborn babies or foster children
☐ Relatives, such as adult children, cousins, or in-laws
☐ Nonrelatives, such as roommates or live-in baby sitters
☐ People staying here temporarily
☐ No additional people

3. Is this house, apartment, or mobile home —
Mark [X] ONE box.
☐ Owned by you or someone in this household with a mortgage or loan? Include home equity loans.
☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
☐ Rented?
☐ Occupied without payment of rent?

4. What is your telephone number? We may call if we don't understand an answer.
Area Code + Number

OMB No. 0607-0919-C: Approval Expires 12/31/2011.

Form D-61 (9-25-2008)

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.
What is Person 1's name? Print name below.
Last Name
First Name MI

6. What is Person 1's sex? Mark [X] ONE box.
☐ Male ☐ Female

7. What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age on April 1, 2010 Month Day Year of birth

➜ NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.

8. Is Person 1 of Hispanic, Latino, or Spanish origin?
☐ No, not of Hispanic, Latino, or Spanish origin
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, another Hispanic, Latino, or Spanish origin — Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.

9. What is Person 1's race? Mark [X] one or more boxes.
☐ White
☐ Black, African Am., or Negro
☐ American Indian or Alaska Native — Print name of enrolled or principal tribe.
☐ Asian Indian ☐ Japanese ☐ Native Hawaiian
☐ Chinese ☐ Korean ☐ Guamanian or Chamorro
☐ Filipino ☐ Vietnamese ☐ Samoan
☐ Other Asian — Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.
☐ Other Pacific Islander — Print race, for example, Fijian, Tongan, and so on.
☐ Some other race — Print race.

10. Does Person 1 sometimes live or stay somewhere else?
☐ No ☐ Yes — Mark [X] all that apply.
☐ In college housing ☐ For child custody
☐ In the military ☐ In jail or prison
☐ At a seasonal or second residence ☐ In a nursing home
☐ For another reason

➜ If more people were counted in Question 1, continue with Person 2.

U S CENSUS BUREAU



**Figure 1.1** The first page of the 2010 census form, mailed to all households in the country. The 2010 census form can be found online at www.census.gov/2010census/about/interactive-form.php. (Source: Census.gov.)

Hispanics can be of any race. Again unable to choose a short name that would satisfy everyone, the Census Bureau decided to ask if you are of "Hispanic, Latino, or Spanish origin."

The fight over "race" reminds us that data reflect society. Race is a social idea, not a biological fact. In the census, you say what race you consider yourself to be. Race is a sensitive issue in the United States, so the fight is no surprise, and the Census Bureau's diplomacy seems a good compromise.

## Observational studies

As Yogi Berra, the former catcher and manager of the New York Yankees who is renowned for his humorous quotes, said, "You can observe a lot by watching." Sometimes all you can do is watch. To learn how chimpanzees in the wild behave, watch. To study how a teacher and young children interact in a schoolroom, watch. It helps if the watcher knows what to look for. The chimpanzee expert may be interested in how males and females interact, in whether some chimps in the troop are dominant, in whether the chimps hunt and eat meat. Indeed, chimps were thought to be vegetarians until Jane Goodall watched them carefully in Gombe National Park, Tanzania. Now it is clear that meat is a natural part of the chimpanzee diet.

At first, the observer may not know what to record. Eventually, patterns seem to emerge, and we can decide what variables we want to measure. How often do chimpanzees hunt? Alone or in groups? How large are hunting groups? Males alone, or both males and females? How much of the diet is meat? Observation that is organized and measures clearly defined variables is more convincing than just watching. Here is an example of highly organized (and expensive) observation.

**EXAMPLE 3** Do power lines cause leukemia in children?

© g bell/Alamy Stock Photo

Electric currents generate magnetic fields. So, living with electricity exposes people to magnetic fields. Living near power lines increases exposure to these fields. Really strong fields can disturb living cells in laboratory studies. What about the weaker fields we experience if we live near power lines? Some data suggested that more children in these locations might develop leukemia, a cancer of the blood cells.

We can't do experiments that deliberately expose children to magnetic fields for weeks and months at a time. It's hard to compare cancer rates among children who happen to live in more and less exposed locations because leukemia is quite rare and locations vary a lot in many ways other than magnetic fields. It is easier to start with children who have leukemia and compare them with children who don't. We can look at lots of possible causes—diet, pesticides, drinking water, magnetic fields, and others—to see where children with leukemia differ from those without. Some of these broad studies suggested a closer look at magnetic fields.

One really careful look at magnetic fields took five years and cost $5 million. The researchers compared 638 children who had leukemia and 620 who did not. They went into the homes and actually measured the magnetic fields in the children's bedrooms, in other rooms, and at the front door. They recorded facts about nearby power lines for the family home and also for the mother's residence when she was pregnant. Result: no evidence of more than a chance connection between magnetic fields and childhood leukemia. Similar conclusions were reached by researchers at the University of Oxford in England who reviewed data from 1962 to 2008.

"No evidence" that magnetic fields are connected with childhood leukemia doesn't prove that there is no risk. It says only that a very careful study could not find any risk that stands out from the play of chance that distributes leukemia cases across the landscape. In other words, the study could not rule out chance as a plausible explanation for what was observed. Critics continue to argue that the study failed to measure some important variables or that the children studied don't fairly represent all children. Nonetheless, a carefully designed observational study is a great advance over haphazard and sometimes emotional counting of cancer cases.

### Response variable and observational study

A **response** is a variable that measures an outcome or result of a study. An **observational study** observes individuals and measures variables of interest but does not intervene in order to influence the responses. The purpose of an observational study is to describe some group or situation.


**You just don't understand** A sample survey of journalists and scientists found quite a communications gap. Journalists think that scientists are arrogant, while scientists think that journalists are ignorant. We won't take sides, but here is one interesting result from the survey: 82% of the scientists agree that the ``media do not understand statistics well enough to explain new findings'' in medicine and other fields.

## Sample surveys

You don't have to eat the entire pot of soup to know it needs more salt. That is the idea of sampling: to gain information about the whole by examining only a part. **Sample surveys** are an important kind of observational study. They survey some group of individuals by studying only some of its members, selected not because they are of special interest but because they represent the larger group. Here is the vocabulary we use to discuss sampling.

### Populations and samples

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

A **sample** is the part of the population from which we actually collect information and is used to draw conclusions about the whole.

Notice that the *population* is the group we want to study. If we want information about all U.S. college students, that is our population—even if students at only one college are available for sampling. To make sense of any sample result, you must know what population the sample represents. Did a preelection poll, for example, ask the opinions of all adults? Or citizens only? Registered voters only? Democrats only? The *sample* consists of the people we actually have information about. If the poll can't contact some of the people it selected, those people aren't in the sample.

The distinction between population and sample is basic to statistics. The following examples illustrate this distinction and also introduce some major uses of sampling. These brief descriptions also indicate the variables measured for each individual in the sample.

### EXAMPLE 4 Public opinion polls

Polls such as those conducted by Gallup and many news organizations ask people's opinions on a variety of issues. The *variables* measured are responses to questions about public issues. Though most noticed at election time, these polls are conducted on a regular basis throughout the year. For a typical opinion poll:

*Population*: U.S. residents 18 years of age and over. Noncitizens and even illegal immigrants are included.

*Sample*: Between 1000 and 1500 people interviewed by telephone.

### EXAMPLE 5 The Current Population Survey

Government economic and social data come from large sample surveys of a nation's individuals, households, or businesses. The monthly Current Population Survey (CPS) is the most important government sample survey in the United States. Many of the *variables* recorded by the CPS concern the employment or unemployment of everyone over 16 years old in a household. The government's monthly unemployment rate comes

from the CPS. The CPS also records many other economic and social variables. For the CPS:

*Population*: The more than 123 million U.S. households. Notice that the individuals are households rather than people or families. A household consists of all people who share the same living quarters, regardless of how they are related to each other.

*Sample*: About 60,000 households interviewed each month.

### EXAMPLE 6 TV ratings

*Market research* is designed to discover what consumers want and what products they use. One example of market research is the television-rating service of Nielsen Media Research. The Nielsen ratings influence how much advertisers will pay to sponsor a program and whether or not the program stays on the air. For the Nielsen national TV ratings:

*Population*: The more than 116 million U.S. households that have a television set.

*Sample*: About 25,000 households that agree to use a "people meter" to record the TV viewing of all people in the household.

The *variables* recorded include the number of people in the household and their age and sex, whether the TV set is in use at each time period, and, if so, what program is being watched and who is watching it.

### EXAMPLE 7 The General Social Survey

Social science research makes heavy use of sampling. The General Social Survey (GSS), carried out every second year by the National Opinion Research Center at the University of Chicago, is the most important social science sample survey. The *variables* cover the subject's personal and family background, experiences and habits, and attitudes and opinions on subjects from abortion to war.

*Population*: Adults (aged 18 and over) living in households in the United States. The population does not include adults in institutions such as prisons and college dormitories. It also does not include persons who cannot be interviewed in English.

*Sample*: About 3000 adults interviewed in person in their homes.

**1.1 Legalizing marijuana.** The Pew Research Center conducted a poll on March 25–29, 2015. They asked:

NOW IT'S YOUR TURN

Do you think the use of marijuana should be made legal or not?

The Pew Research Center reported that the poll consisted of telephone interviews with 1500 randomly selected adult Americans. What do you think the population is? What is the sample?

Most statistical studies use samples in the broad sense. For example, the 638 children with leukemia in Example 3 are supposed to represent all children with leukemia. We usually reserve the dignified term "sample survey" for studies that use an organized plan to choose a sample that represents some specific population. The children with leukemia were patients at centers that specialize in treating children's cancer. Expert judgment says they are typical of all leukemia patients, even though they come only from special types of hospitals. A sample survey doesn't rely on judgment: it starts with an entire population and uses specific, quantifiable methods to choose a sample to represent the population. Chapters 2, 3, and 4 discuss the art and science of sample surveys.

## Census

A sample survey looks at only a part of the population. Why not look at the entire population? A *census* tries to do this.

### Census

A **census** is a sample survey that attempts to include the entire population in the sample.

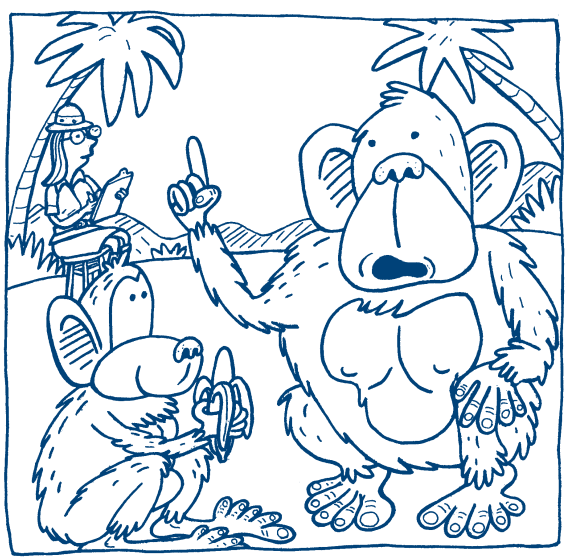
The U.S. Constitution requires a census of the American population every 10 years. A census of so large a population is expensive and takes a long time. Even the federal government, which can afford a census, uses samples such as the Current Population Survey to produce timely data on employment and many other variables. If the government asked every adult in the country about his or her employment, this month's unemployment rate would be available next year rather than next month. In fact, to save money, the 2010 census consisted of only 10 questions. Five of these were general questions, and five required answers for every person living at the address the form was sent to.

**Statistics in Your World**

**Is a census old-fashioned?** The United States has taken a census every 10 years since 1790. Technology marches on, however, and replacements for a national census look promising. Denmark has no census, and France plans to eliminate its census. Denmark has a national register of all its residents, who carry identification cards and change their register entry whenever they move. France will replace its census by a large sample survey that rotates among the nation's regions. The U.S. Census Bureau has a similar idea: the American Community Survey has already started and has eliminated the census "long form." Canada has eliminated its long-form census and replaced it with a controversial voluntary National Household Survey.

So time and money favor samples over a census. Samples can have other advantages as well. If you are testing fireworks or fuses, the sampled items are destroyed. Moreover, a sample can produce more accurate data than a census. A careful sample of an inventory of spare parts will almost certainly give more accurate results than asking the clerks to count all 500,000 parts in the warehouse. Bored people do not count accurately.

The experience of the Census Bureau reminds us that a census can only *attempt* to sample the entire population. The bureau estimated that the 2010 census overcounted the American population by 0.01% but undercounted the black population by 2.1%. A census is not foolproof, even with the resources of the government behind it. Why take a census at all? The government needs block-by-block population figures to create election districts with equal populations. The main function of the U.S. census is to provide this local information.



"Now eat that banana. The nice statistician is watching us."

## Experiments

Our goal in choosing a sample is a picture of the population, disturbed as little as possible by the act of gathering information. All observational studies share the principle "observe but don't disturb." When Jane Goodall first began observing chimpanzees in Tanzania, she set up a feeding station where the chimps could eat bananas. She later said that was a mistake because it might have changed the apes' behavior.

In *experiments*, on the other hand, we want to change behavior. In doing an experiment, we don't just observe individuals or ask them questions. We actively impose some treatment in order to observe the response. Experiments can answer questions such as "Does aspirin reduce the chance of a heart attack?" and "Do a majority of college students prefer Pepsi to Coke when they taste both without knowing which they are drinking?"

## Experiments

An **experiment** deliberately imposes some treatment on individuals in order to observe their responses. The purpose of an experiment is to study whether the treatment causes a change in the response.

### EXAMPLE 8 Helping welfare mothers find jobs

The Urban Institute in Washington DC reports that most adult welfare recipients are single mothers in their 20s and 30s with one or two children. Observational studies of welfare mothers show that many are able to increase their earnings and leave the welfare system. Some take advantage of voluntary job-training programs to improve their skills. Should participation in job-training and job-search programs be required of all able-bodied welfare mothers? Observational studies of the current system cannot tell us what the effects of such a policy would be. Even if the mothers studied are a properly chosen sample of all welfare recipients, those who seek out training and find jobs may differ in many ways from those who do not. They are observed to have more education, for example, but they may also differ in values and motivation, things that cannot be observed.

To see if a required jobs program will help mothers escape welfare, such a program must actually be tried. Choose two similar groups of mothers when they apply for welfare. Require one group to participate in a job-training program, but do not offer the program to the other group. This is an experiment. Comparing the income and work record of the two groups after several years will show whether requiring training has the desired effect.

**NOW IT'S YOUR TURN**

**1.2 Posting lectures on the class website.** To determine what students found most helpful, an educational researcher examined student comments from several classes he taught. Mentioned most often was the fact that copies of lectures were posted on the class website for students to download. The researcher recommended that instructors post lecture notes on their course websites. Was this an observational study or an experiment?

The welfare example illustrates the big advantage of experiments over observational studies: *in principle, experiments can give good evidence for cause and effect*. If we design the experiment properly, we start with two very similar groups of welfare mothers. The *individual* women of course differ from each other in age, education, number of children, and other respects. But the two *groups* resemble each other when we look at the ages, years of education, and number of children for all women in each

group. During the experiment, the women's lives differ, but there is only one systematic difference between the two groups: whether or not they are in the jobs program. All live through the same good or bad economic times, the same changes in public attitudes, and so on. If the training group does much better than the untrained group in holding jobs and earning money, we can say that the training program actually causes this happy outcome.

One of the big ideas of statistics is that experiments can give good evidence that a treatment causes a response. A big idea needs a big caution: statistical conclusions hold "on the average" for groups of individuals. They don't tell us much about one individual. If *on the average* the women in the training program earned more than those who were left out, that says that the program achieved its goal. It doesn't say that every woman in such a program will be helped. And a big idea may also raise big questions: if we hope the training will raise earnings, is it ethical to offer it to some women and not to others? Chapters 5 and 6 explain how to design good experiments, and Chapter 7 looks at ethical issues.

## STATISTICS IN SUMMARY

### Chapter Specifics

- Any statistical study records data about some **individuals** (people, animals, or things) by giving the value of one or more **variables** for each individual.
- Some variables, such as age and income, take numerical values. Others, such as occupation and sex, do not. Be sure the variables in a study really do tell you what you want to know.
- The most important fact about any statistical study is how the data were produced. **Observational studies** try to gather information without disturbing the scene they are observing.
- **Sample surveys** are an important kind of observational study. A sample survey chooses a **sample** from a specific **population** and uses the sample to get information about the entire population.
- A **census** attempts to measure every individual in a population.
- **Experiments** actually do something to individuals in order to see how they respond. The goal of an experiment is usually to learn whether some treatment actually causes a certain response.

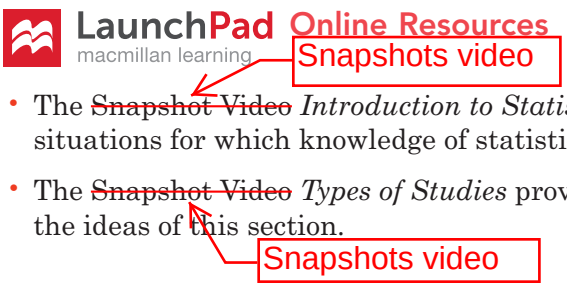

In reasoning from data to a conclusion, we start with the data. Where the data come from is the first step in the argument. The nature and validity of the conclusion are affected by this first step.

Two sources of data are observational studies and experiments. Observational studies are best suited for a conclusion that involves describing some group or situation without disturbing the scene we observe. Sample surveys are a type of observational study in which we draw conclusions about a population by observing only a part of the population (the sample). Experiments are best suited for a conclusion that involves determining if a treatment causes a change in a response.

In the next several chapters, we discuss these sources of data in more detail. We will see what makes for a good observational study and for a good experiment. And we will see how a bad observational study or experiment undermines the validity of the conclusions we wish to make.

**CASE STUDY EVALUATED** Use what you have learned in this chapter to answer some basic questions about the data collected in the MLive poll described in the Case Study that opened the chapter. To participate in the poll, you had to go online to the MLive website and click on one of the possible responses.

1. Is the poll a sample survey, census, or experiment?
2. What is the population of interest?
3. What are the individuals in the poll?
4. For each individual, what variable is measured?
5. Does this variable take numerical values?




- The ~~Snapshot Video~~ *Introduction to Statistics* describes real-world situations for which knowledge of statistical ideas is important.
- The ~~Snapshot Video~~ *Types of Studies* provides a nice introduction to the ideas of this section.

## CHECK THE BASICS

*For Exercise 1.1, see page 11; for Exercise 1.2, see page 13.*

**1.3 Individuals and variables.** A national survey by the Pew Research Center and *USA Today*, conducted August 20–24, 2014, was based on telephone interviews among a national sample of 1501 adults, 18 years of age or older, living in all 50 U.S. states and the District of Columbia. Those interviewed were asked to rate the job performance of police forces across the country for holding officers accountable when misconduct occurs. Possible ratings were "Excellent," "Good," "Only fair," and "Poor." Seventy percent of black respondents gave a rating of "Poor," while only 27% of white

respondents gave a rating of "Poor." For this study,

(a) the individuals are the sample of 1501 adults interviewed.

(b) the variable is the rating a respondent selected.

(c) both (a) and (b).

(d) neither (a) nor (b).

**1.4 Population and sample.** For the survey described in the previous exercise,

(a) the population is the 1501 adults interviewed and the sample are those who gave a rating of poor.

(b) the population is all adults, 18 years of age or older, living in all 50 U.S. states and the District of Columbia, and the sample are those who gave a rating of poor.

(c) the population is all adults, 18 years of age or older, living in all 50 U.S. states and the District of Columbia, and the sample is the black and white respondents.

(d) the population is all adults, 18 years of age or older, living in all 50 U.S. states and the District of Columbia, and the sample is the 1501 adults interviewed.

**1.5 Observational studies and experiments.** Researchers at The Ohio State University studied the effect of acetaminophen on emotion. The study included a total of 167 volunteers. Approximately half were randomly assigned to take acetaminophen and the other half a placebo. All subjects were shown a variety of images intended to invoke emotions. "Whether they saw things likely to make them feel bad or make them feel good, the acetaminophen group had a less-intense response. It's not that they didn't feel sad or happy, just that they felt less (by about 10 to 20%) than their placebo-swallowing peers. It doesn't turn them into zombies. It doesn't turn them into robots," one of the researchers said. Which of the following is true?

(a) This is an observational study, and participants were volunteers.

(b) This is an observational study, and participants were selected at random.

(c) This is an experiment, but participants themselves decided whether to take acetaminophen or the placebo.

(d) This is an experiment, and participants were randomly assigned to treatments.

**1.6 Response variable.** For the study described in the previous exercise, the response variable is which of the following?

(a) Some measure of emotional intensity invoked by the images

(b) Whether a subject received acetaminophen or the placebo

(c) The images

(d) The 167 volunteers

**1.7 A census?** A study is considered to be a census if

(a) the population of interest is very large.

(b) the study attempts to measure every individual in a population.

(c) all units in the study receive some treatment.

(d) it attempts to answer questions about the opinions of all citizens of a particular country.

## CHAPTER 1 EXERCISES

**1.8 Miles per gallon.** Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 2015 model motor vehicles.

(a) What are the individuals in this data set?

(b) For each individual, what variables are given? Which of these variables take numerical values?

| Make and model | Vehicle type | Transmission type | Number of cylinders | City mpg | Highway mpg |
|---|---|---|---|---|---|
| ⋮ | | | | | |
| BMW 328i | Compact car | Manual | 6 | 22 | 34 |
| BMW 335i | Compact car | Manual | 6 | 20 | 30 |
| Buick LaCrosse | Midsize car | Automatic | 6 | 18 | 28 |
| Chevrolet Traverse | Sport utility vehicle (FWD) | Automatic | 6 | 17 | 24 |
| ⋮ | | | | | |

**1.9 Athletes' salaries.** Here is a small part of a data set that describes Major League Baseball players as of opening day of the 2015 season.

(a) What individuals does this data set describe?

(b) In addition to the player's name, how many variables does the data set contain? Which of these variables take numerical values?

(c) What do you think are the *units* in which each of the numerical variables is expressed? For example, what does it mean to give Bryce Harper's annual salary as 3,750? (*Hint*: The average annual salary of a Major League Baseball player on opening day 2015 was, with round-off, $4,250,000.)

| Player | Team | Position | Age | Salary |
|---|---|---|---|---|
| ⋮ | | | | |
| Kershaw, Clayton | Dodgers | Pitcher | 27 | 31,000 |
| Harper, Bryce | Nationals | Outfielder | 22 | 3,750 |
| Sabathia, C. C. | Yankees | Pitcher | 34 | 24,286 |
| Rodriguez, Alex | Yankees | Designated Hitter | 39 | 22,000 |
| ⋮ | | | | |

**1.10 Who recycles?** In Example 1, weight is not a good measure of the participation of households in different neighborhoods in a city recycling program. What variables would you measure in its place?

**1.11 Sampling moms.** Pregnant and breast-feeding women should eat at least 12 ounces of fish and seafood per week to ensure their babies' optimal brain development, according to a coalition of top scientists from

private groups and federal agencies. A nutritionist wants to know whether pregnant women are eating at least 12 ounces of fish per week. To do so, she obtains a list of the 340 members of a local chain of prenatal fitness clubs and mails a questionnaire to 60 of these women selected at random. Only 21 questionnaires are returned. What is the population in this study? What is the sample from which information is actually obtained? What percentage of the women whom the nutritionist tried to contact responded?


**1.12 The death penalty.** A press release by the Gallup News Service says that, based on a poll conducted on October 12–15, 2014, it found that 63% of Americans respond Yes when asked this question: "Are you in favor of the death penalty for a person convicted of murder?" Toward the end of the article, you read: "Results for this Gallup poll are based on 1017 telephone interviews" and "based on a random sample of adults, 18 years and older living in all 50 U.S. states and the District of Columbia." What variable did this poll measure? What population do you think Gallup wants information about? What was the sample?

**1.13 The political gender gap.** There may be a "gender gap" in political party preference in the United States, with women more likely than men to prefer Democratic candidates. A political scientist interviews a large sample of registered voters, both men and women. She asks each voter whether he or she voted for the Democratic or the Republican candidate in the last presidential election. Is this study an experiment? Why or why not? What variables does the study measure?

**1.14 What is the population?** For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not sufficient, complete the description of the population by making reasonable assumptions about any missing information.

(a) An opinion poll contacts 972 American adults and asks them, "Would you rather have a job working for the government or working for business?"

(b) Video adapter cables have pins that plug into slots in a computer monitor. The adapter will not work if pins are bent or broken. A computer store buys video adapter cables in large lots from a supplier. The store chooses five cables from each lot and inspects the pins. If any of the cables have bent or broken pins, the entire lot is sent back.

(c) The American Community Survey contacts 3.5 million households, including some in every county in the United States. This Census Bureau survey asks each household questions about their housing, economic, and social status.

**1.15 What is the population?** For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not sufficient,

complete the description of the population in a reasonable way.

(a) A sociologist is interested in determining what proportion of teens believe the drinking age should be lowered to 18 in all the states. She selects a sample of five high schools in a large city and interviews all 12th-graders in each of the schools.

(b) A medical researcher is interested in the rate of dementia among former NFL football players. From a list of living, former players he selects a sample of 20 and interviews them to determine if signs of dementia are present.

(c) The host of a local radio talk show wonders if people who are actively religious are more likely to trust their neighbors than those who are not. The station receives calls from 51 listeners who voice their opinions.


**1.16 Teens' sleep needs.** A *Washington Post* article reported on a study about the sleep needs of teenagers. In the study, researchers measured the presence of the sleep-promoting hormone melatonin in teenagers' saliva at different times of the day. They learned that the melatonin levels rise later at night than they do in children and adults and remain at a higher level later in the morning. The teenagers who took part in the study were volunteers. Higher levels of melatonin indicate sleepiness. The researchers recommended that high schools start later in the day to accommodate the sleep needs of teens. Is this study an experiment, a sample survey, or an observational study that is not a sample survey? Explain your answer.

**1.17 Power lines and leukemia.** The study of power lines and leukemia in Example 3 compared two groups of individuals and measured many variables that might influence differences between the groups. Explain carefully why this study is *not* an experiment.

**1.18 Treating prostate disease.** A large study used records from Canada's national health care system to compare the effectiveness of two ways to treat prostate disease. The two treatments are traditional surgery and a new method that does not require surgery. The records described many patients whose doctors had chosen one or the other method. The study found that patients treated by the new method were more likely to die within eight years.

(a) Explain why this is an observational study, not an experiment.

(b) Briefly describe the nature of an experiment to compare the two ways to treat prostate disease.

**1.19 Walnuts and cholesterol.** Does eating walnuts increase the level of good cholesterol (HDL) and reduce the level of bad cholesterol (LDL)? Here are two ways to study this question.

1. Researchers in Australia recruited 58 adults with diabetes for a research study. These subjects were randomly assigned to two treatment groups: a low-modified-fat diet group and a low-modified-fat diet group that included a handful of walnuts each day. After six months, researchers compared changes in HDL and LDL levels for the two groups.

2. Another team of researchers recruited 58 adults with diabetes who

regularly eat walnuts as part of their diet. The researchers match each with a similar adult with diabetes who does not regularly eat walnuts. The researchers measured the HDL and LDL for each adult and compared the results for both groups.

(a) Explain why the first is an experiment and the second is an observational study.

(b) Why does the experiment give more useful information about whether walnuts increase HDL and reduce LDL?


**1.20 Alcohol and cancer in women.** A *Washington Post* article reported on a study about alcohol consumption and cancer in women. Since 1996, a team of British researchers has been gathering detailed information from 1.28 million women aged 50 to 64. The researchers recorded how much alcohol the women reported consuming when they volunteered for the study and again three years later. The researchers then examined whether there was any link with the 68,775 cancers the women developed over an average of the next seven years. They found that even among women who consumed as little as 10 grams of alcohol a day on average (the equivalent of about one drink), the risk for cancer of the breast, liver, and rectum was elevated.

(a) Is this an experiment? Explain your answer.

(b) We would prefer a sample survey to using women who volunteer for a study. What population does it appear that the researchers were interested in? What variables did they measure?


**1.21 Bullying.** Researchers tracked 2668 people from early childhood through adulthood and found that 13-year-olds who are frequent targets of bullies were three times more likely than their nonvictimized peers to be depressed later as adults. What is the population in this study? What is the sample? What variables do the researchers measure?

**1.22 Choose your study type.** What is the best way to answer each of the questions below: an experiment, a sample survey, or an observational study that is not a sample survey? Explain your choices.

(a) Is your school's football team called for fewer penalties in home games than in away games?

(b) Are college students satisfied with the cost of textbooks that they are required to purchase?

(c) Do college students who have access to video recordings of course lectures perform better in the course than those who don't?

**1.23 Choose your study purpose.** Give an example of a question about pet owners, their behavior, or their opinions that would best be answered by

(a) a sample survey.

(b) an observational study that is not a sample survey.

(c) an experiment.

**EXPLORING** THE WEB

*Follow the QR code to access exercises.*