



iStock.com/petesaloutos



iStock.com/KeithSzatranski



Al Diaz/Miami Herald/Getty Images

CHAPTER 1

Collecting Data

“You can’t fix by analysis what you bungled by design.”

Richard Light, Judith Singer, and John Willett in *By Design*

CHAPTER OUTLINE

1 Collecting Data 2

- 1.1 The Structure of Data 4
- 1.2 Sampling from a Population 17
- 1.3 Experiments and Observational Studies 31

Questions and Issues

Here are some of the questions and issues we will discuss in this chapter:

- Is there a “sprinting gene”?
- Does tagging penguins for identification purposes harm them?
- Do humans subconsciously give off chemical signals (pheromones)?
- What proportion of people using a public restroom wash their hands?
- If parents could turn back time, would they still choose to have children?
- Why do adolescent spiders engage in foreplay?
- How broadly do experiences of parents affect their future children?
- What percent of college professors consider themselves “above-average” teachers?
- Does giving lots of high fives to teammates help sports teams win?
- Which is better for peak performance: a short mild warm-up or a long intense warm-up?
- Are city dwellers more likely than country dwellers to have mood and anxiety disorders?
- Is there truth to the saying “beauty sleep”?
- What percent of young adults in the US move back in with their parents?
- Does turning up the music in a bar cause people to drink more beer?
- Is your nose getting bigger?
- Does watching cat videos improve mood?
- Does sleep deprivation hurt one’s ability to interpret facial expressions?
- Do artificial sweeteners cause weight gain?
- Does late night eating impair concentration?
- Does eating organic food improve your health?

1.1 THE STRUCTURE OF DATA

We are being inundated with data. It is estimated that the amount of new technical information is doubling every two years, and that over 7.2 zettabytes (that's 7.2×10^{21} bytes) of unique new information is being generated every year.¹ That is more than was generated during the entire 5000-year period before you were born. An incredible amount of data is readily available to us on the Internet and elsewhere. The people who are able to analyze this information are going to have great jobs and are going to be valuable in virtually every field. One of the wonderful things about statistics is that it is relevant in so many areas. Whatever your focus and your future career plans, it is likely that you will need statistical knowledge to make smart decisions in your field and in everyday life. As we will see in this text, effective collection and analysis of data can lead to very powerful results.

Statistics is the science of collecting, describing, and analyzing data. In this chapter, we discuss effective ways to *collect* data. In Chapter 2, we discuss methods to *describe* data. The rest of the chapters are devoted to ways of *analyzing* data to make effective conclusions and to uncover hidden phenomena.

DATA 1.1

A Student Survey

For several years, a first-day survey has been administered to students in an introductory statistics class at one university. Some of the data for a few of the students are displayed in Table 1.1. A more complete table with data for 362 students and 17 variables can be found in the file **StudentSurvey**.² ■

Cases and Variables

The subjects/objects that we obtain information about are called the *cases* or *units* in a dataset. In the **StudentSurvey** dataset, the cases are the students who completed the survey. Each row of the dataset corresponds to a different case.

A *variable* is any characteristic that is recorded for each case. Each column of our dataset corresponds to a different variable. The data in Table 1.1 show eight variables (in addition to the ID column), each describing a different characteristic of the students taking the survey.

Table 1.1 Partial results from a student survey

ID	Sex	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

¹“The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East,” <https://www.emc.com/leadership/digital-universe/2012iview/index.htm>. Accessed June 2020.

²Most datasets used in this text, and descriptions, are available electronically. They can be found at www.wiley.com/college/lock and at www.lock5stat.com. See the Preface for more information. Descriptions of many datasets can also be found in Appendix B.

Cases and Variables

We obtain information about **cases** or **units** in a dataset, and generally record the information for each case in a row of a data table.

A **variable** is any characteristic that is recorded for each case. The variables generally correspond to the columns in a data table.

In any dataset, it is important to understand exactly what each variable is measuring and how the values are coded. For the data in Table 1.1, the first column is ID, to provide an identifier for each of the individuals in the study. In addition, we have:

<i>Sex</i>	M for male and F for female ³
<i>Smoke</i>	Does the student smoke: yes or no
<i>Award</i>	Award the student prefers to win: Academy Award, Olympic gold medal, or Nobel Prize
<i>Exercise</i>	Number of hours spent exercising per week
<i>TV</i>	Number of hours spent watching television per week
<i>GPA</i>	Current grade point average on a 4-point scale
<i>Pulse</i>	Pulse rate in number of beats per minute at the time of the survey
<i>Birth</i>	Birth order: 1 for first/oldest, 2 for second born, etc.

Example 1.1

Explain what each variable tells us about the student with ID 1 in the first row of Table 1.1.

Solution



Student 1 is a male who does not smoke and who would prefer to win an Olympic gold medal over an Academy Award or a Nobel Prize. He says that he exercises 10 hours a week, watches television one hour a week, and that his grade point average is 3.13. His pulse rate was 54 beats per minute at the time of the survey, and he is the fourth oldest child in his family.

Categorical and Quantitative Variables

In determining the most appropriate ways to summarize or analyze data, it is useful to classify variables as either *categorical* or *quantitative*.

Categorical and Quantitative Variables

A **categorical variable** divides the cases into groups, placing each case into exactly one of two or more categories.

A **quantitative variable** measures or records a numerical quantity for each case. Numerical operations like adding and averaging make sense for quantitative variables.

³We acknowledge that this binary dichotomization is not a complete or inclusive representation of reality. However, the binary option is frequently used in data collection for ease of analysis and for confidentiality purposes. We apologize for the lack of inclusivity in situations involving sex and gender in this text.

We may use numbers to code the categories of a categorical variable, but this does not make the variable quantitative unless the numbers have a quantitative meaning. For example, “sex” is categorical even if we choose to record the results as 1 for male and 2 for female, since we are more likely to be interested in how many are in each category rather than an average numerical value. In other situations, we might choose to convert a quantitative variable into categorical groups. For example, “household income” is quantitative if we record the specific values but is categorical if we instead record only an income category (“low,” “medium,” “high”) for each household.

Example 1.2

Classify each of the variables in the student survey data in Table 1.1 as either categorical or quantitative.

Solution



Note that the ID column is neither a quantitative nor a categorical variable. A dataset often has a column with names or ID numbers that are for reference only.

- *Sex* is categorical since it classifies students into categories of male and female.
- *Smoke* is categorical since it classifies students as smokers or nonsmokers.
- *Award* is categorical since students are classified depending on which award is preferred.
- *Exercise*, *TV*, *GPA*, and *Pulse* are all quantitative since each measures a numerical characteristic of each student. It makes sense to compute an average for each variable, such as an average number of hours of exercise a week.
- *Birth* is a somewhat ambiguous variable, as it could be considered either quantitative or categorical depending on how we use it. If we want to find an average birth order, we consider the variable quantitative. However, if we are more interested in knowing how many first-borns, how many second-borns, and so on, are in the data, we consider the variable categorical. Either answer is acceptable.

Investigating Variables and Relationships between Variables

In this book, we discuss ways to describe and analyze a single variable and to describe and analyze relationships between two or more variables. For example, in the student survey data, we might be interested in the following questions, each about a single variable:

- What percentage of students smoke?
- What is the average number of hours a week spent exercising?
- Are there students with unusually high or low pulse rates?
- Which award is the most desired?
- How does the average GPA of students in the survey compare to the average GPA of all students at this university?

Often the most interesting questions arise as we look at relationships between variables. In the student survey data, for example, we might ask the following questions about relationships between variables:

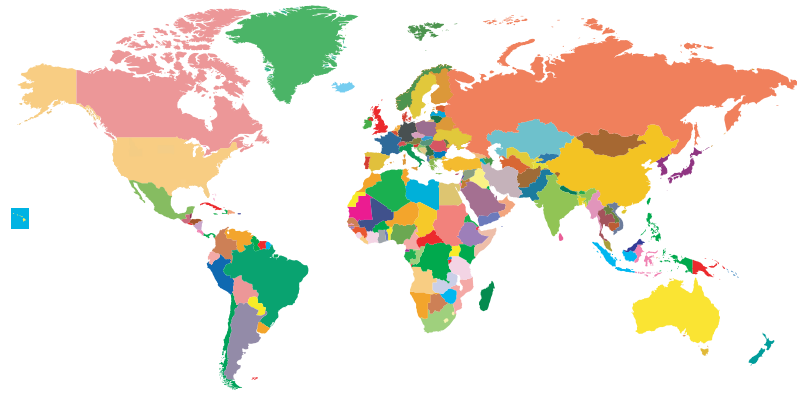
- Who smokes more, males or females?
- Do students who exercise more tend to prefer an Olympic gold medal? Do students who watch lots of television tend to prefer an Academy Award?

- Do males or females watch more television?
- Do students who exercise more tend to have lower pulse rates?
- Do first-borns generally have higher grade point averages?

These examples show that relationships might be between two categorical variables, two quantitative variables, or a quantitative and a categorical variable. In the following chapters, we examine statistical techniques for exploring the nature of relationships in each of these situations.

DATA 1.2**Data on Countries**

As of this writing, there are 217 countries listed by the World Bank.⁴ A great deal of information about these countries (such as energy use, birth rate, life expectancy) is in the full dataset under the name **AllCountries**. ■




redmal/Getty Images

Countries of the world**Example 1.3**

The dataset **AllCountries** includes information on the percent of people in each country with access to the Internet.

- Data from the Principality of Andorra were used to determine that 98.9% of Andorrans have access to the Internet, the highest rate of any country. What are the cases in the data from Andorra? What variable is used? Is it categorical or quantitative?
- In the **AllCountries** dataset, we record the percent of people with access to the Internet for each country. What are the cases in that dataset? What is the relevant variable? Is it categorical or quantitative?

Solution

-  For determining the rate of Internet usage in Andorra, the cases are people in Andorra, and the relevant variable is whether or not each person has access to the Internet. This is a categorical variable.
- In the **AllCountries** dataset, the cases are the countries of the world. The variable is the proportion with access to the Internet. For each country, we record a numerical value. These values range from a low of 1.3% in Eritrea to the high of 98.9% in Andorra, and the average is 54.47%. This is a quantitative variable.

⁴<http://data.worldbank.org/>. Data include information on both countries and economies, accessed June 2019.

As we see in the previous example, we need to think carefully about what the cases are and what is being recorded in each case in order to determine whether a variable is categorical or quantitative.

Example 1.4

In later chapters, we examine some of the following issues using the data in **AllCountries**. Indicate whether each question is about a single variable or a relationship between variables. Also indicate whether the variables are quantitative or categorical.

- How much energy does the average country use in a year?
- Do countries larger in area tend to have a more rural population?
- What is the relationship, if any, between a country's government spending on the military and on health care?
- Is the birth rate higher in developed or undeveloped countries?
- Which country has the highest percent of elderly people?

Solution



- The amount of energy used is a single quantitative variable.
- Both size and percent rural are quantitative variables, so this is a question about a relationship between two quantitative variables.
- Spending on the military and spending on health care are both quantitative, so this is another question about the relationship between two quantitative variables.
- Birth rate is a quantitative variable and whether or not a country is developed is a categorical variable, so this is asking about a relationship between a quantitative variable and a categorical variable.
- Because the cases are countries, percent elderly is a single quantitative variable.

Using Data to Answer a Question

The **StudentSurvey** and **AllCountries** datasets contain lots of information and we can use that information to learn more about students and countries. Increasingly, in this data-driven world, we have large amounts of data and we want to “mine” it for valuable information. Often, however, the order is reversed: We have a question of interest and we need to collect data that will help us answer that question.



iStock.com/petesaloutos

Is there a “sprinting gene”?

Example 1.5*Is There a “Sprinting Gene”?*

A gene called *ACTN3* encodes a protein which functions in fast-twitch muscles. Some people have a variant of this gene that cannot yield this protein. (So we might call the gene variant a possible *non-sprinting* gene.) To address the question of whether this gene is associated with sprinting ability, geneticists tested people from three different groups: world-class sprinters, world-class marathon runners, and a control group of non-athletes. In the samples tested, 6% of the sprinters had the gene variant, compared with 18% of the non-athletes and 24% of the marathon runners. This study⁵ suggests that sprinters are less likely than non-sprinters to have the gene variant.

- What are the cases and variables in this study? Indicate whether each variable is categorical or quantitative.
- What might a table of data look like for this study? Give a table with a possible first two cases filled in.

Solution

- The cases are the people included in the study. One variable is whether the individual has the gene variant or not. Since we record simply “yes” or “no,” this is a categorical variable. The second variable keeps track of the group to which the individual belongs. This is also a categorical variable, with three possible categories (sprinter, marathon runner, or non-athlete). We are interested in the relationship between these two categorical variables.
- The table of data must record answers for each of these variables and may or may not have an identifier column. Table 1.2 shows a possible first two rows for this dataset.

Table 1.2 Possible table to investigate whether there is a sprinter’s gene

Name	Gene Variant	Group
Allan	Yes	Marathon runner
Beth	No	Sprinter
...

Example 1.6*What’s a Habanero?*

A habanero chili is an extremely spicy pepper (roughly 500 times hotter than a jalapeño) that is used to create fiery food. The vice president of product development and marketing for the Carl’s Jr. restaurant chain⁶ is considering adding a habanero burger to the menu. In developing an advertising campaign, one of the issues he must deal with is whether people even know what the term “habanero” means. He identifies three specific questions of interest and plans to survey customers who visit the chain’s restaurants in various parts of the country.

- What proportion of customers know and understand what “habanero” means?
- What proportion of customers are interested in trying a habanero burger?
- How do these proportions change for different regions of the country?

⁵Yang, N., et al., “ACTN3 genotype is associated with human elite athletic performance,” *American Journal of Human Genetics*, September 2003; 73: 627–631.

⁶With thanks to Bruce Frazer.

- (a) Identify the cases for the data he collects.
- (b) Describe three variables that should be included in the dataset.

Solution

- ▶ (a) The cases in the habanero marketing study are the individual customers that respond to the survey request.
- (b) Here are three variables that should be included in the data:
- *Know* = yes or no, depending on whether the customer knows the term “habanero”
 - *Try* = yes or no, to indicate the customer’s willingness to try a habanero burger
 - *Region* = area in the country where the customer lives

All three variables are categorical.

Notice that for each case (customer), we record a value for each of the variables. Each customer would be represented in the dataset by a different row in the data table.



iStock.com/KeithSzafranski

Does tagging penguins harm them?

DATA 1.3

Tagging Penguins

Do metal tags on penguins harm them? Scientists trying to tell penguins apart have several ways to tag the birds. One method involves wrapping metal strips with ID numbers around the penguin’s flipper, while another involves electronic tags. Neither tag seems to physically harm the penguins. However, since tagged penguins are used to study *all* penguins, scientists wanted to determine whether the metal tags have any significant effect on the penguins. Data were collected over a 10-year time span from a sample of 100 penguins that were randomly given either metal or electronic tags. This included information on number of chicks, survival over the decade, and length of time on foraging trips.⁷ ■

Example 1.7


In the study on penguin tags:

- (a) What are the cases? What are the variables? Identify each variable as categorical or quantitative.

⁷Saroux, C., et al., “Reliability of flipper-banded penguins as indicators of climate change,” *Nature*, January 2011; 469: 203–206.

- (b) What information do the scientists hope to gain from the data? Describe at least one question in which they might be interested.

Solution

-  (a) The cases are the tagged penguins. The variables are the type of tag (categorical), number of chicks (quantitative), survival or not (categorical), and length of time on foraging trips (quantitative).
- (b) The scientists want to determine whether there is a relationship between the type of tag and any of the other variables. For example, they might be interested in knowing whether survival rates differ between the penguins with the metal tags and penguins with the electronic tags.

In Example 1.7, we are investigating whether one of the variables (the type of tag) helps us explain or predict values of the other variables. In this situation, we call the type of tag the *explanatory variable* and the other variables the *response variables*. One way to remember these names is the explanatory variable helps explain the response variable, and the response variable responds to the explanatory variable.

Explanatory and Response Variables


If we are using one variable to help us understand or predict values of another variable, we call the former the **explanatory variable** and the latter the **response variable**.

Example 1.8

In Example 1.4, we considered the following three questions about relationships between variables in the **AllCountries** dataset. Identify the explanatory variable and the response variable if it makes sense to do so.

- (a) Do countries larger in area tend to have a more rural population?
 (b) Is the birth rate higher in developed or undeveloped countries?
 (c) What is the relationship, if any, between a country's government spending on the military and on health care?

Solution

-  (a) The question indicates that we think area might influence the percent of a country that is rural, so we call area the explanatory variable and percent rural the response variable.
- (b) The question indicates that we think whether or not a country is developed might influence the birth rate, so the explanatory variable is whether the country is developed or undeveloped and the response variable is birth rate.
- (c) There is no indication in this situation of why we might identify either of the two variables (spending on military and spending on health care) as explanatory or as response. In a relationship between two variables, we don't always identify one as the explanatory variable and one as the response variable.

Different Ways to Answer a Broad Question

A pheromone is a chemical signal given off by one member of a species that influences other members of the species. Many studies (involving data, of course!) provide evidence that different types of animals give off pheromones. It is currently under debate whether humans also communicate subconsciously through pheromones. How might we collect data to answer the question of whether there

are human pheromones? We might start by narrowing the question to one that is not so general. For example, are there pheromones in female tears that affect the behavior of males?

Several studies⁸ suggest that the scent of tears from a crying woman may reduce sexual interest in men. However, to determine whether this effect is caused subconsciously by pheromones (rather than by obvious social influences), we need to think carefully about how to collect data. How might you collect data to answer this question? Three different methods were used in the studies. See what you think of the three methods.

- In one study, 25 men in their twenties had a pad attached to their upper lip that contained either tears collected from women who watched sad films or a salt solution that had been trickled down the same women's faces. Neither substance had a perceptible odor. The men who had tears on the upper lip rated female faces as less sexually alluring than the men who had salt solution on the upper lip.
- In a second study, 50 men who sniffed women's tears showed reduced levels of testosterone relative to levels after sniffing a salt solution.
- In a third study involving 16 men, those who sniffed female tears displayed significantly reduced brain-cell activity in areas that had reacted strongly to an erotic movie, whereas those who sniffed a salt solution did not show the same reduced activity.

Example 1.9

For each of the three studies on women's tears, state the explanatory and response variables.

Solution



In all three studies, the explanatory variable is whether tears or a salt solution is used. In the first study, the response variable is how sexually alluring males rated female faces, in the second study it is testosterone levels, and in the third study it is brain cell activity.

All three of these studies describe data collected in a careful way to answer a question. How to collect data in a way that helps us understand real phenomena is the subject of the rest of this chapter.

We have described several datasets, studies, and questions in this section, involving students, countries, sprinter genes, habanero burgers, penguins, and pheromones. If you are intrigued by any of these questions, keep reading! We examine all of them in more detail in the pages ahead.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- ▶ Recognize that a dataset consists of cases and variables
- ▶ Identify variables as either categorical or quantitative
- ▶ Determine explanatory and response variables where appropriate
- ▶ Describe how data might be used to address a specific question
- ▶ Recognize that understanding statistics allows us to investigate a wide variety of interesting phenomena

⁸Gelstein, S., et al., "Human Tears Contain a Chemosignal," *Science*, January 2011; 331(6014): 226–230.

Exercises for Section 1.1

SKILL BUILDER 1

For the situations described in Exercises 1.1 to 1.6:

- What are the cases?
- What is the variable and is it quantitative or categorical?

1.1 People in a city are asked if they support a new recycling law.

1.2 Record the percentage change in the price of a stock for 100 stocks publicly traded on Wall Street.

1.3 Collect data from a sample of teenagers with a question that asks “Do you eat at least five servings a day of fruits and vegetables?”

1.4 Measure the shelf life of bunches of bananas (the number of days until the bananas go bad) for a large sample.

1.5 Estimate the bending strength of beams by bending 10 beams until they break and recording the force at which the beams broke.

1.6 Record whether or not the literacy rate is over 75% for each country in the world.

SKILL BUILDER 2

In Exercises 1.7 to 1.10, a relationship between two variables is described. In each case, we can think of one variable as helping to explain the other. Identify the explanatory variable and the response variable.

1.7 Lung capacity and number of years smoking cigarettes

1.8 Amount of fertilizer used and the yield of a crop

1.9 Blood alcohol content (BAC) and number of alcoholic drinks consumed

1.10 Year and the world record time in a marathon

1.11 Student Survey Variables Data 1.1 introduced the dataset **StudentSurvey**, and Example 1.2 identified seven of the variables in that dataset as categorical or quantitative. The remaining variables are:

<i>Year</i>	FirstYear, Sophomore, Junior, Senior
<i>Height</i>	In inches
<i>Weight</i>	In pounds
<i>Siblings</i>	Number of siblings the person has
<i>VerbalSAT</i>	Score on the Verbal section of the SAT exam

MathSAT Score on the Math section of the SAT exam

SAT Sum of the scores on the Verbal and Math sections of the SAT exam

HigherSAT Which is higher, Math SAT score or Verbal SAT score?

- Indicate whether each variable is quantitative or categorical.
- List at least two questions we might ask about any one of these individual variables.
- List at least two questions we might ask about relationships between any two (or more) of these variables.

1.12 Countries of the World Information about the world’s countries is given in **AllCountries**, introduced in Data 1.2 on page 7. You can find a description of the variables in Appendix B. For the full dataset:

- Indicate which of the variables are quantitative and which are categorical.
- List at least two questions we might ask about any one of these individual variables.
- List at least two questions we might ask about relationships between any two (or more) of these variables.

1.13 Goldilocks Effect: Read to Your Kids! The American Academy of Pediatrics recommends that parents begin reading to their children soon after birth, and that parents set limits on screen time. A new study⁹ reinforces these recommendations. In the study, 27 four-year-olds were presented with stories in three different formats: audio (sound only), illustrated (sound and pictures), and animated (sound and animation). During the presentations, a magnetic resonance imaging (MRI) machine measured each child’s brain connectivity. The researchers found a “Goldilocks effect,” in which audio was too cold (with low brain connectivity as the children strained to understand) and animation was too hot (with low brain connectivity as the animation did all the work for the children). The highest connectivity (just right!) was found with the illustrated format, which simulates reading a book to a child.

⁹Hutton J et al., “Differences in functional brain network connectivity during stories presented in audio, illustrated, and animated format in preschool-age children,” *Brain Imaging and Behavior*, October 30, 2018.

- (a) What is the explanatory variable? Is it categorical or quantitative?
- (b) What is the response variable? Is it categorical or quantitative?
- (c) How many cases are there?

1.14 Female Gamers Face Sexual Harassment A research firm¹⁰ questioned 1151 female gamers in Great Britain and found that 40% had received obscene messages while playing online. In addition to asking whether they had received obscene messages, the gamers were also asked how many hours a week they played, and whether they felt there were enough strong female characters in games.

- (a) What are the cases in this study?
- (b) What are the variables? Indicate whether each is categorical or quantitative.
- (c) How many rows and how many columns will the dataset have if cases are rows and variables are columns?

1.15 Active Learning vs Passive Learning: Which is Best? Active learning in a classroom implies that students are actively involved and working during class time (either individually, in pairs, or in groups) while passive learning indicates that students are primarily taking notes while the instructor lectures. A recent study¹¹ measured students actual learning under these two formats as well as their feelings of learning. The study was very well designed: students in a college physics course were randomly assigned to a class period with either active learning or passive learning, the same content and handouts were used in both, and both instructors were highly rated. After the class, students were asked to rate how much they thought they had learned (on a 5-point Likert scale) and they also took a 20-question multiple choice exam to test how much they had actually learned. The results were very interesting: students *thought* that they learned more in the passive learning class but they *actually* learned more in the active learning class.

- (a) What are the cases?
- (b) What are the variables? Indicate whether each variable is quantitative or categorical.
- (c) Indicate explanatory and response variables.

¹⁰“Research: One in 3 Female Gamers Face Gender Discrimination, 32% Deal with Sexual Harassment,” *Bryter-research.co.uk*, October 2019.

¹¹Deslauriers L, et al., “Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom,” *PNAS*, 116(39), September 24, 2019.

- (d) There were 154 students in the passive learning lecture and 142 students in the active learning class. Indicate how many rows and how many columns the dataset will have if cases are rows and variables are columns.

1.16 Spider Sex Play Spiders regularly engage in spider foreplay that does not culminate in mating. Male spiders mature faster than female spiders and often practice the mating routine on not-yet-mature females. Since male spiders run the risk of getting eaten by female spiders, biologists wondered why spiders engage in this behavior. In one study,¹² some spiders were allowed to participate in these near-matings, while other maturing spiders were isolated. When the spiders were fully mature, the scientists observed real matings. They discovered that if either partner had participated at least once in mock sex, the pair reached the point of real mating significantly faster than inexperienced spiders did. (Mating faster is, apparently, a real advantage in the spider world.) Describe the variables, indicate whether each variable is quantitative or categorical, and indicate the explanatory and response variables.

1.17 Hormones and Fish Fertility When women take birth control pills, some of the hormones found in the pills eventually make their way into lakes and waterways. In one study, a water sample was taken from various lakes. The data indicate that as the concentration of estrogen in the lake water goes up, the fertility level of fish in the lake goes down. The estrogen level is measured in parts per trillion (ppt) and the fertility level is recorded as the percent of eggs fertilized. What are the cases in this study? What are the variables? Classify each variable as either categorical or quantitative.

1.18 Fast-Twitch Muscles and Race Example 1.5 studied a variant of the gene *ACTN3* which inhibits fast-twitch muscles and seems to be less prevalent in sprinters. A separate study¹³ indicated ethnic differences: Approximately 20% of a sample of Caucasians, approximately 25% of a sample of Asians, and approximately 1% of a sample of Africans had the gene variant. What are the variables in this study? Classify each as categorical or quantitative.

¹²Pruitt, J., paper presented at the Society for Integrative and Comparative Biology Annual Meeting, January 2011, and reported in “For spiders, sex play has its pluses,” *Science News*, January 29, 2011.

¹³North, K., et al., “A common nonsense mutation results in α -actinin-3 deficiency in the general population,” *Nature Genetics*, April 1999; 21(4): 353–354.

1.19 Largest Cities in the World Seven of the ten largest cities in the world are in the Eastern Hemisphere (including the largest: Tokyo, Japan) and three are in the Western Hemisphere.¹⁴ Table 1.3 shows the populations, in millions of people, for these cities.

- How many cases are there in this dataset? How many variables are there and what are they? Is each categorical or quantitative?
- Display the information in Table 1.3 as a dataset with cases as rows and variables as columns.

Table 1.3 *Population, in millions, of the world's largest cities*

Eastern hemisphere:	37, 26, 23, 22, 21, 21, 21
Western hemisphere:	21, 20, 19

1.20 How Fast Do Homing Pigeons Go? Homing pigeons have an amazing ability to find their way home over extremely long distances. How fast do they go on these trips? In the 2019 Midwest Classic, held in Topeka, Kansas, the fastest bird went 1676 YPM (yards per minute), which is about 56 miles per hour.¹⁵ The top seven finishers included three female pigeons (Hens) and four male pigeons (Cocks). Their speeds, in YPM, are given in Table 1.4.

- How many cases are there in this dataset? How many variables are there and what are they? Is each variable categorical or quantitative?
- Display the information as a dataset with cases as rows and variables as columns.

Table 1.4 *Speed of homing pigeons, in yards per minute*

Hens:	1676, 1452, 1449
Cocks:	1458, 1435, 1418, 1413

1.21 Pigeon Racing Exercise 1.20 gives the speed of the top seven finishers in the 2019 Midwest Classic homing pigeon race. In fact, 1412 pigeons finished the race, and their home loft, sex, distance, and speed were all recorded. A loft may have several different pigeons finish the race.

¹⁴<http://www.worldatlas.com/citypops.htm>. Accessed June 2015.

¹⁵Data downloaded from the Midwest Homing Pigeon Association final race report at <http://www.midwesthpa.com/MIDFinalReports.htm>.

- How many cases are in this dataset?
- How many variables are there? How many of these variables are categorical? How many are quantitative?
- How many rows and how many columns will the dataset have?

1.22 Trans-Generational Effects of Diet Can experiences of parents affect future children? New studies¹⁶ suggest that they can: Early life experiences of parents appear to cause permanent changes in sperm and eggs. In one study, some male rats were fed a high-fat diet with 43% of calories from fat (a typical American diet), while others were fed a normal healthy rat diet. Not surprisingly, the rats fed the high-fat diet were far more likely than the normal-diet rats to develop metabolic syndrome (characterized by such things as excess weight, excess fat, insulin resistance, and glucose intolerance.) What surprised the scientists was that the daughters of these rats were also far more likely to develop metabolic syndrome than the daughters of rats fed healthy diets. None of the daughters and none of the mothers ate a high-fat diet and the fathers did not have any contact with the daughters. The high-fat diet of the fathers appeared to cause negative effects for their daughters. What are the two main variables in this study? Is each categorical or quantitative? Identify the explanatory and response variables.

1.23 Trans-Generational Effects of Environment

In Exercise 1.22, we ask whether experiences of parents can affect future children, and describe a study that suggests the answer is yes. A second study, described in the same reference, shows similar effects. Young female mice were assigned to either live for two weeks in an enriched environment or not. Matching what has been seen in other similar experiments, the adult mice who had been exposed to an enriched environment were smarter (in the sense that they learned how to navigate mazes faster) than the mice that did not have that experience. The other interesting result, however, was that the offspring of the mice exposed to the enriched environment were also smarter than the offspring of the other mice, even though none of the offspring were exposed to an enriched environment themselves. What are the two main variables in this study? Is each categorical or quantitative? Identify explanatory and response variables.

¹⁶Begley, S., "Sins of the Grandfathers," *Newsweek*, November 8, 2010; 48–50.

1.24 Pennsylvania High School Seniors The data in **PA Seniors** shows results for a sample of 457 high school seniors in the state of Pennsylvania, selected at random from all students who participated in the Census at Schools project¹⁷ between 2010 and 2019. Each of the questions below relate to information in this dataset. Determine whether the answer to each question gives a value for a quantitative variable, a categorical variable, or is not a value for a variable for this dataset.

- What mode of transportation do you use to get to school?
- Do you have any allergies?
- What proportion of students in this sample are vegetarians?
- How many hours did you spend last week working at a paid job?
- What is the difference between typical hours of sleep you get on school nights and non-school nights?
- What is the maximum time (in minutes) that a student in this sample needs to get to school?
- If you could have a super power would you choose invisibility, telepathy, super strength, ability to fly, or ability to freeze time?

1.25 US College Scorecard The data in **CollegeScores** contains information from the US Department of Education's College Scorecard¹⁸ on all postsecondary educational institutions in the US. Each of the questions below relate to information in this dataset. Determine whether the answer to each question gives a value for a quantitative variable, a categorical variable, or is not a value for a variable for this dataset.

- What is the total tuition and fees for in-state students at the school?
- How many of these schools are located in the Northeast?
- Is the school public, private, or for profit?
- How many undergraduates are enrolled at the school?
- What percentage of undergraduates at the school are part-time students?
- Which school has the highest average faculty salary?

¹⁷Sample data obtained from the Census at Schools random sampler sponsored by the American Statistical Association at <https://www2.amstat.org/censusatschool>.

¹⁸Data downloaded from the US Department of Education's College Scorecard at <https://collegescorecard.ed.gov/data/> (November 2019.)

1.26 Hookahs and Health Hookahs are waterpipes used for smoking flavored tobacco. One study¹⁹ of 3770 university students in North Carolina found that 40% had smoked a hookah at least once, with many claiming that the hookah smoke is safer than cigarette smoke. However, a second study observed people at a hookah bar and recorded the length of the session, the frequency of puffing, and the depth of inhalation. An average session lasted one hour and the smoke inhaled from an average session was equal to the smoke in more than 100 cigarettes. Finally, a third study measured the amount of tar, nicotine, and heavy metals in samples of hookah smoke, finding that the water in a hookah filters out only a very small percentage of these chemicals. Based on these studies and others, many states are introducing laws to ban or limit hookah bars. In each of the three studies, identify the individual cases, the variables, and whether each variable is quantitative or categorical.

1.27 Is Your Nose Getting Bigger? Next time you see an elderly man, check out his nose and ears! While most parts of the human body stop growing as we reach adulthood, studies show that noses and ears continue to grow larger throughout our lifetime. In one study²⁰ examining noses, researchers report "Age significantly influenced all analyzed measurements:" including volume, surface area, height, and width of noses. The sex of the 859 participants in the study was also recorded, and the study reports that "male increments in nasal dimensions were larger than female ones."

- How many variables are mentioned in this description?
- How many of the variables are categorical? How many are quantitative?
- If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

1.28 Don't Text While Studying! For the 2015 Intel Science Fair, two brothers in high school recruited 47 of their classmates to take part in a two-stage study. Participants had to read two different passages and then answer questions on them, and each person's score was recorded for each of the two

¹⁹Quenqua, D., "Putting a Crimp in the Hookah," *New York Times*, May 31, 2011, p A1.

²⁰Sforza, C., Grandi, G., De Menezes, M., Tartaglia, G.M., and Ferrario, V.F., "Age- and sex-related changes in the normal human external nose," *Forensic Science International*, January 30, 2011; 204(1–3): 205.e1–9.

tests. There were no distractions for one of the passages, but participants received text messages while they read the other passage. Participants scored significantly worse when distracted by incoming texts. Participants were also asked if they thought they were good at multitasking (yes or no) but “even students who were confident of their abilities did just as poorly on the test while texting.”²¹

- What are the cases?
- What are the variables? Is each variable categorical or quantitative?
- If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

1.29 Help for Insomniacs A recent study shows that just one session of cognitive behavioral therapy can help people with insomnia.²² In the study, forty people who had been diagnosed with insomnia were randomly divided into two groups of 20 each. People in one group received a one-hour cognitive behavioral therapy session while those in the other group received no treatment. Three months later, 14 of those in the therapy group reported sleep improvements while only 3 people in the other group reported improvements.

²¹Perkins, S., “Studying? Don’t answer that text!” *Science News*, July 23, 2015.

²²Ellis, J.G., Cushing, T., and Germain, A., “Treating acute insomnia: a randomized controlled trial of a ‘single-shot’ of cognitive behavioral therapy for insomnia,” *SLEEP*, 2015; 38(6): 971–978.

- What are the cases in this study?
- What are the relevant variables? Identify each as categorical or quantitative.
- If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

1.30 How Are Age and Income Related? An economist collects data from many people to determine how age and income are related. How the data is collected determines whether the variables are quantitative or categorical. Describe how the information might be recorded if we regard both variables as quantitative. Then describe a different way to record information about these two variables that would make the variables categorical.

1.31 Political Party and Voter Turnout Suppose that we want to investigate the question “Does voter turnout differ by political party?” How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

1.32 Wealth and Happiness Are richer people happier? How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

1.33 Choose Your Own Question Come up with your own question you would like to be able to answer. What is the question? How might you collect data to answer this question? What would the cases be? What would the variable(s) be?

1.2 SAMPLING FROM A POPULATION

While most of this textbook is devoted to analyzing data, the way in which data are *collected* is critical. Data collected well can yield powerful insights and discoveries. Data collected poorly can yield very misleading results. Being able to think critically about the method of data collection is crucial for making or interpreting data-based claims. In the rest of this chapter, we address some of the most important issues that need to be considered when collecting data.

Samples from Populations

The US Census is conducted every 10 years and attempts to gather data about all people living in the US. For example, the census shows that, for people living in the US who are at least 25 years old, 84.6% have at least a high school degree and 27.5% have at least a college bachelor’s degree.²³ The cases in the census dataset are all residents of the US, and there are many variables measured on these cases. The US census attempts to gather information from an entire *population*. In **AllCountries**,

²³<http://factfinder.census.gov>.

introduced as Data 1.2 on page 7, the cases are countries. This is another example of a dataset on an entire population because we have data on every country.

Usually, it is not feasible to gather data for an entire population. If we want to estimate the percent of people who wash their hands after using a public restroom, it is certainly not possible to observe all people all the time. If we want to try out a new drug (with possible side effects) to treat cancer, it is not safe to immediately give it to all patients and sit back to observe what happens. If we want to estimate what percentage of people will react positively to a new advertising campaign, it is not feasible to show the ads to everyone and then track their responses. In most circumstances, we can only work with a *sample* from what might be a very large population.

Samples from Populations

A **population** includes all individuals or objects of interest.

Data are collected from a **sample**, which is a subset of the population.

Example 1.10

To estimate what percent of people in the US wash their hands after using a public restroom, researchers pretended to comb their hair while observing 6000 people in public restrooms throughout the United States. They found that 85% of the people who were observed washed their hands after going to the bathroom.²⁴ What is the sample in this study? What is a reasonable population to which we might generalize?

Solution



The sample is the 6000 people who were observed. A reasonable population to generalize to would be all people in the US. There are other reasonable answers to give for the population, such as all people in the US who use public restrooms or all people in the US who use public restrooms in the cities in which the study was conducted. Also, people might behave differently when alone than when there is someone else in the restroom with them, so we might want to restrict the population to people in a restroom with someone else.

We denote the size of the sample with the letter n . In Example 1.10, $n = 6000$ because there are 6000 people in the sample. Usually, the sample size, n , is much smaller than the size of the entire population.

Since we rarely have data on the entire population, a key question is how to use the information in a sample to make reliable statements about the population. This is called *statistical inference*.

Statistical Inference

Statistical inference is the process of using data from a sample to gain information about the population.

Figure 1.1 diagrams the process of selecting a sample from a population, and then using that sample to make inferences about the population. Much of the data analysis discussed in this text focuses on the latter step, statistical inference. However, the first step, selecting a sample from the population, is critical because the process used to collect the sample determines whether valid inference is even possible.

²⁴Zeizima, K., “For many, ‘Washroom’ seems to be just a name,” *New York Times*, September 14, 2010, p. A14.

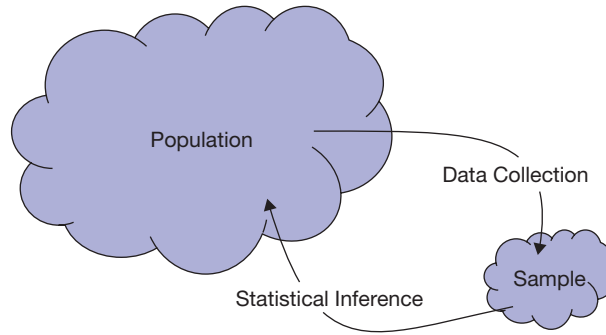


Figure 1.1 From population to sample and from sample to population

Sampling Bias

Example 1.11

Dewey Defeats Truman

The day after the 1948 presidential election, the *Chicago Tribune* ran the headline “Dewey Defeats Truman.” However, Harry S Truman defeated Thomas E. Dewey to become the 33rd president of the United States. The newspaper went to press before all the results had come in, and the headline was based partly on the results of a large telephone poll which showed Dewey sweeping Truman.

- What is the sample and what is the population?
- What did the pollsters want to infer about the population based on the sample?
- Why do you think the telephone poll yielded such inaccurate results?

Solution

- The sample is all the people who participated in the telephone poll. The population is all voting Americans.
- The pollsters wanted to estimate the percentage of all voting Americans who would vote for each candidate.
- One reason the telephone poll may have yielded inaccurate results is that people with telephones in 1948 were not representative of all American voters. People with telephones tended to be wealthier and prefer Dewey while people without phones tended to prefer Truman.



Underwood Archives/Getty Images

A triumphant Harry S Truman holds the *Chicago Tribune* published with the incorrect headline “Dewey Defeats Truman”

The previous example illustrates *sampling bias*, because the method of selecting the sample biased the results by selecting only people with telephones.

Sampling Bias

Sampling bias occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way. If sampling bias exists, then we cannot trust generalizations from the sample to the population.

Example 1.12

After a flight, one of the authors received an email from the airline asking her to fill out a survey regarding her satisfaction with the travel experience. The airline analyzes the data from all responses to such emails.

- (a) What is the sample and in what population is the airline interested?
- (b) Do you expect these survey results to accurately portray customer satisfaction?

Solution



- (a) The sample is all people who choose to fill out the survey and the population is all people who fly this airline.
- (b) The survey results will probably not accurately portray customer satisfaction. Many people won't bother to fill out the survey if the flight was uneventful, while people with a particularly bad or good experience are more likely to fill out the survey.



A sample comprised of volunteers (like the airline survey) often creates sampling bias in opinion surveys, because the people who choose to participate (the sample) often have more extreme opinions than the population.

To avoid sampling bias, we try to obtain a sample that is *representative* of the population. A representative sample resembles the population, only in smaller numbers. The telephone survey in 1948 reached only people wealthy enough to own a telephone, causing the sample to be wealthier than the population, so it was not a representative sample. The more representative a sample is, the more valuable the sample is for making inferences about the population.

Example 1.13

An online poll conducted on *biblegateway.com* asked, “How often do you talk about the Bible in your normal course of conversation?” Over 5000 people answered the question, and 78% of respondents chose the most frequent option: Multiple times a week. Can we infer that 78% of people talk about the bible multiple times a week? Why or why not?

Solution



- No. People who visit the website for *Bible Gateway* and choose to take the poll are probably more likely than the general public to talk about the bible. This sample is not representative of the population of all people, so the results cannot be generalized to all people.

Simple Random Sample

Since a representative sample is essential for drawing valid inference to the population, you are probably wondering how to select such a sample! The key is *random sampling*. We can imagine putting the names of all the cases in the population into a hat and drawing out names to be in our sample. Random sampling avoids sampling bias.

Simple Random Sample

When choosing a **simple random sample** of n units, all groups of size n in the population have the same chance of becoming the sample. As a result, in a simple random sample, each unit of the population has an equal chance of being selected, regardless of the other units chosen for the sample.

Taking a simple random sample avoids sampling bias.

Part of the power of statistics lies in this amazing fact: A simple random sample tends to produce a good representative sample of the population. At the time of writing this book, the population of the United States is more than 300 million people. Although the census collects some data on the entire population, for many questions of interest we are forced to rely on a small sample of individuals. Amazingly, if a simple random sample is selected, even a small sample can yield valid inferences for all 300 million Americans!

Example 1.14

Election Polling

Right before the 2012 presidential election, Google Consumer Surveys²⁵ randomly sampled and collected data on $n = 3252$ Americans. The sample showed Barack Obama ahead of Mitt Romney in the national popular vote by 2.3 percentage points. Can we generalize these results to the entire population of 126 million voters in order to estimate the popular vote in the election?

Solution



Yes! Because the poll included a random sample of voters, the results from the sample should generalize to the population. In the actual election, Obama won the national popular vote by 2.6 percentage points. Of course, the sample data do not perfectly match the population data (in Chapter 3 we will learn how closely we expect sample results to match population results), but the fact that we can get such an accurate guess from sampling only a very small fraction of the population is quite astonishing!

Analogy to Soup



Patti McConville/Alamy Stock Photo

Sampling the soup

²⁵Data from <http://www.fivethirtyeight.com>, “Which Polls Fared Best (and Worst) in the 2012 Presidential Race,” November 10, 2012, and Google Consumer Surveys, November 5, 2012, <http://www.google.com/insights/consumersurveys/view?survey=6iwb56wuu5vc6&question=2&filter=&rw=1>.

Do we need to eat an entire large pot of soup to know what the soup tastes like? No! As long as the soup is well mixed, a few spoonfuls will give us a pretty good idea of the taste of the soup. This is the idea behind sampling: Just sampling a small part of the soup (or population) can still give us a good sense of the whole. However, if the soup is not mixed, so that the broth is at the top, the meat at the bottom, and so on, then a few spoonfuls will *not* give us a good sense of the taste of the soup. This is analogous to taking a biased sample. Mixing the soup randomizes the sample, so that the small part we taste is representative of the entire large pot.

How Do We Select a Random Sample?

You may think that you are capable of “randomly” selecting samples on your own, but you are wrong! Deborah Nolan, a statistics professor, has half of her students flip a coin and write the resulting sequence of heads and tails on the board (flipping a coin is truly random), and the other half of her students generate their own sequence of heads and tails without actually flipping the coin, trying to fake randomness. She can always tell the difference.²⁶ How can she tell? Because students (and their professors and all people!) are very bad at actual randomness.

Similarly, you may think you can select representative samples better than randomness can, but again, you are most likely wrong! Just as humans are surprisingly bad at faking randomness, humans are surprisingly bad at selecting representative samples. We tend to oversample some types of units and undersample less obvious types, even when we are explicitly trying hard not to do so. Luckily, randomness is surprisingly good at selecting a representative sample.

If we can’t do randomness ourselves, how *do* we select a random sample? As mentioned, one way is to draw names out of a hat. A more common way is to use technology. Some forms of technology can automatically draw a sample randomly from a list of the entire population, mimicking the process of drawing names from a hat. Other forms produce random numbers, in which case we give a number to each unit in the population, and our sample becomes the units corresponding to the selected numbers.

Example 1.15

The dataset **AllCountries** contains data on 217 countries or economies. Select a random sample of 5 of these.

Solution



There are many ways to do this, depending on the technology or method available. One way is to number the countries from 1 to 217 and then use a random number generator to select five of the numbers between 1 and 217. Suppose we do this and get the numbers

47 88 155 164 50

As we see in the dataset, the corresponding countries are Costa Rica (47), Hungary (88), Peru (155), Samoa (164), and Cuba (50). These five countries are a random sample from the population of all countries. Of course, the numbers are randomly generated, so each sample generated this way is likely to be different. We talk more about the variability of random samples in Chapter 3.

²⁶Gelman, A. and Nolan, D., *Teaching Statistics: A Bag of Tricks*, Oxford University Press, New York, 2002.



Random Sampling Caution

In statistics, random is NOT the same as haphazard! We cannot obtain a random sample by haphazardly picking a sample on our own. We must use a formal random sampling method such as technology or drawing names out of a hat.

Realities of Random Sampling

While a random sample is ideal, often it may not be achievable. A list of the entire population may not exist, it may be impossible to contact some members of the population, or it may be too expensive or time consuming to do so. Often we must make do with whatever sample is convenient. The study can still be worth doing, but we have to be very careful when drawing inferences to the population and should at least try to avoid obvious sampling bias as much as possible.

Example 1.16

If the *Chicago Tribune* had wanted to more accurately predict the outcome of the 1948 presidential election, what should they have done? Why do you think they didn't do this?

Solution



To more accurately predict the 1948 presidential election, they should have selected a random sample from the list of all registered voters and then asked the selected people who they would vote for. There are several possible reasons they did not select a random sample. They may not have had a list of all registered voters available from which to sample randomly. Also, collecting data from a random sample of voters might have required traveling to homes all over the country, which would have been time consuming and expensive. Sampling only people with telephones was cheaper and more convenient.

When it is difficult to take a random sample from the population of interest, we may have to redefine the population to which we generalize.

Example 1.17

What Proportion of People Are Vegetarian?

To determine what proportion of people are vegetarian, we would need to take a random sample of all people, which would be extremely difficult or impossible. How might we redefine our population and question so that it is possible to obtain an accurate estimate?

Solution



One option is to narrow our population to those living in Boston and ask, "What proportion of Bostonians are vegetarian?" It would be possible to take a random sample of telephone numbers from a Boston phone book and call and ask whether they eat meat. In this case our population would only include people with land line phone numbers listed in the Boston phone book so would not include people who rely only on cell phones or who have no phone at all.

For simplicity we only describe a simple random sample in detail, but there are other types of random samples. If we want to know the average weight of a population and want to ensure that the proportion of males and females in our sample matches that of the population, we may take two simple random samples, one within males and one within females. For a study on high school students, it is

hard to take a simple random sample. We might first take a simple random sample of schools and then, within each of the sampled schools, take a simple random sample of students. These random sampling schemes are more complicated than the simple random sample but can still yield valid inferences.

Other Sources of Bias

Sampling bias is not the only form of bias that can occur when collecting data. Particularly when collecting data on humans, even if we have a good random sample, there are other issues that might bias the results.

Bias

Bias exists when the method of collecting data causes the sample data to inaccurately reflect the population.



Bias can occur when people we have selected to be in our sample choose not to participate. If the people who choose to respond would answer differently than the people who choose not to respond, results will be biased.

Example 1.18

In 1997 in Somerset (a county in England), a study was conducted on lifestyle choices associated with health.²⁷ A random sample of 6009 residents of Somerset were mailed a questionnaire that they were asked to fill out and return, and 57.6% of the people in the sample returned the questionnaire. Do you think health-related behavior such as exercise and smoking are accurately portrayed by the data collected?

Solution



Probably not. People who returned the questionnaire may have been more proud of their responses, or may have been more organized and motivated in general, so more likely to lead a healthy lifestyle.

The researchers followed up with phone interviews for a random sample of those who had not responded. As suspected, the non-responders were quite different regarding health behaviors. For example, only 35.9% of initial responders reported getting no moderate or vigorous physical activity, while this percentage was almost doubled, 69.6%, for non-responders. Using only the data from the initial responders is very misleading.



The way questions are worded can also bias the results. In 1941 Daniel Rugg²⁸ asked people the same question in two different ways. When asked “Do you think that the United States should allow public speeches against democracy?” 21% said the speeches should be allowed. However, when asked “Do you think that the United States should forbid public speeches against democracy?” 39% said the speeches should not be forbidden. Merely changing the wording of the question nearly doubled the percentage of people in favor of allowing (not forbidding) public speeches against democracy.

²⁷Hill, A., Roberts, J., Ewings, P., and Gunnell, D., “Non-response bias in a lifestyle survey,” *Journal of Public Health Medicine*, June 1997; 19(2): 203–207.

²⁸Rugg, D., “Experiments in wording questions,” *Public Opinion Quarterly*, 1941; 5: 91–92.



iStock.com/Carmen Martínez Banús

Would you have children?

DATA 1.4

“If You Had It to Do Over Again, Would You Have Children?”

In 1976, a young couple wrote to the popular columnist Ann Landers, asking for advice on whether or not to have children.²⁹ Ann ran the letter from the young couple (which included various reasons not to have kids, but no positive aspects of parenting) and asked her readers to respond to the question “If you had it to do over again, would you have children?” Her request for data yielded over 10,000 responses, and to her surprise, only 30% of readers answered “Yes.” She later published these results in *Good Housekeeping*, writing that she was “stunned, disturbed, and just plain flummoxed” by the survey results. She again asked readers to answer the exact same question, and this time 95% of responders said “Yes.” ■

Example 1.19

In Data 1.4, why do you think the two percentages, 30% and 95%, are so drastically different?

Solution



The initial request for data was in a column with a letter stating many reasons not to have kids, which may have brought these issues to the minds of the responders. The second request was in an article mentioning Ann Landers’ dismay at parents answering no, which may have influenced responses. The context in which the question is asked can bias answers one way or another.

Sampling bias is also present, since readers of her column in the newspaper and readers of *Good Housekeeping* and readers who choose to respond to each request for data are probably not representative of the population and probably differ from each other. For the first request, people with more negative experiences with children may have been encouraged to respond, while the opposite may have been true in the second case. You may be able to think of additional reasons for the discrepancy in the sample results.

Example 1.20

Suppose you are considering having children and would really like to know whether more parents are happy about having kids or regret their decision. Which percentages in Data 1.4 can you trust? How would you go about collecting data you can trust?

²⁹<http://www.stats.uwo.ca/faculty/bellhouse/stat353annlanders.pdf>.

Solution

- ▶ Since both of these samples only include people who decided to write in (volunteer samples) instead of taking a random sample, both almost definitely contain sampling bias, so neither should be trusted. To collect data you can trust, you should take a random sample of all parents (or perhaps take a random sample of all parents of your nationality).

Newsday took a random sample of all parents in the US, asking the same question as in Data 1.4. In this random sample, 91% said “Yes,” they would have children again if given the choice. This doesn’t mean that exactly 91% of parents are happy they had kids, but because it was a random sample, it does mean that the true percentage is close to 91%. In Chapter 3 we’ll learn how to assess exactly how close we expect it to be. (Notice that the initial sample result of 30% is extremely misleading!)

Bias may also be introduced if people do not answer truthfully. If the sample data cannot be trusted, neither can generalizations from the sample to the population.

Example 1.21

Illicit Drug Use

The National Survey on Drug Use and Health³⁰ selected a random sample of US college students and asked them about illicit drug use, among other things. In the sample, 22.7% of the students reported using illicit drugs in the past year. Do you think this is an accurate portrayal of the percentage of all college students using illicit drugs?

Solution

- ▶ This may be an underestimate. Even if the survey is anonymous, students may be reluctant to report illicit drug use on an official survey and thus may not answer truthfully.

Bias in data collection can result in many other ways not discussed here. The most important message is to always think critically about the way data are collected and to recognize that not all methods of data collection lead to valid inferences. Recognizing sources of bias is often simply common sense, and you will instantly become a more statistically literate individual if, each time you are presented with a statistic, you just stop, inquire, and think about how the data were collected.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- ▶ • Distinguish between a sample and a population
- ▶ • Recognize when it is appropriate to use sample data to infer information about the population
- ▶ • Critically examine the way a sample is selected, identifying possible sources of sampling bias
- ▶ • Recognize that random sampling is a powerful way to avoid sampling bias
- ▶ • Identify other potential sources of bias that may arise in studies on humans

³⁰Substance Abuse and Mental Health Services Administration, Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings (Office of Applied Studies, NSDUH Series H-38A, HHS Publication No. SMA 10-4856Findings), Rockville, MD, 2010, <https://nsduhweb.rti.org/>.

Exercises for Section 1.2

SKILL BUILDER 1

In Exercises 1.34 to 1.37, state whether the data are best described as a population or a sample.

1.34 To estimate size of trout in a lake, an angler records the weight of 12 trout he catches over a weekend.

1.35 A subscription-based music website tracks its total number of active users.

1.36 The US Department of Transportation announces that of the 250 million registered passenger vehicles in the US, 2.1% are electro-gas hybrids.

1.37 A questionnaire to understand athletic participation on a college campus is emailed to 50 college students, and all of them respond.

SKILL BUILDER 2

In Exercises 1.38 to 1.41, describe the sample and describe a reasonable population.

1.38 A sociologist conducting a survey at a mall interviews 120 people about their cell phone use.

1.39 Five hundred Canadian adults are asked if they are proficient on a musical instrument.

1.40 A cell phone carrier sends a satisfaction survey to 100 randomly selected customers.

1.41 The Nielsen Corporation attaches databoxes to televisions in 1000 households throughout the US to monitor what shows are being watched and produce the Nielsen Ratings for television.

SKILL BUILDER 3

In Exercises 1.42 to 1.45, a biased sampling situation is described. In each case, give:

- The sample
- The population of interest
- A population we can generalize to given the sample

1.42 To estimate the proportion of Americans who support changing the drinking age from 21 to 18, a random sample of 100 college students are asked the question “Would you support a measure to lower the drinking age from 21 to 18?”

1.43 To estimate the average number of tweets from all twitter accounts in 2019, one of the authors randomly selected 10 of his followers and counted their tweets.

1.44 To investigate interest across all residents of the US in a new type of ice skate, a random sample of 1500 people in Minnesota are asked about their interest in the product.

1.45 To determine the height distribution of female high school students, the rosters are collected from 20 randomly selected high school girls basketball teams.

SKILL BUILDER 4

In Exercises 1.46 to 1.51, state whether or not the sampling method described produces a random sample from the given population.

1.46 The population is incoming students at a particular university. The name of each incoming student is thrown into a hat, the names are mixed, and 20 names (each corresponding to a different student) are drawn from the hat.

1.47 The population is the approximately 25,000 protein-coding genes in human DNA. Each gene is assigned a number (from 1 to 25,000), and computer software is used to randomly select 100 of these numbers yielding a sample of 100 genes.

1.48 The population is all employees at a company. All employees are emailed a link to a survey.

1.49 The population is adults between the ages of 18 and 22. A sample of 100 students is collected from a local university, and each student at the university had an equal chance of being selected for the sample.

1.50 The population is all trees in a forest. We walk through the forest and pick out trees that appear to be representative of all the trees in the forest.

1.51 The population is all people who visit the website *CNN.com*. All visitors to the website are invited to take part in the daily online poll.

IS IT BIASED?

In Exercises 1.52 to 1.56, indicate whether we should trust the results of the study. Is the method of data collection biased? If it is, explain why.

1.52 Ask a random sample of students at the library on a Friday night “How many hours a week do you study?” to collect data to estimate the average number of hours a week that all college students study.

1.53 Ask a random sample of people in a given school district, “Excellent teachers are essential to the well-being of children in this community, and teachers truly deserve a salary raise this year.

Do you agree?” Use the results to estimate the proportion of all people in the school district who support giving teachers a raise.

1.54 Take 10 apples off the top of a truckload of apples and measure the amount of bruising on those apples to estimate how much bruising there is, on average, in the whole truckload.

1.55 Take a random sample of one type of printer and test each printer to see how many pages of text each will print before the ink runs out. Use the average from the sample to estimate how many pages, on average, all printers of this type will last before the ink runs out.

1.56 Send an email to a random sample of students at a university asking them to reply to the question: “Do you think this university should fund an ultimate frisbee team?” A small number of students reply. Use the replies to estimate the proportion of all students at the university who support this use of funds.

1.57 Do Parents Regret Having Children? In Data 1.4 on page 25, we describe the results of a question asked by a national newspaper columnist: “If you had it to do over again, would you have children?” In addition to those results and a follow-up national survey, the *Kansas City Star* selected a random sample of parents from Kansas City and asked them the same question. In this sample, 94% said “Yes.” To what population can this statistic be generalized?

1.58 Wearing a Uniform to Work The website *fox6now.com* held an online poll in June 2015 asking “What do you think about the concept of having an everyday uniform for work, like Steve Jobs did?” Of the people who answered the question, 24% said they loved the idea, 58% said they hated the idea, and 18% said that they already wore a uniform to work.

- Are the people who answered the poll likely to be representative of all adult workers? Why or why not?
- Is it reasonable to generalize this result and estimate that 24% of all adult workers would like to wear a uniform to work?

1.59 Canadians Stream Music In a random sample of 3500 Canadian consumers, about 71% report that they regularly stream music.³¹

³¹“What Moves Today’s Teenage Canadian Music Fan?,” <http://www.nielsen.com/ca/en/insights/news/2015/what-moves-todays-teenage-canadian-music-fan.html>, Nielsen, Media and Entertainment, June 2, 2015.

- Is the sample likely to be representative of all Canadian consumers? Why or why not?
- Is it reasonable to generalize this result and estimate that about 71% of all Canadian consumers regularly stream music?

1.60 Climate Change In June 2018, a poll asked a random sample of 1000 US adults whether global warming will be a serious problem for the United States.³² The results show that 51% think global warming will be a very serious problem, 27% think it will be a somewhat serious problem, and 21% think it will not be a serious problem.

- What is the sample? What is the intended population?
- Is it reasonable to generalize this result and estimate that 21% of US adults think that global warming will not be a serious problem for the United States?

1.61 Do You Use a Food Delivery App? A 2019 study conducted by eMarketer³³ asked 800 US smartphone users whether they had used a food delivery app at least once in the last month. The survey also asked which food delivery app, if any, was used. The survey showed that 16.3% of respondents had used a food delivery app in the last month. Of those that had used one, 27.6% used DoorDash, 26.7% used Grubhub, 25.2% used UberEats, while the rest used another.

- What is the sample? What is the intended population?
- What are the cases? What are the variables? Classify variables as quantitative or categorical.

1.62 Pennsylvania High School Seniors Exercise 1.24 describes a dataset, stored in **PA Seniors**, for a sample of students who filled out a survey though the US Census at School project. When downloading the sample³⁴ we specified Pennsylvania as the state and Grade 12 as the school year, then the website chose a random sample of 457 students from among all students who matched those criteria. We’d like to generalize results from this sample to a larger population. Discuss whether this would be reasonable for each of the groups listed below.

³²<https://www.rff.org/energy-and-climate/surveying-american-attitudes-toward-climate-change-and-clean-energy/>. Accessed July 2019.

³³“US Food Delivery App Usage Will Approach 40 Million Users in 2019,” www.eMarketer.com, July 2, 2019. Sample size is approximated.

³⁴Sample data obtained Data from U.S. Census at School (<https://www.amstat.org/censusatschool>) used with the permission of the American Statistical Association.

- (a) The 457 students in the original sample
- (b) All Pennsylvania high school seniors who participated in the Census at School survey
- (c) All Pennsylvania high school seniors
- (d) All students in the United States who participated in the Census at School survey

1.63 How Easily are You Influenced? Mentally simulate ten tosses of a coin by writing down a sequence of Heads and Tails that might result from ten flips of a fair coin. (Try this now!) When a random sample of people were asked to do this, over 80% of them wrote down Heads as the first flip. Expanding on this result,³⁵ when the instructions asked for a sequence of “Tails and Heads,” participants were more likely to put Tails as the first flip. Indeed, when they were told that an imaginary coin was purple on one side and orange on the other (with the two colors presented in random order), participants were more likely to start with whichever color was mentioned first.³⁶ Researchers could influence the results just based on the order in which they listed the options.

- (a) Is this an illustration of sampling bias or wording bias or both or neither?
- (b) If you are letting a friend choose between Option *Q* and Option *W*, and you are really hoping that they pick Option *W*, in what order should you present the options?

1.64 Does Chocolate Milk Come from Brown Cows? The US National Dairy Council, a dairy advocacy group, conducted a survey that appears to show that 7% of Americans (which is about 16.4 million people) believe that chocolate milk comes from brown cows. This result was picked up and shared widely, including by CNN.com, the Washington Post, the Today show, and NPR (National Public Radio).³⁷ The goal of the survey was to find some fun facts to share and the advocacy group has not made the full survey results, or the sampling method, publicly available.

- (a) Do you think it is likely that there is sampling bias in this study? Without knowing how the sample was determined, can we know if it is appropriate to generalize to all Americans?

³⁵Bar-Hillel M, Peer E, Acquisti A, “Heads or tails? A Reachability Bias in Binary Choice,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), April 28, 2014.

³⁶This is called the primacy effect, in which the first option given is more likely to be selected.

³⁷Griffin L and Campbell T, “Take that chocolate milk survey with a grain of salt,” *The Conversation*, June 28, 2017.

Can you think of a way in which the sample might have been determined that would create sampling bias?

- (b) While we don’t know how the sample was determined, interviewers on NPR were able to find out how the question was phrased.³⁸ Survey respondents were asked to select one option to the following: “Where does chocolate milk come from? (a) Brown cows, (b) Black and white spotted cows, (c) I don’t know.” Do you think the way the question was worded might have biased the results? Give a different possible way to word the question that might give more accurate results.

1.65 How Many People Wash Their Hands after Using the Washroom? In Example 1.10 on page 18, we introduce a study by researchers from Harris Interactive who were interested in determining what percent of people wash their hands after using the washroom. They collected data by standing in public restrooms and pretending to comb their hair or put on make-up as they observed patrons’ behavior.³⁹ Public restrooms were observed at Turner’s Field in Atlanta, Penn Station and Grand Central Station in New York, the Museum of Science and Industry and the Shedd Aquarium in Chicago, and the Ferry Terminal Farmers Market in San Francisco. Of the over 6000 people whose behavior was observed, 85% washed their hands. Women were more likely to wash their hands: 93% of women washed, while only 77% of men did. The Museum of Science and Industry in Chicago had the highest hand-washing rate, while men at Turner’s Field in Atlanta had the lowest.

- (a) What are the cases? What are the variables? Classify each variable as quantitative or categorical.
- (b) In a separate telephone survey of more than 1000 adults, more than 96% said they always wash their hands after using a public restroom. Why do you think there is such a discrepancy in the percent from the telephone survey compared to the percent observed?

1.66 Teaching Ability In a sample survey of professors at the University of Nebraska, 94% of them described themselves as “above average” teachers.⁴⁰

³⁸Meikle G, “This survey is as murky as chocolate milk,” *Columbia Journalism Review*, June 21, 2017.

³⁹Bakalar, “Study: More people washing hands after using bathroom,” *Salem News*, September 14, 2010.

⁴⁰Cross, P., “Not can, but *will* college teaching be improved?,” *New Directions for Higher Education*, 1977; 17: 115.

- (a) What is the sample? What is the population?
- (b) Based on the information provided, can we conclude that the study suffers from sampling bias?
- (c) Is 94% a good estimate for the percentage of above-average teachers at the University of Nebraska? If not, why not?

1.67 Effects of Alcohol and Marijuana In 1986 the Federal Office of Road Safety in Australia conducted an experiment to assess the effects of alcohol and marijuana on mood and performance.⁴¹ Participants were volunteers who responded to advertisements for the study on two rock radio stations in Sydney. Each volunteer was given a randomly determined combination of the two drugs, then tested and observed. Is the sample likely representative of all Australians? Why or why not?

1.68 What Percent of Young Adults Move Back in with Their Parents? The Pew Research Center polled a random sample of $n = 808$ US residents between the ages of 18 and 34. Of those in the sample, 24% had moved back in with their parents for economic reasons after living on their own.⁴² Do you think that this sample of 808 people is a representative sample of all US residents between the ages of 18 and 34? Why or why not?

1.69 Do Cat Videos Improve Mood? As part of an “internet cat videos/photos” study, Dr. Jessica Gall Myrick posted an on-line survey to Facebook and Twitter asking a series of questions regarding how individuals felt before and after the last time they watched a cat video on the Internet.⁴³ One of the goals of the study was to determine how watching cat videos affects an individual’s energy and emotional state. People were asked to share the link, and everyone who clicked the link and completed the survey was included in the sample. More than 6000 individuals completed the survey, and the study found that after watching a cat video people generally reported more energy, fewer negative emotions, and more positive emotions.

⁴¹Chesher, G., Dauncey, H., Crawford, J., and Horn, K., “The Interaction between Alcohol and Marijuana: A Dose Dependent Study on the Effects on Human Moods and Performance Skills,” Report No. C40, Federal Office of Road Safety, Federal Department of Transport, Australia, 1986.

⁴²Parker, K., “The Boomerang Generation: Feeling OK about Living with Mom and Dad,” Pew Research Center, March 15, 2012.

⁴³Gall Myrick, J., “Emotion regulation, procrastination, and watching cat videos online: Who watches Internet cats, why, and to what effect?,” *Computers in Human Behavior*, June 12, 2015.

- (a) Would this be considered a simple random sample from a target population? Why or why not?
- (b) Ignoring sampling bias, what other ways could bias have been introduced into this study?

1.70 Diet Cola and Weight Gain in Rats A study⁴⁴ fed one group of rats a diet that included yogurt sweetened with sugar, and another group of rats a diet that included yogurt sweetened with a zero-calorie artificial sweetener commonly found in diet cola. The rats that were fed a zero-calorie sweetener gained more weight and more body fat compared to the rats that were fed sugar. After the study was published, many news articles discussed the implication that people who drink diet soda gain more weight. Explain why we cannot conclude that this is necessarily true.

1.71 Armoring Military Planes During the Second World War, the U.S. military collected data on bullet holes found in B-24 bombers that returned from flight missions. The data showed that most bullet holes were found in the wings and tail of the aircraft. Therefore, the military reasoned that more armor should be added to these regions, as they are more likely to be shot. Abraham Wald, a famous statistician of the era, is reported to have argued against this reasoning. In fact, he argued that based on these data more armor should be added to the center of the plane, and NOT the wings and tail. What was Wald’s argument?

1.72 Employment Surveys Employment statistics in the US are often based on two nationwide monthly surveys: the Current Population Survey (CPS) and the Current Employment Statistics (CES) survey. The CPS samples approximately 60,000 US households and collects the employment status, job type, and demographic information of each resident in the household. The CES survey samples 140,000 nonfarm businesses and government agencies and collects the number of payroll jobs, pay rates, and related information for each firm.

- (a) What is the population in the CPS survey?
- (b) What is the population in the CES survey?
- (c) For each of the following statistical questions, state whether the results from the CPS or CES survey would be more relevant.

i. Do larger companies tend to have higher salaries?

⁴⁴Swithers, S.E., Sample, C.H., and Davidson, T.L., “Adverse effects of high-intensity sweeteners on energy intake and weight control in male and obesity-prone female rats.” *Behavioral neuroscience*, 2013; 127(2), 262.

- ii. What percentage of Americans are self-employed?
- iii. Are married men more or less likely to be employed than single men?

1.73 National Health Statistics The Centers for Disease Control and Prevention (CDC) administers a large number of survey programs for monitoring the status of health and health care in the US. One of these programs is the National Health and Nutrition Examination Survey (NHANES), which interviews and examines a random sample of about 5000 people in the US each year. The survey includes questions about health, nutrition, and behavior, while the examination includes physical measurements and lab tests. Another program is the National Hospital Ambulatory Medical Care Survey (NHAMCS), which includes information from hospital records for a random sample of individuals treated in hospital emergency rooms around the country.

- (a) To what population can we reasonably generalize findings from the NHANES?
- (b) To what population can we reasonably generalize findings from the NHAMCS?
- (c) For each of the questions below, indicate which survey, NHANES or NHAMCS, would probably be more appropriate to address the issue.
 - i. Are overweight people more likely to develop diabetes?
 - ii. What proportion of emergency room visits in the US involve sports-related injuries?

- iii. Is there a difference in the average waiting time to be seen by an emergency room physician between male and female patients?
- iv. What proportion of US residents have visited an emergency room within the past year?

1.74 Interviewing the Film Crew on Hollywood Movies There were 1295 movies made in Hollywood between 2012 and 2018. Suppose that, for a documentary about Hollywood film crews, a random sample of 5 of these movies will be selected for in-depth interviews with the crew members. Assuming the movies are numbered 1 to 1295, use a random number generator or table to select a random sample of five movies by number. Indicate which numbers were selected. (If you want to know which movies you selected, check out the dataset **HollywoodMovies**.)

1.75 Sampling Some Starbucks Stores The Starbucks chain has about 24,000 retail stores in 70 countries.⁴⁵ Suppose that a member of the Starbucks administration wishes to visit six of these stores, randomly selected, to gather some first-hand data. Suppose the stores are numbered 1 to 24,000. Use a random number generator or table to select the numbers for 6 of the stores to be in the sample.

⁴⁵<https://www.starbucks.com/about-us/company-information/starbucks-company-profile>.

1.3 EXPERIMENTS AND OBSERVATIONAL STUDIES

Association and Causation

Three neighbors in a small town in northern New York State enjoy living in a climate that has four distinct seasons: warm summers, cold winters, and moderate temperatures in the spring and fall. They also share an interest in using data to help make decisions about questions they encounter at home and at work.

- Living in the first house is a professor at the local college. She's been looking at recent heating bills and comparing them to data on average outside temperature. Not surprisingly, when the temperature is lower, her heating bills tend to be much higher. She wonders, "It's going to be an especially cold winter; should I budget for higher heating costs?"
- Her neighbor is the plant manager for a large manufacturing plant. He's also been looking at heating data and has noticed that when the building's heating plant is used, there are more employees missing work due to back pain or colds and flu. He wonders, "Could emissions from the heating system be having adverse health effects on the workers?"

- The third neighbor is the local highway superintendent. He is looking at data on the amount of salt spread on the local roads and the number of auto accidents. (In northern climates, salt is spread on roads to help melt snow and ice and improve traction.) The data clearly show that weeks when lots of salt is used also tend to have more accidents. He wonders, “Should we cut down on the amount of salt we spread on the roads so that we have fewer accidents?”

Each of these situations involves a relationship between two variables. In each scenario, variations in one of the variables tend to occur in some regular way with changes in the other variable: lower temperatures go along with higher heating costs, more employees have health issues when there is more activity at the heating plant, and more salt goes with more accidents. When this occurs, we say there is an *association* between the two variables.

Association

Two variables are **associated** if values of one variable tend to be related to the values of the other variable.

The three neighbors share a desirable habit of using data to help make decisions, but they are not all doing so wisely. While colder outside temperatures probably force the professor’s furnace to burn more fuel, do you think that using less salt on icy roads will make them safer? The key point is that an association between two variables, even a very strong one, does not imply that there is a *cause and effect* relationship between the two variables.

Causation

Two variables are **causally associated** if changing the value of one variable influences the value of the other variable.

The distinction between association and causation is subtle, but important. In a causal relationship, manipulating one of the variables tends to cause a change in the other. For example, we put more pressure on the gas pedal and a car goes faster. When an association is not causal, changing one of the variables will not produce a predictable change in the other. Causation often implies a particular direction, so colder outside temperatures might cause a furnace to use more fuel to keep the professor’s house warm, but if she increases her heating costs by buying more expensive fuel, we should not expect the outdoor temperatures to fall!

Recall from Section 1.1 that values of an explanatory variable might help predict values of a response variable. These terms help us make the direction of a causal relationship more clear: We say changing the explanatory variable tends to cause the response variable to change. A causal statement (or any association statement) means that the relationship holds as an overall trend—not necessarily in every case.

Example 1.22

For each sentence discussing two variables, state whether the sentence implies no association between the variables, association without implying causation, or association with causation. If there is causation, indicate which variable is the explanatory variable and which is the response variable.

- Studies show that taking a practice exam increases your score on an exam.
- Families with many cars tend to also own many television sets.

- (c) Sales are the same even with different levels of spending on advertising.
- (d) Taking a low-dose aspirin a day reduces the risk of heart attacks.
- (e) Goldfish who live in large ponds are usually larger than goldfish who live in small ponds.
- (f) Putting a goldfish into a larger pond will cause it to grow larger.

Solution



- (a) This sentence implies that, in general, taking a practice exam *causes* an increase in the exam grade. This is association with causation. The explanatory variable is whether or not a practice exam was taken and the response variable is the score on the exam.
- (b) This sentence implies association, since we are told that one variable (number of TVs) tends to be higher when the other (number of cars) is higher. However, it does not imply causation since we do not expect that buying another television set will somehow cause us to own more cars, or that buying another car will somehow cause us to own more television sets! This is association without causation.
- (c) Because sales don't vary in any systematic way as advertising varies, there is no association.
- (d) This sentence indicates association with causation. In this case, the sentence makes clear that a daily low-dose aspirin *causes* heart attack risk to go down. The explanatory variable is taking aspirin and the response variable is heart attack risk.
- (e) This sentence implies association, but it only states that larger fish tend to be in larger ponds, so it does not imply causation.
- (f) This sentence implies association with causation. The explanatory variable is the size of the pond and the response variable is the size of the goldfish.

Contrast the sentences in Example 1.22 parts (e) and (f). Both sentences are correct, but one implies causation (moving to a larger pond makes the fish grow bigger) and one does not (bigger fish just happen to reside in larger ponds). Recognizing the distinction is important, since implying causation incorrectly is one of the most common mistakes in statistics. Try to get in the habit of noticing when a sentence implies causation and when it doesn't.

Many decisions are made based on whether or not an association is causal. For example, in the 1950s, people began to recognize that there was an association between smoking and lung cancer, but there was a debate that lasted for decades over whether smoking *causes* lung cancer. It is now generally agreed that smoking causes lung cancer, and this has led to a substantial decline in smoking rates in the US. The fact that smoking causes lung cancer does not mean that everyone who smokes will get lung cancer, but it does mean that people who smoke are more likely to get it (in fact, 10 to 20 times more likely⁴⁶). Other causal questions, such as whether cell phones cause cancer or whether increasing online advertising would cause sales to increase, remain topics of research and debate. One of the goals of this section is to help you determine when a study can, and cannot, establish causality.

⁴⁶http://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm#1.

Confounding Variables

Why are some variables associated even when they have no cause and effect relationship in either direction? As the next example illustrates, the reason is often the effect of other variables.

DATA 1.5

Vehicles and Life Expectancy

The US government collects data from many sources on a yearly basis. For example, Table 1.5 shows the number of vehicles (in millions) registered in the US⁴⁷ and the average life expectancy (in years) of babies born⁴⁸ in the US every four years from 1970 to 2014. A more complete dataset with values for each of the years from 1970 through 2017 is stored in **LifeExpectancyVehicles**. If we plot the points in Table 1.5, we obtain the graph in Figure 1.2. (This graph is an example of a scatterplot, which we discuss in Chapter 2.) As we see in the table and the graph, these two variables are very strongly associated; the more vehicles that are registered, the longer people are expected to live. ■

Table 1.5 *Vehicle registrations (millions) and life expectancy*

Year	Vehicles	Life Expectancy
1970	108.4	70.8
1974	129.9	72.0
1978	148.4	73.5
1982	159.6	74.5
1986	175.7	74.7
1990	188.8	75.4
1994	198.0	75.7
1998	211.6	76.7
2002	229.6	77.3
2006	244.2	77.7
2010	242.1	78.7
2014	260.4	78.9

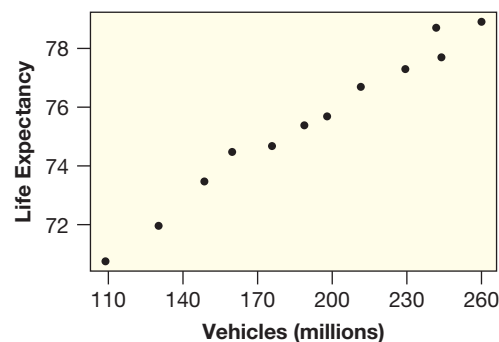


Figure 1.2 *A strong association between vehicles and life expectancy*

⁴⁷Vehicle registrations from US Federal Highway Administration, <https://www.fhwa.dot.gov/policy/information/statistics.cfm>.

⁴⁸Centers for Disease Control and Prevention, National Center for Health Statistics, Health Data Interactive, www.cdc.gov/nchs/hdi.htm, Accessed October 2019.

There is a clear association between vehicle registrations and life expectancy. Is this a *causal* association? If so, which way might it go? Do people live longer because they have a car to drive? When people live longer, do they have time to buy more vehicles? Or is there something else driving this association?

Confounding Variable

A **confounding variable**, also known as a **confounding factor** or **lurking variable**,⁴⁹ is a third variable that is associated with both the explanatory variable and the response variable. A confounding variable can offer a plausible explanation for an association between two variables of interest.

Example 1.23

Describe a possible confounding variable in Data 1.5 about vehicle registrations and life expectancy.

Solution



One confounding variable is the year. As time goes along, the population grows so more vehicles are registered and improvements in medical care help people live longer. Both variables naturally tend to increase as the year increases and may have little direct relationship with each other. The years are an explanation for the association between vehicle registrations and life expectancy.



When faced with a strong association such as that between vehicles and life expectancy, it can be tempting to immediately jump to conclusions of causality. However, it is important to stop and think about whether there are confounding variables which could be explaining the association instead.

Example 1.24

In 2008, the *Los Angeles Times* published a headline⁵⁰ that included “Hospitals... Riskier than a Casino in Event of Cardiac Arrest.” The article, based on a study published in the *New England Journal of Medicine*,⁵¹ states that the widespread availability of defibrillators and bystanders in public places like casinos leads to a higher survival rate than hospitals in the case of cardiac arrest.

- What are the primary variables of interest in this study? Which is the explanatory variable and which is the response variable?
- Give an example of one potential confounding variable in this study.
- If you are having a heart attack, would you go to a hospital or a casino?

Solution



- The two primary variables of interest are the place of cardiac arrest (explanatory) and whether or not the person survives (response).
- A confounding variable is the health of the person at the time of cardiac arrest. Older, frailer, sicker people are more likely to be in the hospital and also less likely to survive (not because they are in a hospital, but just because they are weaker to begin with). Someone in a casino is much more likely to be in better

⁴⁹Some statisticians distinguish between confounding variables and lurking variables. However, for simplicity in this book we treat them as synonymous.

⁵⁰Maugh, T., “Study finds hospitals slow to defibrillate: Researchers say they’re riskier than a casino in event of cardiac arrest,” *Los Angeles Times*, January 3, 2008.

⁵¹Chan, P., Krumholz, H., Nichol, G., and Nallamothu, B., American Heart Association National Registry of Cardiopulmonary Resuscitation Investigators, “Delayed Time to Defibrillation after In-Hospital Cardiac Arrest,” *New England Journal of Medicine*, 2008; 358: 9–17.

physical shape, and thus better able to survive a heart attack. Notice that the confounding variable (health of the person) influences *both* of the variables of interest: where the person might be and whether the person is likely to survive.

- (c) If you are having a heart attack, you should go to a hospital! Even though casinos have a higher survival rate, this can be explained by the confounding variable, and we cannot conclude that being in a casino *causes* a higher survival rate. For a person of a given health status, it is probably safer to be in a hospital under the care of professionals.
-

Many seemingly surprising claims in the media (such as that hospitals are riskier than casinos) can be explained simply by the presence of a confounding variable. Knowing how and when to be on the lookout for confounding variables is essential for statistical literacy and for assessing any data-based claims.

Observational Studies vs Experiments

How can we establish (statistically) when an association represents a causal relationship? The key is in how the data are collected. If we want to study how the explanatory variable influences the response variable, we have to be able to control or specify the values of the explanatory variable to make sure it is not associated with any potential confounding variables.

Note that in data such as **Life Expectancy Vehicles** or the study of cardiac arrest we merely collect available data after the fact. We call data collected in this way, with no effort or ability to manipulate the variables of interest, an *observational study*. With observational data we can never be certain that an apparent association is not due to some confounding variable, and thus the association is not evidence of a causal relationship.

The alternative is to intentionally control one or more of the explanatory variables when producing the data to see how the response variable changes. We call this method of data collection a *statistical experiment*. With a well-designed experiment, we can make conclusions about causation when we see a strong association, since the method for assigning the values of the explanatory variable(s) are not influenced by any confounding variables.

Observational Studies and Experiments

An **experiment** is a study in which the researcher actively controls one or more of the explanatory variables.

An **observational study** is a study in which the researcher does not actively control the value of any variable but simply observes the values as they naturally exist.

Example 1.25

Both studies below are designed to examine the effect of fertilizer on the yield of an apple orchard. Indicate whether each method of collecting the data is an experiment or an observational study.

- (a) Researchers find several different apple orchards and record the amount of fertilizer that had been used and the yield of the orchards.
- (b) Researchers find several different apple orchards and assign different amounts of fertilizer to each orchard. They record the resulting yield from each.

Solution

- ▶ (a) This is an observational study, since data were recorded after the fact and no variables were actively manipulated. Notice that there are many possible confounding variables that might be associated with both the amount of fertilizer used and the yield, such as the quality of soil.
- (b) This is an experiment since the amount of fertilizer was *assigned* to different orchards. The researchers actively manipulated the assignment of the fertilizer variable, in order to determine the effect on the yield variable.



Al Diaz/Miami Herald/Getty Images

Do high fives help teams win?

Example 1.26

Basketball High Fives

In the 2011 NBA (National Basketball Association) finals, the Dallas Mavericks defeated the Miami Heat. One headline on NBC sports⁵² stated, “Miami’s real problem this series: Not enough high fives,” citing a study⁵³ that found that teams exhibiting more “touching,” such as high fives, early in the season had better performance later in the season. Is this study an experiment or an observational study? Does the study provide evidence that additional high fiving improves basketball performance?

Solution

- ▶ The study is an observational study, because researchers did not manipulate or assign the number of high fives. The word “improves” implies causality, but because this was an observational study, confounding variables are likely and causality cannot be established. This study does not provide evidence that giving high fives improves basketball performance.

One possible confounding variable in Example 1.26 is how well a team gets along, which is likely to be associated both with the number of high fives and a team’s performance. While we consider methods to account for some confounding

⁵²<http://probasketballtalk.nbcsports.com/2011/06/09/miami-s-real-problem-this-series-not-enough-high-fives/>.

⁵³Kraus, M., Huang, C., and Keltner, D., “Tactile communication, cooperation, and performance: An ethological study of the NBA,” *Emotion*, 2010; 10(5): 745–749.

variables later in this text, additional confounding variables may still exist. In an observational study, there is no way of guaranteeing that we haven't missed one.



Causation Caution

It is difficult to avoid confounding variables in observational studies. For this reason, observational studies can almost never be used to establish causality.

Randomized Experiments

In an experiment, the researcher controls the assignment of one or more variables. This power can allow the researcher to avoid confounding variables and identify causal relationships, if used correctly. But how can the researcher possibly avoid all potential confounding variables? The key is surprisingly simple: a *randomized experiment*. Just as randomness solved the problem of sampling bias, randomness can also solve the problem of confounding variables.

Randomized Experiment

In a **randomized experiment** the value of the explanatory variable for each unit is determined randomly, before the response variable is measured.

If a randomized experiment yields an association between the two variables, we can establish a causal relationship from the explanatory to the response variable.

In a randomized experiment, we don't expect the explanatory variable to be associated with any other variables at the onset of the study, because its values were determined simply by random chance. If nothing is associated with the explanatory variable, then confounding variables do not exist! For this reason, if an association is found between the explanatory and response variables in a randomized experiment, we can conclude that it indeed is a causal association.

Evaluating Evidence for Causality

There are three explanations for why an association may be observed in sample data:

- (i) There is a causal association.
- (ii) There is an association, but it is due to confounding.
- (iii) There is no actual association; the association seen in the sample data occurred by random chance.

Evaluating evidence *for* the causal explanation requires evaluating evidence *against* the other two competing explanations. The best evidence against (ii), confounding, is the use of random assignment. We'll learn how to evaluate evidence against (iii), "just random chance," in Chapter 4.

Example 1.27*Diet and Depression I*

A meta-analysis is a study that takes existing studies on a single topic and combines them all into one study. A meta-analysis⁵⁴ combined 13 different observational studies that examined the association between diet and depression among adults, and concluded that people who ate a healthy diet were significantly less likely to be depressed.

- What is the explanatory variable? What is the response variable?
- Does this provide convincing evidence that eating a healthy diet causes a reduced risk of depression? Why or why not?
- What might be a confounding variable?

Solution

- The explanatory variable is healthy diet or not; the response variable is depression or not.
- No. These results are based on a combination of observational studies, so we cannot determine whether the observed association is actually causal, or just due to confounding.
- One possible confounding variable is exercise. People who exercise more tend to have better dietary habits, and exercise is known to be associated with lower levels of depression. In the presence of this confounding variable, we cannot determine whether it is the diet or the exercise that is making people less depressed.

Example 1.28*Diet and Depression II*

A group of researchers⁵⁵ in Australia conducted a short (three-week) dietary intervention in a randomized controlled experiment. In the study, 75 college-age students with elevated depression symptoms and relatively poor diet habits were randomly assigned to either a healthy diet intervention group or a control group. The researchers recorded whether the participants were male or female, and also the change over the three-week period on two different numeric scales of depression (the CESD-R scale and the DASS scale). The study concluded that even just a three-week change in diet habits is significantly associated with lower depression levels. Can we conclude from this study that improving dietary habits, even for only three weeks, *causes* lower depression levels? Why or why not?

Solution

Yes! People were *randomly assigned* to the healthy diet intervention or the control group, so we do not expect diet to be associated with anything else, and so do not expect any confounding variables. (The word “significantly” provides evidence against “just random chance,” as we’ll learn in Chapter 4.) Because this was a randomized experiment, we can make causal conclusions!

Confounding and the importance of randomization may be easier to understand in the special case of comparing groups, as in the diet and depression examples, in which case “confounding” means the groups differ to begin with. Without random assignment, the groups will almost surely differ in some way, even before they receive different treatments, in which case it’s very difficult to determine whether

⁵⁴Lai JS, et al., “A systematic review and meta-analysis of dietary patterns and depression in community-dwelling adults,” *The American Journal of Clinical Nutrition*, 99(1), January 2014.

⁵⁵Francis HM, et al., “A brief diet intervention can reduce symptoms of depression in your adults - A randomised controlled trial,” *PLoS ONE*, 14(10), October 2019.

an observed difference in outcomes is caused by the explanatory variable or just the fact that the groups differed to begin with. In Example 1.27, people who choose to eat a healthy diet almost certainly differ from those who don't in many other ways (they are probably more health-conscious in general), so we cannot conclude whether the lower levels of depression were due to the different diets or the differences in the other health-conscious behaviors. If the groups are *randomly assigned*, any pre-treatment differences should be purely random, so the groups should look relatively similar to begin with. In Example 1.28, people were randomly assigned to the healthy diet intervention or the control group, so we expect that these groups will be relatively similar to begin with—before starting the diet, any differences between them will be purely random! In this case, we can conclude that the lower levels of depression were actually due to the diet!

This assessment of the reason for a difference can be generalized to any two groups. Suppose we are comparing *A* versus *B* (two levels of an explanatory variable), and in our sample, the *A* group has different outcomes than the *B* group. The three general explanations for an association can be written more specifically in the special case of a difference between two groups:

- (i) *A* causes different outcomes than *B*.
- (ii) The *A* group and the *B* group differed to begin with.
- (iii) The difference is just due to random chance (not a real difference in outcomes due to *A* and *B*).

Again, random assignment to groups is the best way to rule out explanation (ii).

Recall from Section 1.2 that “random” does not mean haphazard. A formal randomization method (such as flipping a coin, dealing cards, drawing names out of a hat, or using technology) must be used to assign the value of the explanatory variable. This assures that the value of the explanatory variable for each unit is determined by random chance alone and is not influenced by any confounding variables.

The key idea of Section 1.2 was that results from a sample can only be generalized to the population if the sampling units were selected *randomly* from the population. The key idea of this section is that causality can only be established if the values of the explanatory variable are *randomly* assigned to the units. Randomness is the essential ingredient in both cases, but the type of randomness should not be confused. In the first case we are randomly determining which units will be a part of the study. In the second case we are randomly determining which value of the explanatory variable will be assigned to each of the units already selected to be in our sample. Lack of randomness in either stage drastically influences the types of conclusions that can be made: Lack of randomness in sampling prevents generalizing to the population, lack of randomness in assigning the values of the explanatory variable prevents making causal conclusions. See Figure 1.3.

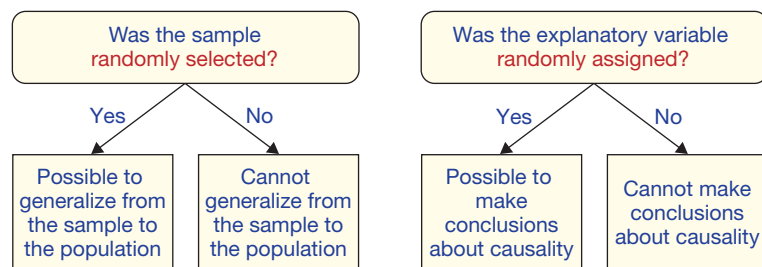


Figure 1.3 Two fundamental questions about data collection

DATA 1.6

Physicians' Health Study

Does anyone you know regularly take a low-dose aspirin? If so, it may be because of a randomized experiment conducted in the 1980s, the Physicians' Health Study.⁵⁶ The study recruited 22,071 male physicians and randomly assigned half of them to take an aspirin every other day for about five years and the other half to take a fake aspirin pill instead. They found that the physicians who took the real aspirin had 44% fewer heart attacks than those taking the fake aspirin. ■

The study in Data 1.6 is a randomized experiment because the researchers randomly determined which physicians received the real aspirin. The physicians themselves had no choice and in fact did not even know which pill they were taking. Because the physicians were split into two groups randomly, the only difference between the groups should be the aspirin. Therefore, we can conclude that the difference in heart attack rates must be *caused* by the aspirin. From this study we can conclude that regularly taking a low-dose aspirin reduces the risk of heart attack.

Many ideas of experimental design were originally developed for medical studies (such as Data 1.6) or agricultural experiments (like the fertilizer example of Example 1.25). For this reason, we often refer to values of the explanatory variable which the researcher controls as *treatments*. In Data 1.6, the treatments are the real aspirin and the fake aspirin.

Example 1.29*Warming Up*

Warming up is a regular part of almost every athlete's pre-game routine, but the optimal amount to warm up is not always known. Cyclists typically have a very intense warm-up, and a study⁵⁷ in 2011 tests whether a shorter, less intense warm-up is better. Ten cyclists were recruited from the Calgary Track Cycling League and completed both a traditional intense warm-up and a shorter, less physically demanding, experimental warm-up. Each cyclist completed each warm-up at different times, and the order in which the warm-ups were performed was randomized. After each warm-up, performance was measured. The study found performance to be better after the shorter warm-up.

- (a) What are the treatments?
- (b) What conclusion can we draw from this study?

Solution

- (a) There are two treatments to be compared: the more intense traditional warm-up and the shorter, less intense, experimental warm-up.
- (b) Because the order of the warm-ups was *randomized*, causal conclusions can be made. The shorter warm-up causes the cyclists to perform better.

⁵⁶The Steering Committee of the Physicians' Health Study Research Group, "Final report on the aspirin component of the ongoing Physicians' Health Study," *New England Journal of Medicine*, 1989; 321: 129–135.

⁵⁷Tomaras, E. and Macintosh, B., "Less is More: Standard Warm-up Causes Fatigue and Less Warm-up Permits Greater Cycling Power Output," *Journal of Applied Physiology*, May 5, 2011.

In Example 1.28 and Data 1.6, it was randomly determined which units got which treatment. In Example 1.29, every unit got both treatments, but the *order* of the treatments was randomly determined. Both ways of randomization yield valid randomized experiments. The former is known as a *randomized comparative experiment* because two groups of units are compared. The latter is known as a *matched pairs experiment*, because each unit forms a pair of data values (one under each treatment), and comparisons are made within each pair. These are only two of many different ways to incorporate randomization into an experiment.

Two Types of Randomized Experiments

In a **randomized comparative experiment**, we randomly assign cases to different treatment groups and then compare results on the response variable(s).

In a **matched pairs experiment**, each case gets both treatments in random order (or cases get paired up in some other obvious way), and we examine individual differences in the response variable between the two treatments.

Example 1.30

Is the Dominant Hand Stronger?

We wish to run an experiment using 30 right-handed people to determine whether gripping strength in the dominant hand is greater than gripping strength in the other hand.

- Describe the experiment if we use a randomized comparative design.
- Describe the experiment if we use a matched pairs design.
- Which design makes more sense in this case?

Solution



- Using a randomized comparative design, we randomly divide the 30 people into two groups of 15 each. We measure gripping strength in the right hand for one of the groups and in the left hand for the other group, and compare results.
- In a matched pairs experiment, we measure the gripping strength in both hands for each of the 30 people. The data are “paired” because we compare the right- and left-handed gripping strength for each person, and examine the difference between the two values. We randomize the order in which participants use the hands: some (randomly determined) doing the right hand first and some the left hand first. Notice that all participants are doing both, unlike in the experiment described in part (a) with two distinct groups each assigned a different treatment.
- A matched pairs experiment makes sense here because hand-gripping strength can vary a great deal between different people and it makes sense to compare a person’s right-hand strength to his or her own left-hand strength.

Control Groups, Placebos, and Blinding

The Physicians’ Health Study illustrates many aspects of a well-designed experiment. The participants who did not take an aspirin pill are an example of a *control group*. Nothing was done to this group that might directly influence the response variable. The control group provides a good comparison for the group that actually got the treatment of interest. Not all good experiments need a control

group. In all cases, however, procedures for handling the groups should match as closely as possible so that effective comparisons can be made.

If people believe they are getting an effective treatment, they may experience the desired effect regardless of whether the treatment is any good. This phenomenon is called the *placebo effect*. Although perhaps not intuitive, the placebo effect has been studied extensively and can be very powerful. A *placebo* is a fake pill or treatment, and placebos are often used to control for the placebo effect in experiments. The fake aspirin pill given to participants in the control group of the Physicians' Health Study is an example of a placebo.

Using a placebo is less helpful, however, if participants know they are not getting the real treatment. This is one of the reasons that *blinding* is so important. In a *single-blind* experiment, the participants are not told which group they are in. In a *double-blind* experiment, the participants are not told which group they are in *and* the people interacting with the participants and recording the results of the response variable also do not know who is in which group. The Physicians' Health Study was double-blind: The people taking the pills did not know whether they were taking an aspirin or a placebo and the doctors treating them and determining who had heart attacks also did not know.

DATA 1.7**Sham Knee Surgery**

For people suffering from arthritis of the knee, arthroscopic surgery has been one possible treatment. In the mid-1990s, a study⁵⁸ was conducted in which 10 men with arthritic knees were scheduled for surgery. They were all treated exactly the same except for one key difference: only some of them actually had the surgery! Once each patient was in the operating room and anesthetized, the surgeon looked at a randomly generated code indicating whether he should do the full surgery or just make three small incisions in the knee and stitch up the patient to leave a scar. All patients received the same post-operative care, rehabilitation, and were later evaluated by staff who didn't know which treatment they had. The result? The men who got the sham knee surgery and the men who got the real knee surgery showed similar and indistinguishable levels of improvement. ■

Example 1.31

Discuss the experiment in Data 1.7. How is randomization used? Is there a placebo? Is the study double-blind? Why did the doctors make incisions in the knees of the men not getting the surgery?

Solution



Randomization was used to divide the men into groups, determining who got the real surgery and who didn't. The placebo was the fake surgery. Because the placebo surgery should match the real surgery as much as possible, those in the placebo group still received incisions and stitches. The men needed similar scars so that both the patients and the staff giving follow-up care were blind as to who actually had surgery done inside their knee. This made the study double-blind.

⁵⁸Talbot, M., "The Placebo Prescription," *The New York Times*, January 9, 2000.

You may wonder whether data from only 10 patients are sufficient to make strong conclusions about the best treatment plan for arthritic knees. That would be a valid concern. In general, we would like to *replicate* each treatment on as many experimental units as is feasible. In many situations a small pilot study, such as the one described in Data 1.7, is used for initial guidance before undertaking a larger, more expensive experiment. In the case of the placebo knee surgery, a follow-up study with 180 patients produced similar results⁵⁹ – indicating that full knee surgery may not be needed for patients with this condition.

Example 1.32

Does an injection of caffeine help rats learn a maze faster? Design an experiment to investigate this question. Incorporate elements of a well-designed experiment.

Solution



We take the rats that are available for the study and *randomly* divide them into two groups. One group will get a shot of caffeine, while the other group will get a shot of saline solution (placebo). We have the rats run the maze and record their times. Don't tell the rats which group they are in! Ideally, all people who come in contact with the rats (the people giving the shots, the people recording the maze times, and so on) should not know which rats are in which group. This makes the study double-blind. Only the statistician analyzing the data will know which rats are in which group. (We describe here a randomized comparative experiment. A matched pairs experiment would also work, and in that case we would also use randomization and a placebo and blinding.)

Realities of Randomized Experiments

Randomization should always be used in designing an experiment. Blinding and the use of a placebo treatment should be used when appropriate and possible. However, there are often ethical considerations that preclude the use of an experiment in any form. For example, imagine designing an experiment to determine whether cell phones cause cancer or whether air pollution leads to adverse health consequences. It would not be appropriate to require people to wear a cell phone on their head for large amounts of time to see if they have higher cancer rates! Similarly, it would not be appropriate to require some people to live in areas with more polluted air. In situations such as these, observational studies can at least help us determine associations.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- ▶ • Recognize that not every association implies causation
- ▶ • Identify potential confounding variables in a study
- ▶ • Distinguish between an observational study and a randomized experiment
- ▶ • Recognize that only randomized experiments can lead to claims of causation
- ▶ • Explain how and why placebos and blinding are used in experiments
- ▶ • Distinguish between a randomized comparative experiment and a matched pairs experiment
- ▶ • Design and implement a randomized experiment

⁵⁹Moseley, J., et al., “A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee,” *The New England Journal of Medicine*, 2002; 347: 81–88.

Exercises for Section 1.3

SKILL BUILDER 1

In Exercises 1.76 to 1.81, we give a headline that recently appeared online or in print. State whether the claim is one of association and causation, association only, or neither association nor causation.

- 1.76** Daily exercise improves mental performance.
- 1.77** Among college students, no link found between number of friends on social networking websites and size of the university.
- 1.78** Cell phone radiation leads to deaths in honey bees.
- 1.79** Wealthy people are more likely than other folks to lie, cheat, and steal.
- 1.80** Cat owners tend to be more educated than dog owners.
- 1.81** Want to lose weight? Eat more fiber!

SKILL BUILDER 2

Exercises 1.82 to 1.87 describe an association between two variables. Give a confounding variable that may help to account for this association.

- 1.82** More ice cream sales have been linked to more deaths by drowning.
- 1.83** The total amount of beef consumed and the total amount of pork consumed worldwide are closely related over the past 100 years.
- 1.84** People who own a yacht are more likely to buy a sports car.
- 1.85** Sales of toboggans tend to be higher when sales of mittens are higher.
- 1.86** Air pollution is higher in places with a higher proportion of paved ground relative to grassy ground.
- 1.87** People with shorter hair tend to be taller.

SKILL BUILDER 3

In Exercises 1.88 to 1.91, we describe data collection methods to answer a question of interest. Are we describing an experiment or an observational study?

- 1.88** To examine whether eating brown rice affects metabolism, we ask a random sample of people whether they eat brown rice and we also measure their metabolism rate.
- 1.89** To examine whether playing music in a store increases the amount customers spend, we randomly assign some stores to play music and some to

stay silent and compare the average amount spent by customers.

1.90 To examine whether planting trees reduces air pollution, we find a sample of city blocks with similar levels of air pollution and we then plant trees in half of the blocks in the sample. After waiting an appropriate amount of time, we measure air pollution levels.

1.91 To examine whether farm-grown salmon contain more omega-3 oils if water is more acidic, we collect samples of salmon and water from multiple fish farms to see if the two variables are related.

REVISITING QUESTIONS FROM SECTION 1.1

Exercises 1.92 to 1.94 refer to questions of interest asked in Section 1.1 in which we describe data collection methods. Indicate whether the data come from an experiment or an observational study.

- 1.92** “Is there a sprinting gene?” Introduced in Example 1.5 on page 9.
- 1.93** “Do metal tags on penguins harm them?” Introduced in Data 1.3 on page 10.
- 1.94** “Are there human pheromones?” Introduced on page 11. Three studies are described; indicate whether each of them is an experiment or an observational study.

1.95 Salt on Roads and Accidents Three situations are described at the start of this section, on page 31. In the third bullet, we describe an association between the amount of salt spread on the roads and the number of accidents. Describe a possible confounding variable and explain how it fits the definition of a confounding variable.

1.96 Height and Reading Ability In elementary school (grades 1 to 6), there is a strong association between a child’s height and the child’s reading ability. Taller children tend to be able to read at a higher level. However, there is a very significant confounding variable that is influencing both height and reading ability. What is it?

1.97 Golfing for a Long Life? A recent quote⁶⁰ on an online health site tells us “golfers’ average life expectancy was 5 years longer than other people’s.”

⁶⁰www.webmd.com, September 4, 2019.

- (a) What are the cases in this reported result? What is/are the variable(s)?
- (b) Does this information appear to come from an experiment or an observational study?
- (c) Give a confounding variable and explain how it meets the definition of a confounding variable.

1.98 How to Debate a Science Denier Science deniers oppose robust and valid results of scientific inquiry. A recent study⁶¹ investigates the most effective ways to debunk scientific misinformation. Science advocates can respond to misinformation using topic rebuttal (providing scientific facts on the subject) or technique rebuttal (explaining more generally the false techniques used by science deniers). In the study, 1773 participants first listened to a science denier and were then randomly assigned to one of four conditions: no rebuttal, topic rebuttal, technique rebuttal, or both topic and technique rebuttal. Participants' attitudes toward the science topic were measured and recorded three times: before participation, after the science denier, and after the rebuttal. Results indicate the importance of having a rebuttal: participants on average were influenced by the science denier, but the influence was mitigated by having any rebuttal. The study further showed that topic or technique rebuttal worked equally well, and having both provided no additional benefit.

- (a) What are the cases in this study?
- (b) What is the explanatory variable? Is it categorical or quantitative?
- (c) Is this an observational study or an experiment?
- (d) How many rows and how many columns will the dataset have, if we use cases as rows and variables as columns?

1.99 Music Volume and Beer Consumption A study⁶² was conducted measuring the impact that music volume has on beer consumption. The researchers went into bars, controlled the music volume, and measured how much beer was consumed. The article states that “the sound level of the environmental music was manipulated according to

⁶¹Schmid P and Betsch C, “Effective Strategies for Rebutting Science Denialism in Public Discussion,” *Nature Human Behaviour*, June 24, 2019.

⁶²Gueguen, N., Jacob, C., Le Guellec, H., Morineau, T., and Lourel, M., “Sound Level of Environmental Music and Drinking Behavior: A Field Experiment with Beer Drinkers,” *Alcoholism: Clinical and Experimental Research*, 2008; 32: 1795–1798.

a randomization scheme.” It was found that louder music corresponds to more beer consumption. Does this provide evidence that louder music causes people to drink more beer? Why or why not?

1.100 Nuts and Cholesterol Several studies have been performed to examine the relationship between nut consumption and cholesterol levels. Here we consider two such studies. In Study 1,⁶³ participants were assigned into two groups: one group was given nuts to eat each day, and the other group was told to consume a diet without nuts. In Study 2,⁶⁴ participants were free to follow their own diet, and reported how many nuts they consumed. Cholesterol levels were measured for all participants, and both studies found that nut consumption was associated with lower levels of LDL (“bad”) cholesterol. Based on the information above, which study do you think provides better evidence that nut consumption reduces LDL cholesterol? Explain your answer.

1.101 Antibiotics in Infancy and Obesity in Adults “Antibiotics in infancy may cause obesity in adults,” claims a recent headline.⁶⁵ A study in mice randomly assigned infant mice to either be given antibiotics or not, and the mice given antibiotics were more likely to be obese as adults. A separate study in humans found that children who had been given antibiotics before they were a year old (for example, for an ear infection) were more likely to be obese as adults. (Researchers believe the effect may be due to changes in the gut microbiome.) Based on these studies, is the headline an appropriate conclusion to make:

- (a) For mice?
- (b) For humans?

1.102 Fast-Food Diet and Depression A study⁶⁶ involving 84 middle-school students investigates the link between a diet high in processed foods and rates of depression. Researchers collected the urine of the participants and measured levels of sodium and potassium. A diet high in processed food, such

⁶³Morgan, W.A., and Clayshulte, B.J., “Pecans lower low density lipoprotein cholesterol in people with normal lipid levels.” *Journal of the American Dietetic Association*, 2000; 100(3), 312–318.

⁶⁴Li, T.Y., Brennan, A.M., Wedick, N.M., Mantzoros, C., Rifai, N., and Hu, F.B. “Regular consumption of nuts is associated with a lower risk of cardiovascular disease in women with type 2 diabetes.” *The Journal of Nutrition*, 2009; 139(7), 1333–1338.

⁶⁵Saey, T.H., “Antibiotics in infancy may cause obesity in adults,” *Science News*, September 20, 2014.

⁶⁶LaMotte S, “Fast food may contribute to teen depression, study says,” *CNN Health*, August 29, 2019.

as fast food, is high in sodium and low in potassium. The students were also evaluated on a 100-point scale for depressive symptoms. The study found that high levels of sodium and low levels of potassium were both associated with higher depression levels.

- Is this an experiment or an observational study?
- Can we conclude that eating a diet high in sodium and low in potassium increases the likelihood of having depression? Why or why not?
- Give a possible confounding variable in this study, and explain how it meets the definition of a confounding variable.

1.103 Potassium Levels in Spinach A study examining spinach leaves from a variety of different locations finds that spinach grown in soil with high amounts of iron tends to have lower levels of potassium.⁶⁷

- What are the two variables? Is each quantitative or categorical?
- Which is the explanatory variable?
- Does the result appear to come from an observational study or an experiment?
- Can we assume that higher levels of iron in the soil cause spinach to have less potassium?
- Describe anything that might be a possible confounding variable.

1.104 Encourage Kids to Read! Researchers⁶⁸ asked parents of 19 American children, ages 8 to 12, to complete surveys on the number of hours their children spent reading and on the number of hours they spent on screen-based media time. Each of the children also underwent magnetic resonance imaging to assess connectivity in the brain. More time spent reading was associated with higher connectivity, while more screen time was associated with lower connectivity in the brains of the children.

- How many cases are there in this study? How many variables are there?
- Identify each of the variables as quantitative or categorical. Also, identify each of the variables as an explanatory variable or a response variable.
- Is this an experiment or an observational study?

⁶⁷Shea K, *Health & Nutrition Update*, Tufts University, August 11, 2019.

⁶⁸Horowitz-Kraus T and Hutton J, “Brain connectivity in children is increased by the time they spend reading books and decreased by the length of exposure to screen-based media,” *Acta Paediatrica*, 107, 2018.

- Can we conclude that more time spent reading increases connectivity in children’s brains?
- Identify a possible confounding variable and explain how it meets the definition of a confounding variable.

1.105 Want to Be Healthier? Take a Vacation! A study⁶⁹ measured health characteristics (including triglyceride level, cholesterol, and blood glucose) for each of 63 adult participants and also asked participants how many vacations they had taken in the last year. Results showed that those who had taken more vacations had better results on the health measures.

- Is this an experiment or an observational study?
- Give a possible confounding variable and explain how it meets the definition of a confounding variable.
- Indicate whether each of the following is an appropriate conclusion (Yes or No) from this study:
 - People who are healthier tend to take more vacations.
 - People who take vacations tend to be healthier.
 - Taking vacations will improve a person’s health.
 - Improving one’s health will allow a person to take additional vacations.
- Is the title of this exercise an appropriate conclusion from this study?

1.106 Do Online Cat Videos Improve Mood? Exercise 1.69 on page 30 introduced a study on cat videos, in which people who clicked on the link were asked questions regarding their mood before and after the most recent time they watched a cat video. Overall, participants reported that after watching a cat video they had significantly more energy, fewer negative emotions, and more positive emotions. Can we conclude from this study that watching cat videos increases energy and improves emotional state?

1.107 Green Spaces Make Kids Smarter A recent article⁷⁰ claims that “Green Spaces Make Kids Smarter.” The study described in the article involved 2623 schoolchildren in Barcelona. The

⁶⁹Voelker R, “Vacationing More Often May Reduce Metabolic Syndrome Risk,” *JAMA*, August 14, 2019.

⁷⁰Khazan, O., “Green Spaces Make Kids Smarter,” *The Atlantic*, June 16, 2016.

researchers measured the amount of greenery around the children's schools, and then measured the children's working memories and attention spans. The children who had more vegetation around their schools did better on the memory and attention tests.

- What are the cases in this study?
- What is the explanatory variable?
- What is the response variable?
- Does the headline imply causation?
- Is the study an experiment or an observational study?
- Is it appropriate to conclude causation in this case?
- Suggest a possible confounding variable, and explain why it meets the requirements of a confounding variable.

1.108 Infections Can Lower IQ A headline in June 2015 proclaims "Infections can lower IQ."⁷¹ The headline is based on a study in which scientists gave an IQ test to Danish men at age 19. They also analyzed the hospital records of the men and found that 35% of them had been in a hospital with an infection such as an STI or a urinary tract infection. The average IQ score was lower for the men who had an infection than for the men who hadn't.

- What are the cases in this study?
- What is the explanatory variable? Is it categorical or quantitative?
- What is the response variable? Is it categorical or quantitative?
- Does the headline imply causation?
- Is the study an experiment or an observational study?
- Is it appropriate to conclude causation in this case?

1.109 Sitting Is the New Smoking A 2014 headline reads "Sitting Is the New Smoking: Ways a Sedentary Lifestyle is Killing You,"⁷² and explains the mounting evidence for ways in which sitting is bad for you. A more recent large 2015 study⁷³

⁷¹"Infections can lower IQ," *The Week*, June 12, 2015, p. 18.

⁷²"Sitting Is the New Smoking: Ways a Sedentary Lifestyle Is Killing You," http://www.huffingtonpost.com/the-active-times/sitting-is-the-new-smokin_b_5890006.html, September 29, 2014, Accessed July 17, 2015.

⁷³Patel, A.V., et al., "Leisure-time spent sitting and site-specific cancer incidence in a large US cohort," *Cancer Epidemiology, Biomarkers & Prevention*, June 30, 2015, doi:10.1158/1055-9965.EPI-15-0237.

contributed to this evidence by following 69,260 men and 77,462 women and finding that for women, those who spent more leisure time sitting were significantly more likely to get cancer.

- What are the explanatory and response variables for the 2015 study?
- Is the 2015 study an observational study or a randomized experiment?
- Can we conclude from the 2015 study that spending more leisure time sitting causes cancer in women? Why or why not?
- Can we conclude from the 2015 study that spending more leisure time sitting does not cause cancer in women?

1.110 Late Night Eating It is well-known that lack of sleep impairs concentration and alertness, and this might be due partly to late night food consumption. A study⁷⁴ took 44 people aged 21 to 50 and gave them unlimited access to food and drink during the day, but allowed them only 4 hours of sleep per night for three consecutive nights. On the fourth night, all participants again had to stay up until 4 am, but this time participants were randomized into two groups; one group was only given access to water from 10 pm until their bedtime at 4 am while the other group still had unlimited access to food and drink for all hours. The group forced to fast from 10 pm on performed significantly better on tests of reaction time and had fewer attention lapses than the group with access to late night food.

- What are the explanatory and response variables?
- Is this an observational study or a randomized experiment?
- Can we conclude that eating late at night worsens some of the typical effects of sleep deprivation (reaction time and attention lapses)?
- Are there likely to be confounding variables? Why or why not?

1.111 Does Early Language Reduce Tantrums? A recent headline reads "Early Language Skills Reduce Preschool Tantrums, Study Finds,"⁷⁵ and

⁷⁴University of Pennsylvania School of Medicine. "Eating less during late night hours may stave off some effects of sleep deprivation." *ScienceDaily*, June 4, 2015 www.sciencedaily.com/releases/2015/06/150604141905.htm.

⁷⁵"Early Language Skills Reduce Preschool Tantrums, Study Finds," *US News and World Report*, <http://health.usnews.com/health-news/news/articles/2012/12/20/early-language-skills-reduce-preschool-tantrums-study-finds>, 20 December 2012, Accessed July 17, 2015.

the article offers a potential explanation for this: “Verbalizing their frustrations may help little ones cope.” The article refers to a study that recorded the language skill level and the number of tantrums of a sample of preschoolers.

- (a) Is this an observational study or a randomized experiment?
- (b) Can we conclude that “Early Language Skills Reduce Preschool Tantrums”? Why or why not?
- (c) Give a potential confounding variable.

1.112 Does Buying Organic Food Improve Your Health? The National Health and Nutrition Examination Survey (NHANES)⁷⁶ surveys a large national random sample of Americans about a wide variety of health related topics. Among many questions, the 2009–2010 survey included a question about eating organic (“In the past 30 days, did you buy any food that had the word ‘organic’ on the package?”), and also a question about self-reported health status (“Would you say your health in general is Excellent, Very good, Good, Fair, or Poor?”). For the 5060 people who answered both questions, 46% of the people who had bought organic food reported their health as very good or excellent, as opposed to only 33% of the people who had not bought organic food. Remember that there are three possible explanations for this observed difference:

- (i) Causal association; buying (and then, presumably eating) organic food improves general health status.
- (ii) Association due to confounding; the groups (people who buy organic and those who don’t) differed to begin with.
- (iii) No actual association; just random chance.

Also recall that evaluating evidence *for* the causal explanation, (i), requires evaluating evidence *against* the alternative explanations, (ii) and (iii). In Chapter 6 we rule out explanation (iii). What about explanation (ii)?

- (a) Does this study provide evidence against confounding and groups differing to begin with? Why or why not?
- (b) If your answer to part (a) was no, name a potential confounding variable or a way in which the groups may differ to begin with.

⁷⁶<https://www.cdc.gov/nchs/nhanes/index.htm>

- (c) Based on your answer to part (a), does this study provide convincing evidence that buying organic food improves general health status? Why or why not?

1.113 Does Eating Organic Food Make Fruit Flies Live Longer? For a high school science project, a 16-year-old girl randomly divided fruit flies into two groups, and fed one group organic food and the other group conventional (non-organic) food.⁷⁷ The flies fed organic food lived an average of 3.25 days longer than the flies fed conventional food (which is a long time in the lifespan of a fruit fly!). Remember that there are three possible explanations for this observed difference:

- (i) Causal association; eating organic food causes fruit flies to live longer.
- (ii) Association due to confounding; the groups (organic fed and conventional fed fruit flies) differed to begin with.
- (iii) No actual association; just random chance.

Also recall that evaluating evidence *for* the causal explanation, (i), requires evaluating evidence *against* the alternative explanations, (ii) and (iii). In Chapter 5 we rule out explanation (iii). What about explanation (ii)?

- (a) Does this study provide evidence against confounding and groups differing to begin with? Why or why not?
- (b) If your answer to part (a) was no, name a potential confounding variable or a way in which the groups may differ to begin with.
- (c) Based on your answer to part (a), does this study provide convincing evidence that eating organic food causes fruit flies to live longer? Why or why not?

1.114 Goldilocks Effect: Read to Your Kids! (Part 2) Exercise 1.13 introduces a study in which 27 four-year-old children are presented with stories three different ways: audio only, illustrated, and animated. Every child was presented with all three formats while an MRI machine measured their brain connectivity. The researchers found a “Goldilocks effect,” in which both audio (too cold) and animation (too hot) showed low brain connectivity, while the illustrated format (similar to reading a book to a child) showed the highest connectivity (just right!).

⁷⁷Chhabra R, Kolli S, Bauer JH (2013) Organically Grown Food Provides Health Benefits to *Drosophila melanogaster*. PLoS ONE 8(1): e52988.

- (a) Is this an observational study, a randomized controlled experiment or a matched pairs type experiment?
- (b) Can we conclude that reading to a child increases connectivity in their brain more than listening to a story or watching a movie does?
- (c) How should randomization be used in the design of this study?

1.115 Sleep and Recognition of Facial Expressions

The ability to recognize and interpret facial expressions is key to successful human interaction. Could this ability be compromised by sleep deprivation? A study⁷⁸ took 18 healthy young adult volunteers and exposed them to 70 images of facial expressions, ranging from friendly to threatening. They were each shown images both after a full night of sleep and after sleep deprivation (24 hours of being awake), and whether each individual got a full night of sleep or was kept awake first was randomly determined. The study found that people were much worse at recognizing facial expressions after they had been kept awake.

- (a) What are the explanatory and response variables?
- (b) Is this an observational study or a randomized experiment? If it is a randomized experiment, is it a randomized comparative experiment or a matched pairs experiment?
- (c) Can we conclude that missing a night of sleep hinders the ability to recognize facial expressions? Why or why not?
- (d) In addition, for the people who had slept, the study found a strong positive association between quality of Rapid Eye Movement (REM) sleep and ability to recognize facial expressions. Can we conclude that better quality of REM sleep improves ability to recognize facial expressions? Why or why not? (*Hint*: What is the explanatory variable in this case? Was it randomly assigned?)

1.116 Diet Cola and Weight Gain in Humans A study⁷⁹ found that American senior citizens who

⁷⁸Goldstein-Piekarski, A., et al., "Sleep Deprivation Impairs the Human Central and Peripheral Nervous System Discrimination of Social Threat," *The Journal of Neuroscience*, July 15, 2015; 35(28): 10135–10145; doi: 10.1523/JNEUROSCI.5254-14.2015

⁷⁹Fowler, S.P., Williams, K., and Hazuda, H.P. "Diet Soda Intake Is Associated with Long-Term Increases in Waist Circumference in a Bioethnic Cohort of Older Adults: The San Antonio Longitudinal Study of Aging," *Journal of the American Geriatrics Society*, 2015; 63(4), 708–715.

report drinking diet soda regularly experience a greater increase in weight and waist circumference than those who do not drink diet soda regularly.

- (a) From these results, can we conclude that drinking diet soda causes weight gain? Explain why or why not.
- (b) Consider the results of this study on senior citizens, and the randomized experiment on rats introduced in Exercise 1.70 on page 30, which showed a similar association. Discuss what these two studies together might imply about the likelihood that diet cola causes weight gain in humans.

1.117 Urban Brains and Rural Brains A study published in 2010 showed that city dwellers have a 21% higher risk of developing anxiety disorders and a 39% higher risk of developing mood disorders than those who live in the country. A follow-up study published in 2011 used brain scans of city dwellers and country dwellers as they took a difficult math test.⁸⁰ To increase the stress of the participants, those conducting the study tried to humiliate the participants by telling them how poorly they were doing on the test. The brain scans showed very different levels of activity in stress centers of the brain, with the urban dwellers having greater brain activity than rural dwellers in areas that react to stress.

- (a) Is the 2010 study an experiment or an observational study?
- (b) Can we conclude from the 2010 study that living in a city increases a person's likelihood of developing an anxiety disorder or mood disorder?
- (c) Is the 2011 study an experiment or an observational study?
- (d) In the 2011 study, what is the explanatory variable and what is the response variable? Indicate whether each is categorical or quantitative.
- (e) Can we conclude from the 2011 study that living in a city increases activity in stress centers of the brain when a person is under stress?

1.118 Split the Bill? When the time comes for a group of people eating together at a restaurant to pay their bill, sometimes they might agree to split the costs equally and other times will pay individually. If this decision were made in advance, would it affect what they order? Suppose that you'd like to do an experiment to address this question. The variables you will record are the type of *payment*

⁸⁰"A New York state of mind," *The Economist*, June 25, 2011, p. 94.

(split or individual), *sex* of each person, number of *items* ordered, and the *cost* of each person's order. Identify which of these variables should be treated as *explanatory* and which as *response*. For each explanatory variable, indicate whether or not it should be randomly assigned.

1.119 Be Sure to Get Your Beauty Sleep! New research⁸¹ supports the idea that people who get a good night's sleep look more attractive. In the study, 23 subjects ages 18 to 31 were photographed twice, once after a good night's sleep and once after being kept awake for 31 hours. Hair, make-up, clothing, and lighting were the same for both photographs. Observers then rated the photographs for attractiveness, and the average rating under the two conditions was compared. The researchers report in the *British Medical Journal* that "Our findings show that sleep-deprived people appear less attractive compared with when they are well rested."

- What is the explanatory variable? What is the response variable?
- Is this an experiment or an observational study? If it is an experiment, is it a randomized comparative design or a matched pairs design?
- Can we conclude that sleep deprivation *causes* people to look less attractive? Why or why not?

1.120 Do Antidepressants Work? Following the steps given, design a randomized comparative experiment to test whether fluoxetine (the active ingredient in Prozac pills) is effective at reducing depression. The participants are 50 people suffering from depression and the response variable is the change on a standard questionnaire measuring level of depression.

- Describe how randomization will be used in the design.
- Describe how a placebo will be used.
- Describe how to make the experiment double-blind.

⁸¹Stein, R., "Beauty sleep no myth, study finds," *Washington Post*, washingtonpost.com, Accessed December 15, 2010.

1.121 Do Children Need Sleep to Grow? About 60% of a child's growth hormone is secreted during sleep, so it is believed that a lack of sleep in children might stunt growth.⁸²

- What is the explanatory variable and what is the response variable in this association?
- Describe a randomized comparative experiment to test this association.
- Explain why it is difficult (and unethical) to get objective verification of this possible causal relationship.

1.122 Carbo Loading It is commonly accepted that athletes should "carbo load," that is, eat lots of carbohydrates, the day before an event requiring physical endurance. Is there any truth to this? Suppose you want to design an experiment to find out for yourself: "Does carbo loading actually improve athletic performance the following day?" You recruit 50 athletes to participate in your study.

- How would you design a randomized comparative experiment?
- How would you design a matched pairs experiment?
- Which design do you think is better for this situation? Why?

1.123 Alcohol and Reaction Time Does alcohol increase reaction time? Design a randomized experiment to address this question using the method described in each case. Assume the participants are 40 college seniors and the response variable is time to react to an image on a screen after drinking either alcohol or water. Be sure to explain how randomization is used in each case.

- A randomized comparative experiment with two groups getting two separate treatments
- A matched pairs experiment

1.124 Causation and Confounding Causation does not necessarily mean that there is no confounding variable. Give an example of an association between two variables that have a causal relationship AND have a confounding variable.

⁸²Rochman, B., "Please, Please, Go to Sleep," *Time Magazine*, March 26, 2012, p. 46.