Thinkstock

CHAPTER 3

Producing Data

Introduction

In Chapters 1 and 2, we learned some basic tools of *data analysis*. We used graphs and numbers to describe data. When we do exploratory data analysis, we rely heavily on plotting the data. We look for patterns that suggest interesting conclusions or questions for further study. However, *exploratory analysis alone can rarely provide convincing evidence for its conclusions because striking patterns that we find in data can arise from many sources.*

CAUTION

exploratory data analysis

The validity of the conclusions that we draw from an analysis of data depends not only on the use of the best methods to perform the analysis, but also on the quality of the data. Therefore, Section 3.1 begins this chapter with a short overview on sources of data.

The two main sources for quality data are designed experiments and sample surveys. We study these two sources in Sections 3.2 and 3.3, respectively.

Should an experiment or sample survey that could possibly provide interesting and important information always be performed? How can we safeguard the privacy of subjects in a sample survey? What constitutes the mistreatment of people or animals who are studied in an experiment? These are questions of **ethics**. In Section 3.4, we address ethical issues related to the design of studies and the analysis of data.

ethics

3.1 Sources of Data

When you complete this section, you will be able to:

- Identify anecdotal data and, using specific examples, explain why they have limited value.
- Identify available data and explain how they can be used in specific examples.
- Identify data collected from sample surveys and explain how they can be used in specific examples.
- Identify data collected from experiments and explain how they can be used in specific examples.
- Distinguish data that are from experiments, from observational studies that are sample surveys, and from observational studies that are not sample surveys.
- Identify the treatment in an experiment.

There are many sources of data. Some data are very easy to collect, but they may not be very useful. Other data require careful planning and need professional staff to gather. These can be much more useful. Whatever the source, a good statistical analysis will start with a careful study of the source of the data. Here is one type of source.

Anecdotal data

It is tempting to simply draw conclusions from our own experience, making no use of more broadly representative data. A magazine article about Pilates says that men need this form of exercise even more than women do. The article describes the benefits that two men received from taking Pilates classes. A newspaper ad states that a particular brand of window is "considered to be the best" and says that "now is the best time to replace your windows and doors." These types of stories, or *anecdotes*, sometimes provide data. However, this type of data does not give us a sound basis for drawing conclusions.

ANECDOTAL DATA

Anecdotal data represent individual cases, which often come to our attention because they are striking in some way. These cases are not necessarily representative of any larger group of cases.

USE YOUR KNOWLEDGE

3.1 The best instructor? A friend tells you that the instructor in her statistics class is the best teacher in the college. Can you conclude that this teacher is better than all of the other instructors in the college? Explain your answer.

3.2 Describe an anecdote. Find an example from some recent experience where anecdotal evidence was used to draw a conclusion that is not justified. Describe the example and explain why the anecdote should not be used in this way.

3.3 Opposition to a new requirement. Your student newspaper ran a story describing interviews with three students who were strongly opposed to a proposed new requirement that all students take a course on ethics. Can you conclude that most students are opposed to this requirement? Explain your answer.

3.4 Are all vehicles this good? A friend has driven a Toyota Camry for more than 200,000 miles and with only the usual service maintenance expenses. Explain why not all Camry owners can expect this kind of performance.

Not all anecdotal data are bad. The experiences of an individual or a small group of individuals might suggest an interesting study that could be performed using more carefully collected data.

Available data

Occasionally, data are collected for a particular purpose but can also serve as the basis for drawing sound conclusions about other research questions. We use the term *available data* for this type of data.

AVAILABLE DATA

Available data are data that were produced for some other purpose but that may help answer a question of interest.

The library and the Internet can be good sources of available data. Because producing new data is expensive, we all use available data whenever possible. Here are two examples.

EXAMPLE 3.1

How Americans use their time. If you visit the U.S. Bureau of Labor Statistics website, **bls.gov**, you will find many interesting sets of data and statistical summaries. The American Time Use Survey[1] recently reported that men spend an average of 5.71 hours per day on leisure and sports activities, while women spend an average of 4.93 hours on these activities.

EXAMPLE 3.2

Math skills. At the website of the National Center for Education Statistics, **nces.ed.gov**, you will find full details about the math skills of schoolchildren as determined by the latest National Assessment of Educational Progress (Figure 3.1). Mathematics scores have slowly but steadily increased since 1990. Across all racial/ethnic groups, both boys and girls in most states are getting better in math.

FIGURE 3.1 Websites of government statistical offices are prime sources of data. Here is a page from the website of the National Center for Education Statistics. (Source: U.S. Department of Education Institute of Education Sciences National Center for Education Statistics)

Many nations have a single national statistical office, such as Statistics Canada (**statcan.gc.ca**) and Mexico's INEGI (**inegi.org.mx**). More than 70 different U.S. agencies collect data. You can reach most of them through the U.S. government's FedStats site (**fedstats.sites.usa.gov**).

USE YOUR KNOWLEDGE

3.5 What more do you need? A website claims that Millennial generation consumers are very loyal to the brands that they prefer. What additional information do you need to evaluate this claim?

A survey of college athletes is designed to estimate the percent who gamble. Do restaurant patrons give higher tips when their server repeats their order carefully? The validity of our conclusions from the analysis of data collected to address these issues rests on a foundation of carefully collected data.

In this chapter, we will develop the skills needed to produce trustworthy data and to judge the quality of data produced by others. The techniques for producing data that we will study require no formulas, but they are among the most important ideas in statistics. Statistical designs for producing data rely on either *sampling* or *experiments*.

Sample surveys and experiments

sample surveys

How have the attitudes of Americans, on issues ranging from abortion to work, changed over time? **Sample surveys** are the usual tool for answering questions like these.

EXAMPLE 3.3

The General Social Survey. One of the most important sample surveys is the General Social Survey (GSS) conducted by the National Opinion Research Center (NORC), an organization affiliated with the University of Chicago.[2] The GSS interviews about 3000 adult residents of the United States every other year.

sample population

The GSS selects a **sample** of adults to represent the larger **population** of all English-speaking adults living in the United States. The idea of *sampling* is to study a part in order to gain information about the whole. Data are often produced by sampling a population of people or things. Opinion polls, for example, may report the views of the entire country based on interviews with a sample of about 1000 people. Government reports on employment and unemployment are produced from a monthly sample of about 60,000 households. The quality of manufactured items is monitored by inspecting small samples each hour or each shift.

USE YOUR KNOWLEDGE

3.6 **Check out the General Social Survey.** Visit the General Social Survey website at **gss.norc.org**. Write a short summary of one of their reports, paying particular attention to the methods used to collect the data.

census

In all our examples, the expense of examining every item in the population makes sampling a practical necessity. Timeliness is another reason for preferring a sample to a **census**, which is an attempt to contact every individual in the population. We want information on current unemployment and public opinion next week, not next year. Moreover, a carefully conducted sample is often more accurate than a census. Accountants, for example, sample a firm's inventory to verify the accuracy of the records. Attempting to count every last item in the warehouse would be not only expensive, but also inaccurate. Bored people do not count carefully.

If conclusions based on a sample are to be valid for the population, a sound design for selecting the sample is required. Sampling designs are the topic of Section 3.3.

A sample survey collects information about a population by selecting and measuring a sample from the population. The goal is a picture of the population, disturbed as little as possible by the act of gathering information. Sample surveys are one kind of *observational study*.

OBSERVATION VERSUS EXPERIMENT

In an **observational study**, we observe individuals and measure variables of interest but do not attempt to influence the responses.

In an **experiment**, we deliberately impose some condition on individuals and we observe their responses.

EXAMPLE 3.4

Baseball players have strong bones in their throwing arms. A study of young baseball players measured the strength of the bones in their throwing arms. A control group of subjects who were matched with the baseball players based on age were also measured. This is an example of an observational study that is not a sample survey. The study reported that bone strength was 30% higher in the baseball players.[3]

What can we conclude from this study? If you start to play baseball, will you have stronger bones in your throwing arm?

EXAMPLE 3.5

Is there a cause-and-effect relationship? Example 3.4 describes an observational study. People choose to participate in baseball or not. Is it possible that those who choose to play baseball have stronger arms than those who do not? The study does not address this question.

We can imagine an experiment that would remove these difficulties. From a large group of subjects, require some to play baseball and forbid the rest from playing. This is an experiment because the condition (playing baseball or not) is imposed on the subjects. Of course, this particular experiment is neither practical nor ethical.

EXAMPLE 3.6

Baseball and bones. Example 3.4 compared the arm bone strengths of baseball players with those of age-matched controls. Although the study tells us something about baseball players, the results are particularly interesting because they suggest that certain kinds of exercise can help us to build strong bones.

USE YOUR KNOWLEDGE

3.7 Available data. Can available data be from an observational study? Can available data be from an experiment? Explain your answers.

3.8 Picky eaters. A study of 2049 children in grades 4 to 6 in 33 schools recorded their behaviors in the lunchroom. One of the conclusions of the study was that girls discarded more food than boys.[4] Is this an observational study or an experiment? Is it a sample survey? If it is an experiment, what is the treatment? Explain your answers.

3.9 Automatic soap dispensers. A study compared several brands of automatic soap dispensers. For one test, the dispensers were run until their AA batteries failed. The times to failure were compared for the different brands.[5] Is this an observational study or an experiment? Is it a sample survey? If it is an experiment, what is the condition? Explain your answers.

An observational study, even one based on a carefully chosen sample, is a poor way to determine what will happen if we change something. The best way to see the effects of a change is to do an **intervention**—where we actually impose the change. When our goal is to understand cause and effect, experiments are the only source of fully convincing data.

intervention

LOOK BACK
confounding, p. 150
explanatory variable, p. 82

Confounding occurs when an explanatory variable is related to one or more other variables that have an influence on the response variable. When this happens, we sometimes attribute a relationship to an explanatory when the effect is fully or partly due to the confounding variables.

In Example 3.4, the effect of baseball playing on arm bone strength is confounded with (mixed up with) other characteristics of the subjects in the study. Observational studies that examine the effect of a single variable on an outcome can be misleading when the effects of the explanatory variable are confounded with those of other variables.

Because experiments allow us to isolate the effects of specific variables, we generally prefer them. Here is an example.

EXAMPLE 3.7

Which web page design sells more? A company that sells products on the Internet wants to decide which of two possible web page designs to use. During a two-week period, they will use both designs and collect data on sales. They randomly select one of the designs to be used on the first day and then alternate the two designs on each of the following days. At the end of this period, they compare the sales for the two designs.

Experiments usually require some sort of randomization, as in this example. We begin the discussion of statistical designs for data collection in Section 3.2 with the principles underlying the design of experiments.

USE YOUR KNOWLEDGE

3.10 Software for teaching creative writing. An educational software company wants to compare the effectiveness of its computer animation for teaching creative writing with that of a textbook presentation. The company tests the creative-writing skills of a number of second-year college students and then randomly divides them into two groups. One group uses the animation, and the other studies the text. The company retests all the students and compares the increase in creative-writing skills in the two groups. Is this an experiment? Why or why not? What are the explanatory and response variables?

LOOK BACK
response variable, p. 82

3.11 Apples or apple juice? Food rheologists study different forms of foods and how the form of a food affects how full we feel when we eat it. One study prepared samples of apple juice and samples of apples with the same number of calories. Half of the subjects were fed apples on one day followed by apple juice on a later day; the other half received the apple juice followed by the apples. After eating, the subjects were asked about how full they felt. Is this an experiment? Why or why not? What are the explanatory and response variables?

SECTION 3.1 SUMMARY

- **Anecdotal data** come from stories or reports about cases that do not necessarily represent a larger group of cases.
- **Available data** are data that were produced for some other purpose but that may help answer a question of interest.
- A **sample survey** collects data from a **sample** of cases that represent some larger **population** of cases.
- A **census** collects data from all cases in the population of interest.
- In an **experiment**, a **treatment** (an **intervention)** is imposed and the responses are recorded.
- **Confounding** occurs when the effects of two or more variables are related in such a way that we need to take care in assigning the effect on the response variable to one or to the other.

SECTION 3.1 EXERCISES

For Exercises 3.1 to 3.4, see pages 164–165; for Exercise 3.5, see page 166; for Exercise 3.6, see page 167; for Exercises 3.7 and 3.9, see pages 168–169; and for Exercises 3.10 and 3.11, see pages 169–170.

In several of the following exercises, you are asked to identify the type of data that is described. Possible answers include anecdotal data, available data, observational data that are from sample surveys, observational data that are not from sample surveys, and experimental data. It is possible for some data to be classified in more than one category.

3.12 Not enough tuna. You like to eat tuna fish sandwiches. Recently you have noticed that there does not seem to be as much tuna as you expect when you open the can. Identify the type of data that this represents, and describe how it can or cannot be used to reach a conclusion about the amount of tuna in cans of tuna fish. Is this an experiment? If yes, what is the treatment?

3.13 More about tuna. According to a story in *Consumer Reports*, three major producers of canned tuna agreed to pay \$3,300,000 to settle claims in California that the amount of tuna in their cans was less than the amount printed on the label of the cans.[6] What kind of data do you think was used in this situation to convince the producers to pay this amount of money to settle the claims? Explain your answer fully.

3.14 What's wrong? Explain what is wrong in each of the following statements.

(a) Available data is always anecdotal.

(b) A census collects information on a subset of the population of interest.

(c) A sample survey usually involves a treatment.

3.15 Satisfaction with allocation of concert tickets. Your college sponsored a concert that sold out.

(a) After the concert, an article in the student newspaper reported interviews with three students who were unable to get tickets and were very upset. What kind of data does this represent? Explain your answer.

(b) A week later the student organization that sponsored the concert set up a website where students could rank their satisfaction with the way that the tickets were allocated using a 5-point scale with values "very satisfied," "satisfied," "neither satisfied nor unsatisfied," "dissatisfied," and "very dissatisfied." The website was open to any students who chose to provide their opinion. How would you classify these data? Give reasons for your answer.

(c) Suppose that the website in part (b) was changed so that only a sample of students from the college were invited by a text message to respond, and those who did not respond within three days were sent an additional text message reminding them to respond. How would your answer to part (b) change, if at all?

(d) Is the description in part (c) an experiment? If yes, what is the treatment?

(e) Write a short summary contrasting different types of data using your answers to parts (a), (b), (c), and (d) of this exercise.

3.16 Does echinacea reduce the severity of the common cold? In a study designed to evaluate the benefits of taking echinacea when you have a cold, 719 patients were randomly divided into four groups. The groups were (1) no pills, (2) pills that had no echinacea, (3) pills that had echinacea but the subjects did not know whether or not the pills contained echinacea, and (4) pills that had echinacea and the bottle containing the pills stated that the contents included echinacea. The outcome was a measure of the severity of the cold.[7]

(a) Identify the type of data collected in this study. Give reasons for your answer.

(b) Is this an experiment? If yes, what is the treatment?

3.2 Design of Experiments

When you complete this section, you will be able to:

- Identify experimental units, subjects, treatments, and outcomes for an experiment.
- Identify a comparative experiment.
- Describe a placebo effect in an experiment.
- Identify bias in an experiment.
- Explain the need for a control group in an experiment.
- Explain the need for randomization in an experiment.
- When evaluating an experiment, apply the basic principles of experimental design: compare, randomize, and repeat.
- Use a table of random digits to randomly assign experimental units to treatments in an experiment.
- Use software to randomly assign experimental units to treatments in an experiment.
- Identify a matched pairs design.
- Identify a block design.

An experiment is a study in which we actually do something to people, animals, or objects in order to observe the response. Here is the basic vocabulary of experiments.

EXPERIMENTAL UNITS, SUBJECTS, TREATMENTS, AND OUTCOMES

The individuals on which the experiment is done are the **experimental units**. When the units are human beings, they are called **subjects**. Experimental conditions applied to the units are called **treatments**. The **outcomes** are the measured variables that are used to compare the treatments.

Because the purpose of an experiment is to reveal the response of one variable to changes in one or more other variables, the distinction between explanatory and response variables is important. The explanatory variables factors

level of a factor

in an experiment are often called **factors**. Many experiments study the joint effects of several factors. In such an experiment, each treatment is formed by combining a specific value (often called a **level**) of each of the factors.

EXAMPLE 3.8

Are smaller class sizes better? Do smaller classes in elementary school really benefit students in areas such as scores on standard tests, staying in school, and going on to college? We might do an observational study that compares students who happened to be in smaller classes with those who happened to be in larger classes in their early school years. Small classes are expensive, so they are more common in schools that serve richer communities. Students in small classes tend to also have other advantages: their schools have more resources, their parents are better educated, and so on. Confounding makes it impossible to isolate the effects of small classes.

The Tennessee STAR program was an experiment on the effects of class size. It has been called "one of the most important educational investigations ever carried out." The *subjects* were 6385 students who were beginning kindergarten. Each student was assigned to one of three *treatments*: regular class (22 to 25 students) with one teacher, regular class (22 to 25 students) with a teacher and a full-time teacher's aide, and small class (13 to 17 students). These treatments are levels of a single *factor*, the type of class. The students stayed in the same type of class for four years, then all returned to regular classes. In later years, students from the small classes had higher scores on the *outcomes*, standard tests. The benefits of small classes were greatest for minority students.[8]

LOOK BACK

lurking variables, p. 130

Example 3.8 illustrates the big advantage of experiments over observational studies. **In principle, experiments can give good evidence for causation.** In an experiment, we study the specific factors we are interested in while controlling the effects of **lurking variables**. All the students in the Tennessee STAR program followed the usual curriculum at their schools. Because students were assigned to different class types within their schools, school resources and family backgrounds were not confounded with class type. The only systematic difference was the type of class. When students from the small classes did better than those in the other two types, we can be confident that class size made the difference.

EXAMPLE 3.9

Alamy

Repeated exposure to advertising. What are the effects of repeated exposure to an advertising message? The answer may depend both on the length of the ad and on how often it is repeated. An experiment investigated this question using undergraduate students as *subjects*. All subjects viewed a 40-minute television program that included ads for a digital camera. Some subjects saw a 30-second commercial; others, a 90-second version. The same commercial was shown either one, three, or five times during the program.

This experiment has two *factors*: length of the commercial, with two levels, and repetitions, with three levels. The six combinations of one level of each factor form six *treatments*. Figure 3.2 shows the layout of the treatments. After viewing the TV program, all the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it. These are the *outcomes*.

		Factor B Repetitions		
		1 time	3 times	5 times
Factor A Length	30 seconds	1	2	3
	90 seconds	4	5	6

FIGURE 3.2 The treatments in the study of advertising, Example 3.9. Combining the levels of the two factors forms six treatments.

Example 3.9 shows how experiments allow us to study the combined effects of more than one factor. The interaction of several factors can produce effects that cannot be predicted from looking at the effects of each factor alone. Perhaps longer commercials increase interest in a product, and more commercials also increase interest, but if we both make a commercial longer and show it more often, viewers get annoyed and their interest in the product drops. The two-factor experiment in Example 3.9 will help us find out.

USE YOUR KNOWLEDGE

3.17 Calcium and bones. Calcium is important for the growth of bone for children. In a study designed to understand how calcium is processed by the body, 40 young girls attended a summer camp where they were fed a controlled diet. The camp ran for two 3-week periods. For one period, the diet included a low amount of calcium. For the other period, there was a high amount of calcium in the diet. The researchers recorded the amount of calcium retained in the body for each girl. Explain why this study is an experiment and identify the experimental units, the treatments, and the response variable. Describe the factor and its levels.

3.18 Does echinacea reduce the severity of the common cold? In a study designed to evaluate the benefits of taking echinacea when you have a cold, 719 patients were randomly divided into four groups. The groups were (1) no pills, (2) pills that had no echinacea, (3) pills that had echinacea but the subjects did not know whether or not the pills contained echinacea, and (4) pills that had echinacea and the bottle containing the pills stated that the contents included echinacea. The outcome was a measure of the severity of the cold.[9] Identify the experimental units, the treatments, and the outcome. Describe the factor and its levels. The study subjects were aged 12 to 80 years. To what extent do you think the results of this experiment can be generalized to young children?

Comparative experiments

Laboratory experiments in science and engineering often have a simple design with only a single treatment, which is applied to all experimental units. The design of such an experiment can be outlined as

$$\textbf{Treatment} \longrightarrow \textbf{Observe response}$$

For example, we may subject a beam to a load (treatment) and measure its deflection (observation). We rely on the controlled environment of the laboratory to protect us from lurking variables. When experiments are conducted outside the laboratory or with living subjects, such simple designs often yield invalid data. That is, we cannot tell whether the response was due to the treatment or to lurking variables.

EXAMPLE 3.10

Will writing about it reduce test anxiety? A study designed to reduce test anxiety had students write an essay about their feelings concerning an upcoming exam.[10] The scores on this exam, the second of the semester, were compared with those on the first exam in the course. The mean scores on the second exam were higher than the mean scores on the first exam.

Write about feelings ⟶ Observe exam scores

The test anxiety experiment of Example 3.10 was poorly designed to evaluate the effect of the writing exercise. Perhaps exam scores would have increased on the second exam because the students became more familiar with the exam style of this particular instructor even without the writing exercise. Another possible explanation is that people typically respond to the personal attention that the students received by the person who explained how to write about their feelings regarding the exam.

placebo effect

In medical settings, this phenomenon is called the **placebo effect**. In medicine, a placebo is a dummy treatment, such as a sugar pill. People respond favorably to personal attention or to any treatment that they hope will help them. On the other hand, the writing exercise may have been very effective in improving exam scores. For this experiment, we don't know whether the change was due to writing the essay, to the personal contacts with the study personnel, or to greater familiarity with the way the instructor designed exams.

comparative experiment

The test anxiety experiment gave inconclusive results because the effect of writing the essay was confounded with other factors that could have had an effect on exam scores. The best way to avoid confounding is to do a **comparative experiment**. Think about a study in which some students performed the writing exercise and others did not. A comparison of the exam scores of these two groups of students would provide an evaluation of the effect of the writing exercise.

control group
treatment group

In medical settings, it is standard practice to randomly assign patients either to a **control group** or a **treatment group**. All patients are treated the same in every way except that the treatment group receives the product that is being evaluated.

CAUTION

Uncontrolled experiments (that is, experiments that don't include a control group) in medicine and the behavioral sciences can be dominated by such influences as the details of the experimental arrangement, the selection of subjects, and the placebo effect. The result is often *bias*.

BIAS

The design of a study is **biased** if it systematically favors certain outcomes.

An uncontrolled study of a new medical therapy, for example, is biased in favor of finding the treatment effective because of the placebo effect. Uncontrolled studies in medicine give new therapies a much higher success rate than proper comparative experiments do. Well-designed experiments usually compare several treatments.

USE YOUR KNOWLEDGE

3.19 Does aspirin cure headaches? A study enrolled 100 college students who had frequent headaches to participate in a study to examine the effects of aspirin on their headaches. The students were instructed to take aspirin when they had a headache and to report whether there was a substantial relief from the headache pain within an hour.

(a) Explain why this study is biased.

(b) How would you change the study to remove the bias? Explain your answer.

3.20 Are the teacher evaluations biased? The evaluations of two instructors by their students are compared when it is time to decide raises for the coming year. One teacher always hands out the evaluation forms in class when the grades on the first exam are given to the students. The other instructor always hands out the evaluation forms at the end of a class in which a very interesting film clip is shown. Discuss the possibility of bias in this context.

Randomization

experimental design

The **design of an experiment** first describes the response variable or variables, the factors (explanatory variables), and the treatments, with comparison as the leading principle. Figure 3.2 (page 173) illustrates this aspect of the design of a study of response to advertising. The second aspect of experimental design is how the experimental units are assigned to the treatments. Comparison of the effects of several treatments is valid only when all treatments are applied to similar groups of experimental units. If one corn variety is planted on more fertile ground or if one cancer drug is given to more seriously ill patients, comparisons among treatments are meaningless. If groups assigned to treatments are quite different in a comparative experiment, we should be concerned that our experiment will be biased. How can we assign experimental units to treatments in a way that is fair to all treatments?

matching

Experimenters often attempt to match groups by elaborate balancing acts. Medical researchers, for example, try to match the patients in a "new drug" experimental group and a "standard drug" control group by age, sex, physical condition, smoker or not, and so on. **Matching** is helpful but not adequate—there are too many lurking variables that might affect the outcome. The experimenter is unable to measure some of these variables and will not think of others until after the experiment.

Some important variables, such as how advanced a cancer patient's disease is, are so subjective that they can't be measured. In other cases, an experimenter might unconsciously bias a study by assigning those patients who seemed the sickest to a promising new treatment in the (unconscious) hope that it would help them.

The statistician's remedy is to rely on chance to make an assignment that does not depend on any characteristic of the experimental units and that does not rely on the judgment of the experimenter in any way. The use of chance can be combined with matching, but the simplest experimental design creates groups by chance alone. Here is an example.

EXAMPLE 3.11

Xaume Olleros/Bloomberg via Getty Images

Which smartphone should be marketed? Two teams have each prepared a prototype for a new smartphone. Before deciding which one will be marketed, the smartphones will be evaluated by college students. Forty students will receive a new phone. They will use it for two weeks and then answer some questions about how well they like the phone. The 40 students will be randomized, with 20 receiving each phone.

This experiment has a single factor (prototype) with two levels. The researchers must divide the 40 student subjects into two groups of 20. To do this in a completely unbiased fashion, put the names of the 40 students in a hat, mix them up, and draw 20. These students will receive Phone 1, and the remaining 20 will receive Phone 2. Figure 3.3 outlines the design of this experiment.

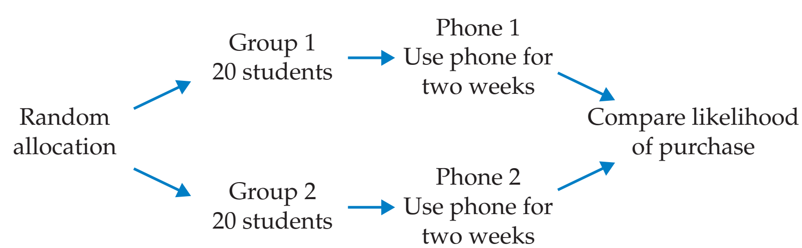

FIGURE 3.3 Outline of a randomized comparative experiment, Example 3.11.

randomization

The use of chance to divide experimental units into groups is called **randomization**. The design in Figure 3.3 combines comparison and randomization to arrive at the simplest randomized comparative design. This "flowchart" outline presents all the essentials: randomization, the sizes of the groups and which treatment they receive, and the response variable. There are, as we will see later, statistical reasons for using treatment groups that are about equal in size.

USE YOUR KNOWLEDGE

3.21 Diagram the echinacea experiment. Refer to Exercise 3.16 (page 171). Draw a diagram similar to Figure 3.3 that describes the experiment.

3.22 Diagram the aspirin experiment. Draw a diagram similar to Figure 3.3 that describes the experiment you suggested in part (b) of Exercise 3.19 (page 175).

Randomized comparative experiments

The logic behind the randomized comparative design in Figure 3.3 is as follows:

- Randomization produces two groups of subjects that we expect to be similar in all respects before the treatments are applied.
- Comparative design helps ensure that influences other than the characteristics of the smartphone operate equally on both groups.
- Therefore, differences in the satisfaction with the smartphone must be due either to the characteristics of the phone or to the chance assignment of subjects to the two groups.

That "either-or" deserves more comment. We cannot say that *all* the difference in the satisfaction with the two smartphones is caused by the characteristics of the phones. There would be some difference even if both groups used the same phone. Some students would be more likely to be highly favorable of any new phone. Chance can assign more of these students to one of the phones so that there is a chance difference between the groups. We would not trust an experiment with just one subject in each group, for example. The results would depend too much on which phone got lucky and received the subject who was more likely to be highly satisfied. If we assign many students to each group, however, the effects of chance will average out. There will be little difference in the satisfaction between the two groups unless the phone characteristics causes a difference. "Use enough subjects to reduce chance variation" is the third big idea of statistical design of experiments.

PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles of statistical design of experiments are

1. **Compare** two or more treatments. This will control the effects of lurking variables on the response.
2. **Randomize**—use chance to assign experimental units to treatments.
3. **Repeat** each treatment on many units to reduce chance variation in the results.

How to randomize

The idea of randomization is to assign subjects to treatments by drawing names from a hat. In practice, experimenters use software to carry out randomization. For example, most statistical software can choose five out of a list of 10 at random. The list might contain the names of 10 human subjects to be randomly assigned to two groups. The five chosen form one group, and the five that remain form the second group. The *Simple Random Sample* applet on the text website makes it particularly easy to choose treatment groups at random.

label

When we randomize, we first give a **label** to each in the collection of items to be randomized. The label could be the name of a subject in a clinical study or simply a numerical identification number. We then perform the randomization using software or a table of random numbers. To illustrate these methods, let's randomize 10 subjects for a study that will compare a treatment with a placebo control. We will randomly select the five subjects for the treatment group, and the remaining subjects will receive the placebo. We start by labeling the subjects with the numbers 1 through 10.

Randomization using software

LOOK BACK uniform distribution, p. 71

Here is an example of one way to do the randomization using Excel. We start with a spreadsheet that has 10 rows corresponding to 10 subjects to be randomized to treatment or placebo.

The basic idea is that we generate a uniform random variable for each subject. In Excel, we use the RAND() function for this step. Then we sort the spreadsheet by the column with the uniform numbers and select the first five labels to be the treatment group and the remaining labels to be the placebo controls.

This process is essentially the same as writing the labels on a deck of 10 cards. We then shuffle the cards and deal five cards to form the treatment group.

EXAMPLE 3.12

Using software for the randomization. First create a data set with the numbers 1 to 10 in the first column. See Figure 3.4(a). Then we use RAND() to generate 10 random numbers in the second column. See Figure 3.4(b). Finally, we sort the data set based on the numbers in the second column. See Figure 3.4(c). The first five labels (7, 3, 4, 2, and 9) will receive the treatment. The remaining five labels (10, 6, 8, 1, and 5) will receive the placebo control.

Excel

	A	B
1	ID	
2	1	
3	2	
4	3	
5	4	
6	5	
7	6	
8	7	
9	8	
10	9	
11	10	

(a)

Excel

	A	B
1	ID	u
2	1	0.912229
3	2	0.372308
4	3	0.110239
5	4	0.141886
6	5	0.921569
7	6	0.649107
8	7	0.002333
9	8	0.739519
10	9	0.605416
11	10	0.632499

(b)

FIGURE 3.4 Randomization of 10 experimental units using an Excel spreadsheet, Example 3.12: (a) labels; (b) random numbers; (c) sorted list of labels; (d) labels with group assignments.

Excel

	A	B	C
1	ID	u	Group
2	7	0.002333	Treatment
3	3	0.110239	Treatment
4	4	0.141886	Treatment
5	2	0.372308	Treatment
6	9	0.605416	Treatment
7	10	0.632499	Control
8	6	0.649107	Control
9	8	0.739519	Control
10	1	0.912229	Control
11	5	0.921569	Control

(c)

Excel

	A	B
1	ID	Group
2	1	Control
3	2	Treatment
4	3	Treatment
5	4	Treatment
6	5	Control
7	6	Control
8	7	Treatment
9	8	Control
10	9	Treatment
11	10	Control

(d)

FIGURE 3.4 *Continued*

If you want to save the uniform numbers that you generated in your file, you should copy them to another column using the "paste values" option before you perform the sort. Note that we have added a column called Group to the spreadsheet, which gives the group to which each subject is assigned. With this variable included, we can now sort the file on ID and delete the column with the random numbers. The result is shown in Figure 3.4(d). The spreadsheet in this form can now be used as a template for entering data.

Randomization using random digits

You can randomize without software by using a *table of random digits*. Thinking about random digits helps you to understand randomization even if you will use software in practice. Table B at the back of the book is a table of random digits.

RANDOM DIGITS

A **table of random digits** is a list of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 that has the following properties:

1. The digit in any position in the list has the same chance of being any one of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
2. The digits in different positions are independent in the sense that the value of one has no influence on the value of any other.

You can think of Table B as the result of asking an assistant (or a computer) to mix the digits 0 to 9 in a hat, draw one, then replace the digit drawn, mix again, draw a second digit, and so on. The assistant's mixing and drawing save us the work of mixing and drawing when we need to randomize. Table B begins with the digits 19223950340575628713. To make the table easier to read, the digits appear in groups of five and in numbered rows. The groups and rows have no meaning—the table is just a long list of digits having Properties 1 and 2 described earlier.

Our goal is to use random digits for experimental randomization. We need the following facts about random digits, which are consequences of Properties 1 and 2:

- Any *pair* of random digits has the same chance of being any of the 100 possible pairs: 00, 01, 02, . . . , 98, 99.
- Any *triple* of random digits has the same chance of being any of the 1000 possible triples: 000, 001, 002, . . . , 998, 999.
- . . . and so on for groups of four or more random digits.

EXAMPLE 3.13

Randomize the subjects. Let's use random digits to perform the randomization that we performed using Excel in Example 3.12. Because the labels range from 1 to 10, we can use two digits for our labels

01, 02, 03, 04, 05, 06, 07, 08, 09, 10

when we select random digits from Table B. We could also have changed our labels to 0 through 9 and then we would only need to use single digits from Table B.

Start anywhere in Table B and read two-digit groups. Suppose we begin at line 175, which is

80011 09937 57195 33906 94831 10056 42211 65491

The first 10 two-digit groups in this line are

80 01 10 99 37 57 19 53 39 06

Each of these two-digit groups is a label. The labels 00 and 11 to 99 are not used in this example, so we ignore them. The first 10 labels between 01 and 10 that we encounter in the table choose subjects who will receive the treatment. Of the first 10 labels in line 175, we ignore seven because they are too high (over 10). The others are 01, 10, and 06. Continue across line 175 and 176 and verify that the next two subjects selected correspond to labels 03 and 04. Our randomization has selected subjects 1, 3, 4, 6, and 10 to receive the treatment. The remaining subjects, 2, 5, 7, 8, and 9 will receive the placebo control.

completely randomized design

When all experimental units are allocated at random among all treatments, as in Examples 3.12 and 3.13, the experimental design is **completely randomized**. Completely randomized designs can compare any number of treatments. The treatments can be formed by levels of a single factor or by more than one factor.

EXAMPLE 3.14

Randomization for the TV commercial experiment. Figure 3.2 (page 173) displays six treatments formed by the two factors in an experiment on response to a TV commercial. Suppose that we have 150 students who are willing to serve as subjects. We must assign 25 students at random to each group. Figure 3.5 outlines the completely randomized design.

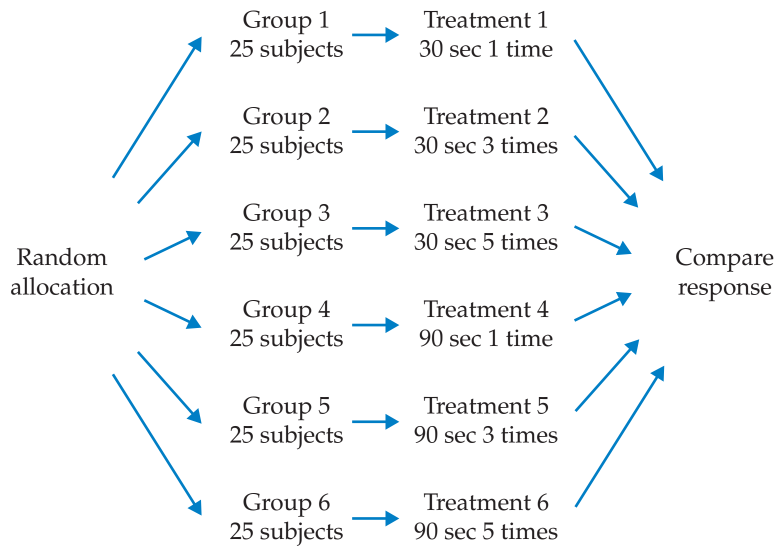

FIGURE 3.5 Outline of a completely randomized design comparing six treatments, Example 3.14.

To carry out the random assignment, label the 150 students 001 to 150. (Three digits are needed to label 150 subjects.) Using Excel, we would generate a uniform random variable for each label and sort the file as we did in Example 3.12. The first 25 students in this sorted file will receive Treatment 1, the next 25 will receive Treatment 2, etc.

Using random digits, we could enter Table B and read three-digit groups until you have selected 25 students to receive Treatment 1 (a 30-second ad shown once). If you start at line 140, the first few labels for Treatment 1 subjects are 129, 048, and 003.

Continue in Table B to select 25 more students to receive Treatment 2 (a 30-second ad shown three times). Then select another 25 for Treatment 3 and so on until you have assigned 125 of the 150 students to Treatments 1 through 5. The 25 students who remain get Treatment 6.

The randomization is straightforward but very tedious to do by using random digits. We strongly recommend that you use software, such as Excel or the *Simple Random Sample* applet. Exercise 3.37 (page 187) shows how to use the applet to do the randomization for this example.

USE YOUR KNOWLEDGE

3.23 Do the randomization. Use computer software to carry out the randomization in Example 3.14. Show your work by including the random uniform numbers in your final spreadsheet.

Cautions about experimentation

The logic of a randomized comparative experiment depends on our ability to treat all the experimental units identically in every way except for the actual treatments being compared. Good experiments, therefore, require careful attention to details. The ideal situation is where a study is **double-blind**—neither the

double-blind

subjects themselves nor the experimenters know which treatment any subject has received. The double-blind method avoids unconscious bias by, for example, a doctor who doesn't think that "just a placebo" can benefit a patient.

Many—perhaps most—experiments have some weaknesses in detail. The environment of an experiment can influence the outcomes in unexpected ways. Although experiments are the gold standard for evidence of cause and effect, really convincing evidence usually requires that a number of studies in different places with different details produce similar results. Here are some brief examples of what can go wrong.

EXAMPLE 3.15

Placebo for a marijuana experiment. A study of the effects of marijuana recruited young men who used marijuana. Some were randomly assigned to smoke marijuana cigarettes, while others were given placebo cigarettes. This failed: the control group recognized that their cigarettes were phony and complained loudly. It may be quite common for blindness to fail because the subjects can tell which treatment they are receiving.[11]

lack of realism

The most serious potential weakness of experiments is **lack of realism**. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study. Here is an example.

EXAMPLE 3.16

Layoffs and feeling bad. How do layoffs at a workplace affect the workers who remain on the job? To try to answer this question, psychologists asked student subjects to proofread text for extra course credit, then "let go" some of the workers (who were actually accomplices of the experimenters). Some subjects were told that those let go had performed poorly (Treatment 1). Others were told that not all could be kept and that it was just luck that they were kept and others let go (Treatment 2). We can't be sure that the reactions of the students are the same as those of workers who survive a layoff in which other workers lose their jobs. Many behavioral science experiments use student subjects in a campus setting. Do the conclusions apply to the real world?

Lack of realism can limit our ability to apply the conclusions of an experiment to the settings of greatest interest. Most experimenters want to generalize their conclusions to some setting wider than that of the actual experiment. *Statistical analysis of an experiment cannot tell us how far the results will generalize to other settings.* Nonetheless, the randomized comparative experiment, because of its ability to give convincing evidence for causation, is one of the most important ideas in statistics.

Matched pairs designs

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of control, randomization, and repetition. However, completely randomized designs are often inferior to more elaborate statistical designs. In particular, matching the subjects in various ways can produce more precise results than simple randomization.

matched pairs design

The simplest use of matching is a **matched pairs design**, which compares just two treatments. The subjects are matched in pairs. For example, an experiment to compare two advertisements for the same product might use pairs of subjects with the same age, sex, and income. The idea is that matched subjects are more similar than unmatched subjects so that comparing responses within a number of pairs is more efficient than comparing the responses of groups of randomly assigned subjects. Randomization remains important: which one of a matched pair sees the first ad is decided at random. One common variation of the matched pairs design imposes both treatments on the same subjects so that each subject serves as his or her own control. Here is an example.

EXAMPLE 3.17

Matched pairs for the smartphone prototype experiment. Example 3.11 describes an experiment to compare two prototypes of a new smartphone. The experiment compared two treatments: Phone 1 and Phone 2. The response variable is the satisfaction of the college student participant with the new smartphone. In Example 3.11, 40 student subjects were assigned at random, 20 students to each phone. This is a completely randomized design, outlined in Figure 3.3. Subjects differ in how satisfied they are with smartphones in general. The completely randomized design relies on chance to create two similar groups of subjects.

If we wanted to do a matched pairs version of this experiment, we would have each college student use each phone for two weeks. An effective design would randomize the *order* in which the phones are evaluated by each student. This will eliminate bias due to the possibility that the first phone evaluated will be systematically evaluated higher or lower than the second phone evaluated.

The completely randomized design uses chance to decide which subjects will evaluate each smartphone prototype. The matched pairs design uses chance to decide which 20 subjects will evaluate Phone 1 first. The other 20 will evaluate Phone 2 first. This experiment is called a **cross-over** experiment. Situations where there are more than two treatments and all subjects receive all treatments can also be performed in this way.

cross-over

Block designs

The matched pairs design of Example 3.17 uses the principles of comparison of treatments, randomization, and repetition on several experimental units. However, the randomization is not complete (all subjects randomly assigned to treatment groups) but is restricted to assigning the order of the treatments for each subject. *Block designs* extend the use of "similar subjects" from pairs to larger groups.

BLOCK DESIGN

A **block** is a group of experimental units or subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments. In a **block design**, the random assignment of units to treatments is carried out separately within each block.

Block designs can have blocks of any size. A block design combines the idea of creating equivalent treatment groups by matching with the principle of forming treatment groups at random. Blocks are another form of *control*. They control the effects of some outside variables by bringing those variables into the experiment to form the blocks. Here are some typical examples of block designs.

EXAMPLE 3.18

Blocking in a cancer experiment. The progress of a type of cancer differs in women and men. A clinical experiment to compare three therapies for this cancer then treats sex as a blocking variable. Two separate randomizations are done, one assigning the female subjects to the treatments and the other assigning the male subjects. Figure 3.6 outlines the design of this experiment. Note that there is no randomization involved in making up the blocks. They are groups of subjects who differ in some way (sex in this case) that is apparent before the experiment begins.

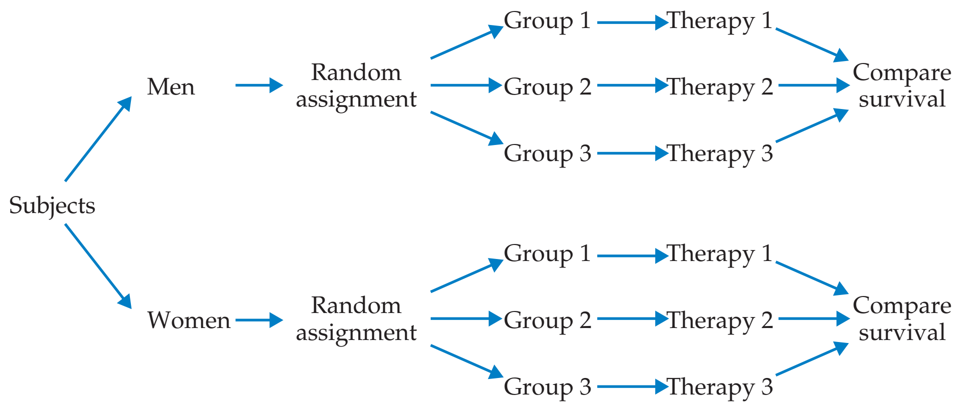

FIGURE 3.6 Outline of a block design, Example 3.18. The blocks consist of male and female subjects. The treatments are the three therapies for cancer.

EXAMPLE 3.19

Blocking in an agriculture experiment. The soil type and fertility of farmland differ by location. Because of this, a test of the effect of tillage type (two types) and pesticide application (three application schedules) on soybean yields uses small fields as blocks. Each block is divided into six plots, and the six treatments are randomly assigned to plots separately within each block.

EXAMPLE 3.20

Blocking in an education experiment. The Tennessee STAR class size experiment (Example 3.8, page 172) used a block design. It was important to compare different class types in the same school because the children in a school come from the same neighborhood, follow the same curriculum, and have the same school environment outside class. In all, 79 schools across Tennessee participated in the program. That is, there were 79 blocks. New kindergarten students were randomly placed in the three types of class separately within each school.

Blocks allow us to draw separate conclusions about each block, for example, about men and women in the cancer study in Example 3.18. Blocking also allows more precise overall conclusions because the systematic differences between men and women can be removed when we study the overall effects of the three therapies. The idea of blocking is an important additional principle of statistical design of experiments. A wise experimenter will form blocks based on the most important unavoidable sources of variability among the experimental units. Randomization will then average out the effects of the remaining variation and allow an unbiased comparison of the treatments.

SECTION 3.2 SUMMARY

- In an experiment, one or more **treatments** are imposed on the **experimental units** or **subjects**. Each treatment is a combination of **levels** of the explanatory variables, which we call **factors**. **Outcomes** are the measured variables that are used to compare the treatments.

- The **design** of an experiment refers to the choice of treatments and the manner in which the experimental units or subjects are assigned to the treatments.

- The basic principles of statistical design of experiments are **compare**, **randomization**, and **repetition**.

- The simplest form of control is **comparison**. Experiments should compare two or more treatments in order to prevent **confounding** the effect of a treatment with other influences, such as lurking variables.

- **Randomization** uses chance to assign subjects to the treatments. Randomization creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias**, or systematic favoritism, in experiments.

- You can carry out randomization by giving numerical labels to the experimental units and using a **table of random digits** to choose treatment groups.

- **Repetition** of the treatments on many units reduces the role of chance variation and makes the experiment more sensitive to differences among the treatments.

- Good experiments require attention to detail as well as good statistical design. Many behavioral and medical experiments are **double-blind**. **Lack of realism** in an experiment can prevent us from generalizing its results.

- In addition to comparison, a second form of control is to restrict randomization by forming **blocks** of experimental units that are similar in some way that is important to the response. Randomization is then carried out separately within each block.

- **Matched pairs** are a common form of blocking for comparing just two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, the subjects are matched in pairs as closely as possible, and one subject in each pair receives each treatment.

SECTION 3.2 EXERCISES

For Exercises 3.17 and 3.18, see page 173; for Exercises 3.19 and 3.20, see page 175; for Exercises 3.21 and 3.22, see page 176; and for Exercise 3.23, see page 181.

3.24 Blueberries and bones. A study of the effects of blueberries on the bones of mice compared diets containing no blueberries, blueberries as 5% of the diet, and blueberries as 10% of the diet. Ten mice were randomly assigned to each diet. The mice were fed the diets for 30 days, and the total body bone mineral density (TBBMD) was measured at the end of the feeding period. What are the experimental units, the treatments, and the outcomes for this experiment? Would you use the term *subjects* for the experimental units? Explain your answers.

3.25 Online homework. Thirty students participated in a study designed to evaluate a new online homework system. None of the students had used an online homework system in the past. After using the system for a month, they were asked to rate their satisfaction with the system using a five-point scale.

(a) What are the experimental units, the treatment, and the outcome for this experiment? Can we use the term *subjects* for the experimental units? Explain your answers.

(b) Is this a comparative experiment? If your answer is Yes, explain why. If your answer is No, describe how you would change the design so that it would be a comparative experiment.

(c) Suggest some different outcomes that you think would be appropriate for this experiment.

3.26 Do magnets reduce pain? Some claim that magnets can be used to reduce pain. Design a double-blind experiment to test this claim. Write a proposal requesting funding for your study giving all the important details, including the number of subjects, issues concerning randomization, and how you will make the study double-blind.

3.27 Online sales of running shoes. A company that sells running shoes online wants to compare two new marketing strategies. They will test the strategies on 10 weekdays. In the morning of each day, a web page describing the comfort of the running shoes will be displayed. In the afternoon of each day, a web page describing the discounted price for the shoes will be displayed. Sales of the featured running shoes in the morning will be compared with sales in the afternoon at the end of the experiment.

(a) What are the experimental units, the treatments, and the outcomes for this experiment? Explain your answers.

(b) Is this a comparative experiment? Why or why not?

(c) Could the experiment be improved by using randomization? Explain your answer.

(d) Could the experiment be improved by using a placebo treatment? Explain your answer.

3.28 Online sales of running shoes. Refer to the previous exercise. Suppose that for each day, you randomized the web pages, showing one in the morning and the other in the afternoon.

(a) Can you view this experiment as a block design? Explain your answer.

(b) Do you prefer this experiment or the one in the previous exercise? Give reasons for your answer.

3.29 Online sales of running shoes. Refer to Exercise 3.27. Here is another way in which the experiment could be designed. Suppose that you alternate the display each time a customer visits the website. Can you view this experiment as a matched pairs design? Explain your answer.

3.30 Randomize the web pages for the running shoes. Refer to Exercise 3.28. Use software or Table B to randomize the treatments. Give a step-by-step detailed description of how you performed the randomization.

3.31 What is needed? Explain what is deficient in each of the following proposed experiments and explain how you would improve the experiment.

(a) Two product promotion offers are to be compared. The first, which offers two items for $2, will be used in a store on Friday. The second, which offers three items for $3, will be used in the same store on Saturday.

(b) A study compares two marketing campaigns to encourage individuals to eat more fruits and vegetables. The first campaign is launched in Florida at the same time that the second campaign is launched in Minnesota.

(c) You want to evaluate the effectiveness of a new investment strategy. You try the strategy for one year and evaluate the performance of the strategy.

3.32 The *Madden* curse. Some people believe that individuals who appear on the cover of the football game *Madden NFL* will soon have a serious injury. Can you evaluate this belief with an experiment? Explain your answer.

3.33 Evaluate a new orientation program. Your company runs a two-day orientation program Monday and Tuesday each week for new employees. A new program is to be compared with the current one. Set up an experiment to compare the new program with the old. Be sure to provide details regarding randomization and what outcome variables you will measure.

3.34 What is wrong? Explain what is wrong with each of the following randomization procedures, and describe how you would do the randomization correctly.

(a) Twenty students are to be used to evaluate a new treatment. Ten men are assigned to receive the treatment, and 10 women are assigned to be the controls.

(b) Ten subjects are to be assigned to two treatments, five to each. For each subject, a coin is tossed. If the coin comes up heads, the subject is assigned to the first treatment; if the coin comes up tails, the subject is assigned to the second treatment.

(c) An experiment will assign 40 rats to four different treatment conditions. The rats arrive from the supplier in batches of 10, and the treatment lasts two weeks. The first batch of 10 rats is randomly assigned to one of the four treatments, and data for these rats are collected. After a one-week break, another batch of 10 rats arrives and is assigned to one of the three remaining treatments. The process continues until the last batch of rats is given the treatment that has not been assigned to the three previous batches.

3.35 Calcium and vitamin D. Vitamin D is needed for the body to use calcium. An experiment is designed to study the effects of calcium and vitamin D supplements on the bones of first-year college students. The outcome measure is the total body bone mineral content (TBBMC), a measure of bone health. Three doses of calcium will be used: 0, 250, and 500 milligrams per day (mg/day). The doses of vitamin D will be 0, 75, and 150 international units (IU) per day. The calcium and vitamin D will be given in a single tablet. All tablets, including those with no calcium and no vitamin D, will look identical. Subjects for the study will be 45 men and 45 women.

(a) What are the factors and the treatments for this experiment?

(b) Draw a picture explaining how you would randomize the 90 college students to the treatments.

(c) Use a spreadsheet to carry out the randomization.

(d) Is there a placebo in this experiment? Explain your answer.

3.36 Use the *Simple Random Sample* applet. You can use the *Simple Random Sample* applet to choose a group at random once you have labeled the subjects. Example 3.12 (page 178) uses Excel to choose five students from a group of 10 to receive a treatment in an experiment. The remaining five students will receive a placebo control.

(a) Use the applet to choose five students. Which students were selected?

(b) Compare using Excel, as we did in Example 3.12, with the applet that you used for this exercise. Which do you prefer? Give reasons for your answer.

3.37 Use the *Simple Random Sample* applet. The *Simple Random Sample* applet allows you to randomly assign experimental units to more than two groups without difficulty. Example 3.14 (page 180) describes a randomized comparative experiment in which 150 students are randomly assigned to six groups of 25.

(a) Use the applet to randomly choose 25 out of 150 students to form the first group. Which students are in this group?

(b) The "population hopper" now contains the 125 students who were not chosen, in scrambled order. Click "Sample" again to choose 25 of these remaining students to make up the second group. Which students were chosen?

(c) Click "Sample" three more times to choose the third, fourth, and fifth groups. Don't take the time to write down these groups. Check that there are only 25 students remaining in the "population hopper." These subjects get Treatment 6. Which students are they?

3.38 Use the *Simple Random Sample* applet. The *Simple Random Sample* applet can demonstrate how randomization works to create similar groups for comparative experiments. Suppose that (unknown to the experimenters) the 20 even-numbered students among the 40 subjects for the smartphone study in Example 3.11 (page 176) tend to send more text messages than the odd-numbered students. We would like the two groups to be similar with respect to text messaging. Use the applet to choose 10 samples of size 20 from the 40 students. (Be sure to click "Reset" after each sample.) Record the counts of even-numbered students in each of your 10 samples. You see that there is considerable chance variation but no systematic bias in favor of one or the other group in assigning the fast-reacting students. Larger samples from larger populations will, on the average, do a better job of making the two groups equivalent.

3.39 Health benefits of bee pollen. "Bee pollen is effective for combating fatigue, depression, cancer, and colon disorders." So says a website that offers the pollen for sale. We wonder if bee pollen really does prevent colon disorders. Here are two ways to study this question. Explain why the first design will produce more trustworthy data.

(a) Find 400 women who do not have colon disorders. Randomly assign 200 to take bee pollen capsules and the other 200 to take placebo capsules that are identical in appearance. Follow both groups for five years.

(b) Find 200 women who take bee pollen regularly. Match each with a woman of the same age, race, and occupation who does not take bee pollen. Follow both groups for five years.

3.40 Random digits. Table B is a table of random digits. Which of the following statements are true of a table of random digits, and which are false? Explain your answers.

(a) There are exactly four 0s in each row of 40 digits.

(b) Each pair of digits has chance 1/100 of being 00.

(c) The digits 0000 can never appear as a group because this pattern is not random.

3.41 Calcium and the bones of young girls. Calcium is important to the bone development of young girls. To study how the bodies of young girls process calcium, investigators used the setting of a summer camp. Calcium was given in punch at either a high or a low level. The camp diet was otherwise the same for all girls. Suppose that there are 30 campers.

(a) Outline a completely randomized design for this experiment.

(b) Use software or Table B to do the randomization. Explain in step-by-step detail how you carried out how you performed the randomization.

(c) Make a table giving the treatment that each camper will receive.

3.42 Calcium and the bones of young girls. Refer to the previous exercise.

(a) Outline a matched pairs design in which each girl receives both levels of calcium (with a "washout period" in which no calcium supplementation was given between the two treatment periods).

(b) What is the advantage of the matched pairs design over the completely randomized design?

(c) The same randomization can be used in different ways for both designs. Explain why this is true.

(d) Use software or Table B to do the randomization. Explain what each subject will do for the matched pairs design.

3.43 Measuring water quality in streams and lakes. Water quality of streams and lakes is an issue of concern to the public. Although trained professionals typically are used to take reliable measurements, many volunteer groups are gathering and distributing information based on data that they collect.[12] You are part of a team to train volunteers to collect accurate water quality data. Design an experiment to evaluate the effectiveness of the training. Write a summary of your proposed design to present to your team. Be sure to include all the details that they will need to evaluate your proposal.

3.3 Sampling Design

When you complete this section, you will be able to:

- Distinguish between a population and a sample.
- Use the response rate to evaluate a survey.
- Use Table B to generate a simple random sample (SRS).
- Use software to generate a simple random sample.
- Construct a stratified random sample using Table B or software to select the samples from the strata.
- Identify voluntary response samples, simple random samples, stratified random samples, and multistage random samples.
- Identify characteristics of samples that limit their usefulness, including undercoverage, nonresponse, response bias, and the wording of questions.

A political scientist wants to know what percent of college-age adults consider themselves conservatives. An automaker hires a market research firm to learn what percent of adults aged 18 to 35 recall seeing television advertisements for a new sports utility vehicle. Government economists inquire about average household income.

In all these cases, we want to gather information about a large group of individuals. We will not, as in an experiment, impose a treatment in order to observe the response. Also, time, cost, and inconvenience forbid contacting every individual. In such cases, we gather information about only part of the group—a *sample*—in order to draw conclusions about the whole. **Sample surveys** are an important kind of observational study.

sample survey

POPULATION AND SAMPLE

The entire group of individuals that we want information about is called the **population**.

A **sample** is a part of the population that we actually examine in order to gather information.

Notice that "population" is defined in terms of our desire for knowledge. If we wish to draw conclusions about all U.S. college students, that group is our population even if only local students are available for questioning. The sample is the part from which we draw conclusions about the whole. The **design of a sample survey** refers to the method used to choose the sample from the population.

sample design

EXAMPLE 3.21

hartcreations/iStockphoto

The Reading Recovery program. The Reading Recovery (RR) program has specially trained teachers work one-on-one with at-risk first-grade students to help them learn to read. A study was designed to examine the relationship between the RR teachers' beliefs about their ability to motivate students and the progress of the students whom they teach.[13] The Reading Recovery International Data Evaluation Center website (**www.idecweb.us**) says that there are 13,823 RR teachers. The researchers send a questionnaire to a random sample of 200 of these. The population consists of all 13,823 RR teachers, and the sample is the 200 that were randomly selected.

Unfortunately, our idealized framework of population and sample does not exactly correspond to the situations that we face in many cases. In Example 3.21, the list of teachers was prepared at a particular time in the past. It is very likely that some of the teachers on the list are no longer working as RR teachers today. New teachers have been trained in RR methods and are not on the list. Despite these difficulties, we still view the list as the population. Also, we may have out-of-date addresses for some who are still working as RR teachers, and some teachers may choose not to respond to the survey questions.

In reporting the results of a sample survey, it is important to include all details regarding the procedures used. Follow-up mailings or phone calls to those who do not initially respond can help increase the response rate. The proportion of the original sample who actually provide usable data is called the **response rate** and should be reported for all surveys. If only 150 of the teachers who were sent questionnaires provided usable data, the response rate would be 150/200, or 75%.

response rate

USE YOUR KNOWLEDGE

3.44 Are they satisfied? An educational research team wanted to examine the relationship between faculty participation in decision making and job satisfaction in Mongolian public universities. They are planning to randomly select 250 faculty members from a list of 2000 faculty members in these universities. The Job Descriptive Index will be used to measure job satisfaction, and the Conway Adaptation of the

Alutto-Belasco Decisional Participation Scale will be used to measure decision participation.

(a) Describe the population for this study.

(b) Describe the sample for this study.

(c) How would you determine the response rate for this study? Can you calculate it from the information given? If your answer is yes, calculate it.

3.45 What is the impact of the taxes? A study was designed to assess the impact of taxes on forest land usage in part of the Upper Wabash River Watershed in Indiana.[14] A survey was sent to 772 forest owners from this region, and 348 were returned.

(a) What is the sample for this study?

(b) What is the population for this study?

(c) How would you determine the response rate for this study? Can you calculate it from the information given? If your answer is yes, calculate it.

Poor sample designs can produce misleading conclusions. Here is an example.

EXAMPLE 3.22

Sampling pieces of steel. A mill produces large coils of thin steel for use in manufacturing home appliances. The quality engineer wants to submit a sample of 5-centimeter squares to detailed laboratory examination. She asks a technician to cut a sample of 10 such squares. Wanting to provide "good" pieces of steel, the technician carefully avoids the visible defects in the coil material when cutting the sample. The laboratory results are wonderful, but the customers complain about the material they are receiving.

LOOK BACK
bias, p. 74

In Example 3.22, the sample was selected in a manner that guaranteed that it would not be representative of the entire population. This sampling scheme displays *bias,* or systematic error, in favoring some parts of the population over others.

Online polls use *voluntary response samples,* a particularly common form of biased sample. The sample who respond are not representative of the population at large. People who take the trouble to respond to an open invitation are not representative of the entire population.

VOLUNTARY RESPONSE SAMPLE

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples are biased because people with strong opinions, especially negative opinions, are most likely to respond.

The remedy for bias in choosing a sample is to allow chance to do the choosing so that there is neither favoritism by the sampler (Example 3.22) nor voluntary response (online opinion polls). Random selection of a sample eliminates bias by giving all individuals an equal chance to be chosen, just as randomization eliminates bias in assigning experimental units.

Simple random samples

The simplest sampling design amounts to placing names in a hat (the population) and drawing out a handful (the sample). This is *simple random sampling*.

SIMPLE RANDOM SAMPLE

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

Each treatment group in a completely randomized experimental design is an SRS drawn from the available experimental units. We select an SRS by labeling all the individuals in the population and using software or a table of random digits to select a sample of the desired size, just as in experimental randomization. Notice that an SRS not only gives every possible sample an equal chance to be chosen, but also gives each individual an equal chance to be chosen. There are other random sampling designs that give each individual, but not each sample, an equal chance. One such design, systematic random sampling, is described in Exercise 3.64 (page 202).

How to select a simple random sample

The basic ideas needed to select a simple random sample are very similar to those that we discussed when we randomized subjects to treatments (page 177). We first assign a label to each case in our population. Then we perform the randomization using software or random digits from Table B.

Selection of a simple random sample using software The World Bank collects information about starting businesses throughout the world. In Example 1.23 (page 26) and several other examples in Chapter 1, we examined the time to start a business in a subset of these countries. For those exercises, we used a subset of the data because it was easier to show some details about our calculations with a smaller amount of data.

Now, suppose we want to collect additional information about countries that would help us to understand the processes of starting a business. The complete data set contains entries for 189 countries, and the time required to collect the additional information on all these would be too much. Let's use Excel to select a sample of 25 countries for a more detailed examination of these countries.

EXAMPLE 3.23

TTS

Select an SRS of countries using Excel. The data file TTS includes columns for the country name and a three-letter country code for each of the 189 countries. We could use either of these for our label. We will use the three letter codes in our screen shots to save space.

Figure 3.7(a) shows the codes for the first ten countries. In Figure 3.7(b), we show the uniform numbers generated with the RAND() function (and

Excel

	B	
1	CountryCode	
2	AFG	
3	AGO	
4	ALB	
5	ARE	
6	ARG	
7	ARM	
8	ATG	
9	AUS	
10	AUT	
11	AZE	

(a)

Excel

	B	C
1	CountryCode	u
2	AFG	0.529712245
3	AGO	0.638512682
4	ALB	0.188561334
5	ARE	0.634620202
6	ARG	0.089508274
7	ARM	0.535028604
8	ATG	0.52024832
9	AUS	0.896995737
10	AUT	0.897389837
11	AZE	0.278114291

(b)

Excel

	B	C
1	CountryCode	u
2	CHL	0.006224199
3	BEL	0.022409567
4	SLB	0.022793342
5	FJI	0.023409422
6	GRD	0.029450432
7	BIH	0.029970504
8	BLR	0.0394511
9	KHM	0.068092371
10	TUR	0.081259656
11	RWA	0.084609957

(c)

FIGURE 3.7 Selection of a simple random sample of countries from the population of 189 countries, Example 3.23.

then pasted into column C). The first three countries are Afghanistan (AFG), Angola (AGO), and Albania (ALB).

Figure 3.7(c) shows the file after we sort on the uniform random numbers in column C. Our sample is the first 25 countries in the sorted file. The first three selected are Chile (CHL), Belgium (BEL), and the Solomon Islands (SLB). Note that Excel does not display the last two digits for Belarus (BLR) because they are zero.

The *Simple Random Sample* applet on the text website is another convenient way to automate this task.

USE YOUR KNOWLEDGE

3.46 Select an SRS. Use the *Simple Random Sample* applet or Excel to select an SRS of five countries from the TTS data file. Include a step-by-step detailed description of how you selected the countries.

TTS

Selection of a simple random sample using random digits We illustrate the procedure by selecting an SRS of countries from the population of 189 countries in the data file TTS. Recall that we used Excel to select such a sample in Example 3.23.

EXAMPLE 3.24

Select an SRS of countries using random digits. To use Table B, we need a numeric label. We could create such a label by adding a column to the data file TTS containing the numbers 1 to 189. An alternative requiring less work would be to use the numbers in the leftmost part of the spreadsheet. Notice in Figure 3.7(a), for example, that there is a 1 in the first row of the spreadsheet where we have entered the names of the variables in the columns. Therefore, the numbers corresponding to countries run from 2 through 190. We will use these numbers as our label.

We will examine the entries in Table B in sets of three. Three digit numbers between 2 and 190 will correspond to selected countries. We will ignore three digit numbers equal to 000, 001, or greater than 190. Let's start our selection at line 106 in Table B. The entries on this line are

68417 35013 15529 72765 85089 57067 50211 47487

If we arrange these into sets of three, we have

684 173 501 315 529 727 658 508 057 067 502 114 748 7

The selected labels from this set of random digits are 173, 057, and 067. Checking the spreadsheet, we see that these numbers correspond to Turkey, France, and Greece.

Note that we do not use the last digit on line 106 to select the country with the label 7. We should combine this single digit with the first two digits from line 107 of Table B. This gives us the three-digit number 782, which is a number that we ignore. We complete our selection of the additional 22 countries that we need in our SRS using additional lines from Table B as needed.

USE YOUR KNOWLEDGE

3.47 Find the next three countries to be selected. Continue the process described in Example 3.24 to select the next three countries for the SRS. Show your work.

3.48 Listen to three rock songs. The walk to your statistics class takes about 10 minutes, about the amount of time needed to listen to three songs on your iPod or smartphone. You decide to take a simple random sample of songs from the top 10 listed on a Billboard Hot Rock Songs.[15] Here is the list:

Shut Up and Dance	Uma Thurman	Renegades	Ex's & Oh's
Centuries	Cecilia and the Satellite	Tear in My Heart	Brother
Stressed Out	Shots		

Select your three hot rock songs using a simple random sample. Show your work.

3.49 Listen to three songs. Refer to the previous exercise. Suppose that you like to include more variety in your music, so you look at the Billboard Top 100 songs.[16] Here are the top 10 on this list:

Cheerleader	Can't Feel My Face	Watch Me	Lean On
The Hills	Good for You	Fight Song	679
Trap Queen	Shut Up and Dance		

Select the three songs for your iPod or smartphone using a simple random sample. Show your work.

EXAMPLE 3.25

Select an SRS of countries using JMP. Refer to Example 3.23, where we selected an SRS of countries from a population of 189 countries. Also see Example 3.24, where we used the random digits in Table B to select the SRS. We can also use JMP to select an SRS. The output for the first 17 countries selected is displayed in Figure 3.8. JMP provides a file with the selected countries with all columns in the original data file. Note that the selected files are listed in the order that they appear in the original file, which is alphabetically in this case.

JMP

3/0 25/0	CountryName	CountryCode
1	Bulgaria	BGR
2	Bahrain	BHR
3	Bahamas, The	BHS
4	Bolivia	BOL
5	Barbados	BRB
6	Bhutan	BTN
7	Costa Rica	CRI
8	Dominican Republic	DOM
9	Micronesia, Fed. Sts.	FSM
10	Equatorial Guinea	GNQ
11	Jordan	JOR
12	Japan	JPN
13	Kiribati	KIR
14	Korea, Rep.	KOR
15	St. Lucia	LCA
16	Marshall Islands	MHL
17	Malaysia	MYS

FIGURE 3.8 JMP output for selecting an SRS of 25 countries from a population of 189 countries, Example 3.25.

AU/DE/PE: add to Fig caption: "Only the first 17 are displayed"?

Stratified random samples

The general framework for designs that use chance to choose a sample is a *probability sample*.

PROBABILITY SAMPLE

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Some probability sampling designs (such as an SRS) give each member of the population an *equal* chance to be selected. This may not be true in more elaborate sampling designs. In every case, however, the use of chance to select the sample is the essential principle of statistical sampling.

Designs for sampling from large populations spread out over a wide area are usually more complex than an SRS. For example, it is common to sample important groups within the population separately, then combine these samples. This is the idea of a *stratified sample*.

STRATIFIED RANDOM SAMPLE

To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Choose the strata based on facts known before the sample is taken. For example, a population of election districts might be divided into urban, suburban, and rural strata.

A stratified design can produce more exact information than an SRS of the same size by taking advantage of the fact that individuals in the same stratum are similar to one another. Think of the extreme case in which all individuals in each stratum are identical: just one individual from each stratum is then enough to completely describe the population.

Strata for sampling are similar to blocks in experiments. We have two names because the idea of grouping similar units before randomizing arose separately in sampling and in experiments.

EXAMPLE 3.26

A stratified sample of countries. In Examples 3.23 and 3.24, we selected SRSs of size 25 from the population of 189 countries in the World Bank file with data on starting businesses. Let's think about using a stratified sample. You still want to select 25 companies to examine in detail.

Let's classify each of the countries as located in Asia, Africa, Europe, North or South America, and Other. We have five strata, and we want a total of 25 countries to examine in detail. Therefore, we need to sample five countries from each stratum. We take an SRS of size 5 from each of these strata.

Multistage random samples

multistage random sample

Another common means of restricting random selection is to choose the sample in stages. These designs are called **multistage designs**. They are widely used in national samples of households or people. For example, data on employment and unemployment are gathered by the government's Current Population Survey, which conducts interviews in about 60,000 households each month. The cost of sending interviewers to the widely scattered households in an SRS would be too high. Moreover, the government wants data broken down by states and large cities.

Thus, the Current Population Survey uses a multistage random sampling design. The final sample consists of groups of nearby households, called clusters

clusters, that an interviewer can easily visit. Most opinion polls and other national samples are also multistage, though interviewing in most national samples today is done by telephone rather than in person, eliminating the economic need for clustering. The Current Population Survey sampling design is roughly as follows:[17]

Stage 1. Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. PSUs do not cross state lines. Select a sample of 754 PSUs. This sample includes the 428 PSUs with the largest population and a stratified sample of 326 of the others.

Stage 2. Divide each PSU selected into smaller areas called "blocks." Stratify the blocks using ethnic and other information, and take a stratified sample of the blocks in each PSU.

Stage 3. Sort the housing units in each block into clusters of four nearby units. Interview the households in a probability sample of these clusters.

Analysis of data from sampling designs more complex than an SRS takes us beyond basic statistics. But the SRS is the building block of more elaborate designs, and analysis of other designs differs more in complexity of detail than in fundamental concepts.

Cautions about sample surveys

Random selection eliminates bias in the choice of a sample from a list of the population. Sample surveys of large human populations, however, require much more than a good sampling design.[18] To begin, we need an accurate and complete list of the population. Because such a list is rarely available, most samples suffer from some degree of *undercoverage*. A sample survey of households, for example, will miss not only homeless people, but also prison inmates and students in dormitories. An opinion poll conducted by telephone will miss the large number of American households without residential phones. The results of national sample surveys, therefore, have some bias if the people not covered—who most often are poor people—differ from the rest of the population.

A more serious source of bias in most sample surveys is *nonresponse*, which occurs when a selected individual cannot be contacted or refuses to cooperate. Nonresponse to sample surveys often reaches 50% or more, even with careful planning and several callbacks. Because nonresponse is higher in urban areas, most sample surveys substitute other people in the same area to avoid favoring rural areas in the final sample. If the people contacted differ from those who are rarely at home or who refuse to answer questions, some bias remains.

UNDERCOVERAGE AND NONRESPONSE

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or does not cooperate.

EXAMPLE 3.27

Nonresponse in the Current Population Survey. How bad is nonresponse? The Current Population Survey (CPS) has the lowest nonresponse rate of any poll we know: only about 5% of the households in the CPS sample refuse to take part, and another 2% or 3% can't be contacted.[19] People are more likely to respond to a government survey such as the CPS, and the CPS contacts its sample in person before doing later interviews by phone.

The General Social Survey (Figure 3.9) is the nation's most important social science research survey. The GSS also contacts its sample in person, and it is run by a university. Despite these advantages, its most recent survey had a 30% rate of nonresponse.[20]

FIGURE 3.9 Part of the home page for the General Social Survey (GSS). The GSS has assessed attitudes on a wide variety of topics since 1972. Its continuity over time makes the GSS a valuable source for studies of changing attitudes. (Source: GSS)

What about polls done by the media and by market research and opinion-polling firms? Often, we don't know their rates of nonresponse because they won't say. That itself is a bad sign.

EXAMPLE 3.28

Change in nonresponse in Pew surveys. The Pew Research Center conducts research using surveys on a variety of issues, attitudes, and trends.[21] A study by the center examined the decline in the response rates to their surveys over time. The changes are dramatic, and there is a consistent pattern over time. Here are some data from the report:[22]

Year	1997	2000	2003	2006	2009	2012
Nonresponse rate	64%	72%	75%	79%	85%	91%

The center is devising alternative methods that show some promise of improving the response rates of their surveys.

Most sample surveys, and almost all opinion polls, are now carried out by telephone. This and other details of the interview method can affect the results. When presented with several options for a reply, such as "completely agree," "mostly agree," "mostly disagree," and "completely disagree," people tend to be a little more likely to respond to the first one or two options presented.

response bias

The behavior of the respondent or of the interviewer can cause **response bias** in sample results. Respondents may lie, especially if asked about illegal or unpopular behavior. The race or sex of the interviewer can influence responses to questions about race relations or attitudes toward feminism. Answers to questions that ask respondents to recall past events are often inaccurate because of faulty memory. For example, many people "telescope" events in the past, bringing them forward in memory to more recent time periods. "Have you visited a dentist in the last six months?" will often elicit a Yes from someone who last visited a dentist eight months ago.

wording of questions

The **wording of questions** is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and even minor changes in wording can change a survey's outcome. Here are some examples.

EXAMPLE 3.29

The form of the question is important. In response to the question "Are you heterosexual, homosexual, or bisexual?" in a social science research survey, one woman answered, "It's just me and my husband, so bisexual." The issue is serious, even if the example seems silly: reporting about sexual behavior is difficult because people understand and misunderstand sexual terms in many ways.

How do Americans feel about government help for the poor? Only 13% think we are spending too much on "assistance to the poor," but 44% think we are spending too much on "welfare." How do the Scots feel about the movement to become independent from England? Well, 51% would vote for "independence for Scotland," but only 34% support "an independent Scotland separate from the United Kingdom." It seems that "assistance to the poor" and "independence" are nice, hopeful words. "Welfare" and "separate" are negative words.[23]

The statistical design of sample surveys is a science, but this science is only part of the art of sampling. Because of nonresponse, response bias, and the difficulty of posing clear and neutral questions, you should hesitate to fully trust reports about complicated issues based on surveys of large human populations. *Insist on knowing the exact questions asked, the rate of nonresponse, and the date and method of the survey before you trust a poll result.*

BEYOND THE BASICS

Capture-Recapture Sampling

Sockeye salmon return to reproduce in the river where they were hatched four years earlier. How many salmon survived natural perils and heavy fishing to make it back this year? How many mountain sheep are there in Colorado? Are migratory songbird populations in North America decreasing or holding their own? These questions concern the size of animal populations. Biologists address them with a special kind of repeated sampling, called *capture-recapture sampling*.

EXAMPLE 3.30

Estimate the number of least flycatchers. You are interested in the number of least flycatchers migrating along a major route in the north-central United States. You set up "mist nets" that capture the birds but do not harm them. The birds caught in the net are fitted with a small aluminum leg band and released. Last year, you banded and released 200 least flycatchers. This year, you repeat the process. Your net catches 120 least flycatchers, 12 of which have tags from last year's catch.

The proportion of your second sample that have bands should estimate the proportion in the entire population that are banded. So if N is the unknown number of least flycatchers, we should have approximately

$$\text{proportion banded in sample} = \text{proportion banded in population}$$

$$\frac{12}{120} = \frac{200}{N}$$

Solve for N to estimate that the total number of flycatchers migrating while your net was up this year is approximately

$$N = 200 \times \frac{120}{12} = 2000$$

The capture-recapture idea extends the use of a sample proportion to estimate a population proportion. The idea works well if both samples are SRSs from the population and the population remains unchanged between samples. In practice, complications arise because, for example, some of the birds tagged last year died before this year's migration.

Variations on capture-recapture samples are widely used in wildlife studies and are now finding other applications. One way to estimate the census undercount in a district is to consider the census as "capturing and marking" the households that respond. Census workers then visit the district, take an SRS of households, and see how many of those counted by the census show up in the sample. Capture-recapture estimates the total count of households in the district. As with estimating wildlife populations, there are many practical pitfalls. Our final word is as before: the real world is less orderly than statistics textbooks imply.

SECTION 3.3 SUMMARY

- A sample survey selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data about the sample.
- The **design** of a sample refers to the method used to select the sample from the population. **Probability sampling designs** use impersonal chance to select a sample.
- The basic probability sample is a **simple random sample (SRS)**. An SRS gives every possible sample of a given size the same chance to be chosen.
- Choose an SRS using software. This can also be done using a **table of random digits** to select the sample.
- To choose a **stratified random sample**, divide the population into **strata**, groups of individuals that are similar in some way that is important to the response. Then choose a separate SRS from each stratum, and combine them to form the full sample.
- **Multistage random samples** select successively smaller groups within the population in stages, resulting in a sample consisting of clusters of individuals. Each stage may employ an SRS, a stratified sample, or another type of sample.
- Failure to use probability sampling often results in **bias**, or systematic errors in the way the sample represents the population. **Voluntary response** samples, in which the respondents choose themselves, are particularly prone to large bias.
- In human populations, even probability samples can suffer from bias due to **undercoverage** or **nonresponse**, from **response bias** due to the behavior of the interviewer or the respondent, or from misleading results due to **poorly worded questions**.

SECTION 3.3 EXERCISES

For Exercises 3.44 and 3.45, see pages 189–190; for Exercise 3.46, see page 192; and for Exercises 3.47, 3.48, and 3.49, see page 193.

3.50 How many text messages? You would like to know something about how many text messages you will receive in the next 100 days. Counting the number for each of the 100 days would take more time than you would like to spend on this project, so you randomly select 10 days from the hundred to count.

(a) Describe the population for this setting.

(b) What is the sample?

3.51 Response rate? A survey designed to assess satisfaction with food items sold at a college's football games was sent to 150 fans who had season tickets. The total number of fans who have season tickets is 5674. Responses to the survey were received from 98 fans.

(a) Describe the population for this survey.

(b) What is the sample?

(c) What is the response rate?

(d) What is the nonresponse rate?

(e) Suggest some ways that could be used in a future survey to increase the response rate.

3.52 Interview some students. You are a teaching assistant for an introductory statistics class. The instructor would like you to interview some of the students in the class to find out their opinion regarding some new interactive activities that she has introduced to the course. There are 123 students in the class, so you cannot interview all of them. You decide to select eight students to interview.

(a) What is the population for this setting?

(b) What is the sample?

(c) Make a spreadsheet with the numeric labels for the 123 students in the class.

(d) Use Excel to select the labels of the eight students to be interviewed from the spreadsheet.

(e) Explain the steps that you used in sufficient detail so that another person could repeat your work.

3.53 Interview some students. Refer to the previous exercise.

(a) Use Table B to select the students. Give details.

(b) Compare the use of Table B with software for selecting the students. Which do you prefer? Give reasons for your answer.

3.54 What kind of sample? In each of the following situations, identify the sample as an SRS, a stratified random sample, a multistage random sample, or a voluntary response sample. Explain your answers.

(a) There are seven sections of an introductory statistics course. A random sample of three sections is chosen, and then random samples of eight students from each of these sections are chosen.

(b) A student organization has 55 members. A table of random numbers is used to select a sample of five.

(c) An online poll asks people who visit this site to choose their favorite television show.

(d) Separate random samples of male and female first-year college students in an introductory psychology course are selected to receive a one-week alternative instructional method.

3.55 What's wrong? Explain what is wrong in each of the following scenarios.

(a) The population consists of all individuals selected in a simple random sample.

(b) In a poll of an SRS of residents in a local community, respondents are asked to indicate the level of their concern about the dangers of dihydrogen monoxide, a substance that is a major component of acid rain and that, in its gaseous state, can cause severe burns. (*Hint:* Ask a friend who is majoring in chemistry about this substance or search the Internet for information about it.)

(c) Students in a class are asked to raise their hands if they have cheated on an exam one or more times within the past year.

3.56 What's wrong? Explain what is wrong with each of the following random selection procedures, and explain how you would do the randomization correctly.

(a) To determine the reading level of an introductory statistics text, you evaluate all the written material in the third chapter.

(b) You want to sample student opinions about a proposed change in procedures for changing majors. You hand out questionnaires to 100 students as they arrive for class at 7:30 A.M.

(c) A population of subjects is put in alphabetical order, and a simple random sample of size 10 is taken by selecting the first 10 subjects in the list.

AU: set as (a), (b), and (c)?

3.57 Importance of students as customers. A committee on community relations in a college town plans to survey local businesses about the importance of students as customers. From telephone book listings, the committee chooses 80 businesses at random. Of these, 46 return the questionnaire mailed by the committee. What is the population for this sample survey? What is the sample? What is the rate (percent) of nonresponse?

3.58 Identify the populations. For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not complete, complete the description of the population in a reasonable way.

(a) A college has changed its core curriculum and wants to obtain detailed feedback information from the students during each of the first 12 weeks of the coming semester. Each week, a random sample of five students will be selected to be interviewed.

(b) The American Community Survey (ACS) replaced the census "long form" starting with the 2010 census. The ACS contacts 250,000 addresses by mail each month, with follow-up by phone and in person if there is no response. Each household answers questions about their housing, economic, and social status.

(c) An opinion poll contacts 1161 adults and asks them, "Which political party do you think has better ideas for leading the country in the twenty-first century?"

3.59 Interview residents of apartment complexes. You are planning a report on apartment living in a college town. You decide to select eight apartment complexes at random for in-depth interviews with residents. Select a simple random sample of eight of the following apartment complexes. If you use Table B, start at line 136. RESIDEN

Ashley Oaks	Country View	Mayfair Village
Bay Pointe	Country Villa	Nobb Hill
Beau Jardin	Crestview	Pemberly Courts
Bluffs	Del-Lynn	Peppermill
Brandon Place	Fairington	Pheasant Run
Briarwood	Fairway Knolls	Richfield
Brownstone	Fowler	Sagamore Ridge
Burberry	Franklin Park	Salem Courthouse
Cambridge	Georgetown	Village Manor
Chauncey Village	Greenacres	Waterford Court
Country Squire	Lahr House	Williamsburg

3.60 Using GIS to identify mint field conditions. A Geographic Information System (GIS) is to be used to distinguish different conditions in mint fields. Ground observations will be used to classify regions of each field as either healthy mint, diseased mint, or weed-infested mint. The GIS divides mint-growing areas into regions called pixels. An experimental area contains 200 pixels. For a random sample of 15 pixels, ground measurements will be made to determine the status of the mint, and these observations will be compared with information obtained by the GIS. Select the random sample. If you use Table B, start at line 132 and choose only the first 15 pixels in the sample.

3.61 Use the *Simple Random Sample* applet. After you have labeled the individuals in a population, the *Simple Random Sample* applet automates the task of choosing an SRS. Use the applet to choose the sample in the previous exercise.

3.62 Select a simple random sample. There are 38 active telephone area codes in California. You want to choose an SRS of 10 of these area codes for a study of available telephone numbers. Label the codes 01 to 38 and use the *Simple Random Sample* applet, Table B, or software to choose your sample. (If you use Table B, start at line 131.)

3.63 Stratified samples for attitudes about alcohol. At a party, there are 25 students over age 21 and 15 students under age 21. You choose at random five of those over 21 and separately choose at random three of those under 21 to interview about attitudes toward alcohol. You have given every student at the party the same chance to be interviewed: what is that chance? Why is your sample not an SRS?

3.64 Systematic random samples. Systematic random samples are often used to choose a sample of apartments in a large building or dwelling units in a block at the last stage of a multistage sample. An example will illustrate the idea of a systematic sample. Suppose that we must choose five addresses out of 125. Because 125/5 = 25, we can think of the list as five lists of 25 addresses. Choose one of the first 25 at random, using software or Table B. The sample contains this address and the addresses 25, 50, 75, and 100 places down the list from it. If 13 is chosen, for example, then the systematic random sample consists of the addresses numbered 13, 38, 63, 88, and 113.

(a) A study of dating among college students wanted a sample of 200 of the 8000 single male students on campus. The sample consisted of every 40th name from a list of the 8000 students. Explain why the survey chooses every 40th name.

(b) Use software or Table B at line 112 to choose the starting point for this systematic sample.

3.65 Systematic random samples versus simple random samples. The previous exercise introduces systematic random samples. Explain carefully why a systematic random sample *does* give every individual the same chance to be chosen but is *not* a simple random sample.

3.66 Random digit telephone dialing. An opinion poll in California uses random digit dialing to choose telephone numbers at random. Numbers are selected separately within each California area code. The size of the sample in each area code is proportional to the population living there. AREACOD

(a) What is the name for this kind of sampling design?

(b) California area codes, in rough order from north to south, are

209	213	310	323	341	369	408	415	424	442
510	530	559	562	619	626	627	628	650	657
661	669	707	714	747	752	760	764	805	818
831	858	909	916	925	935	949	951		

Another California survey does not call numbers in all area codes but starts with an SRS of eight area codes. Choose such an SRS. If you use Table B, start at line 132.

3.67 Select club members to go to a convention. A club has 30 student members and 10 faculty members. The students are

Abel	Fisher	Huber	Moran	Reinmann
Carson	Golomb	Jimenez	Moskowitz	Santos
Chen	Griswold	Jones	Neyman	Shaw
David	Hein	Kiefer	O'Brien	Thompson
Deming	Hernandez	Klotz	Pearl	Utts
Elashoff	Holland	Liu	Potter	Vlasic

and the faculty members are

Andrews	Fernandez	Kim	Moore	Rabinowitz
Besicovitch	Gupta	Lightman	Phillips	Yang

The club can send seven students and three faculty members to a convention and decides to choose those who will go by random selection. Select a stratified random sample of seven students and three faculty members.

3.68 Stratified samples for accounting audits. Accountants use stratified samples during audits to verify a company's records of such things as accounts receivable. The stratification is based on the dollar amount of the item and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 100 are in amounts over \$50,000; 500 are in amounts between \$1000 and \$50,000; and the remaining 4400 are in amounts under \$1000. Using these groups as strata, you decide to verify all of the largest accounts and to sample 5% of the midsize accounts and 1% of the small accounts. How would you label the two strata from which you will sample? Use software or Table B, starting at line 125, to select the first six accounts from each of these strata.

3.69 The sampling frame. The list of individuals from which a sample is actually selected is called the **sampling frame**. Ideally, the frame should list every individual in the population, but in practice this is often difficult. A frame that leaves out part of the population is a common source of undercoverage.

(a) Suppose that a sample of households in a community is selected at random from the telephone directory. What households are omitted from this frame? What types of people do you think are likely to live in these households? These people will probably be underrepresented in the sample.

(b) It is usual in telephone surveys to use random digit dialing equipment that selects the last four digits of a telephone number at random after being given the area code and the exchange. The exchange is the first three digits of the telephone number. Which of the households that you mentioned in your answer to part (a) will be included in the sampling frame by random digit dialing?

3.70 Survey questions. Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

(a) "Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?"

(b) "Do you agree that a national system of health insurance should be favored because it would provide health insurance for everyone and would reduce administrative costs?"

(c) "In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?"

3.4 Ethics

When you complete this section, you will be able to:

- Describe the purpose of an institutional review board and describe what kinds of expertise its members require.
- Describe informed consent and evaluate whether or not it has been given in specific examples.
- Determine when data have been kept confidential in a study.
- Evaluate a clinical trial from the viewpoint of ethics.

The production and use of data, like all human endeavors, raise ethical questions. We won't discuss the telemarketer who begins a telephone sales pitch with "I'm conducting a survey." Such deception is clearly unethical. It enrages legitimate survey organizations, which find the public less willing to talk with them. Neither will we discuss those few researchers who, in the pursuit of professional advancement, publish fake data. There is no ethical question here—faking data to advance your career is just wrong. It will end your career when uncovered.

But just how honest must researchers be about real, unfaked data? Here is an example that suggests the answer is "More honest than they often are."

EXAMPLE 3.31

Provide all the critical information. Papers reporting scientific research are supposed to be short, with no extra baggage. But brevity can allow the researchers to avoid complete honesty about their data. Did they choose their subjects in a biased way? Did they report data on only some of their subjects? Did they try several statistical analyses and report only the ones that looked best?

The statistician John Bailar screened more than 4000 medical papers in more than a decade as consultant to the *New England Journal of Medicine*. He says, "When it came to the statistical review, it was often clear that critical information was lacking, and the gaps nearly always had the practical effect of making the authors' conclusions look stronger than they should have."[24] The situation is no doubt worse in fields that screen published work less carefully.

The most complex issues of data ethics arise when we collect data from people. The ethical difficulties are more severe for experiments that impose some treatment on people than for sample surveys that simply gather information. Trials of new medical treatments, for example, can do harm as well as good to their subjects. Here are some basic standards of data ethics that must be obeyed by any study that gathers data from human subjects, whether sample survey or experiment.

BASIC DATA ETHICS

The organization that carries out the study must have an **institutional review board** that reviews all planned studies in advance in order to protect the subjects from possible harm.

All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential**. Only statistical summaries for groups of subjects may be made public.

The law requires that studies funded by the federal government obey these principles. But neither the law nor the consensus of experts is completely clear about the details of their application.

Institutional review boards

The purpose of an institutional review board is not to decide whether a proposed study will produce valuable information or whether it is statistically sound. The board's purpose is, in the words of one university's board, "to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities." When protocols are greater than minimal risk, a statistician is often included on the board to help determine benefits.

The board reviews the plan of the study and can require changes. It reviews the consent form to be sure that subjects are informed about the nature of the study and about any potential risks. Once research begins, the board monitors its progress at least once a year.

The most pressing issue concerning institutional review boards is whether their workload has become so large that their effectiveness in protecting subjects drops. There are shorter review procedures for projects that involve only minimal risks to subjects, such as most sample surveys. When a board is overloaded, there is a temptation to put more proposals in the minimal-risk category to speed the work.

USE YOUR KNOWLEDGE

The exercises in this section on ethics are designed to help you think about the issues that we are discussing and to formulate some opinions. In general, there are no wrong or right answers, but you need to give reasons for your answers.

3.71 Do these proposals involve minimal risk? You are a member of your college's institutional review board. You must decide whether several research proposals qualify for lighter review because they involve only minimal risk to subjects. Federal regulations say that "minimal risk" means that the risks are no greater than "those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." That's vague. Which of these do you think qualifies as "minimal risk"? Explain your choices.

(a) Draw a drop of blood by pricking a finger in order to measure blood sugar.

(b) Draw blood from the arm for a full set of blood tests.

(c) Insert a tube that remains in the arm so that blood can be drawn regularly.

3.72 Who should be on an institutional review board? Government regulations require that institutional review boards consist of at least five people, including at least one scientist, one nonscientist, and one person from outside the institution. Most boards are larger, but many contain just one outsider.

(a) Why should review boards contain people who are not scientists?

(b) Do you think that one outside member is enough? How would you choose that member? (For example, would you prefer a medical doctor? A member of the clergy? An activist for patients' rights?)

Informed consent

Both words in the phrase "informed consent" are important, and both can be controversial. Subjects must be *informed* in advance about the nature of a study and any risk of harm it may bring. In the case of a sample survey, physical harm is not possible. The subjects should be told what kinds of questions the survey will ask and about how much of their time it will take. Experimenters must tell subjects the nature and purpose of the study and outline possible risks. Subjects must then *consent* in writing.

EXAMPLE 3.32

Who can give informed consent? Are there some subjects who can't give informed consent? It was once common, for example, to test new vaccines on prison inmates who gave their consent in return for good-behavior credit. Now we worry that prisoners are not really free to refuse, and the law forbids medical experiments in prisons.

Young children can't give fully informed consent, so the usual procedure is to ask their parents. A study of new ways to teach reading is about to start at a local elementary school, so the study team sends consent forms home to parents. Many parents don't return the forms. Can their children take part in the study because the parents did not say No, or should we allow only children whose parents returned the form and said Yes?

What about research into new medical treatments for people with mental disorders? What about studies of new ways to help emergency room patients who may be unconscious or have suffered a stroke? In most cases, there is not time even to get the consent of the family. Does the principle of informed consent bar realistic trials of new treatments for unconscious patients?

These are questions without clear answers. Reasonable people differ strongly on all of them. There is nothing simple about informed consent.[25]

The difficulties of informed consent do not vanish even for capable subjects. Some researchers, especially in medical trials, regard consent as a barrier to getting patients to participate in research. They may not explain all possible risks; they may not point out that there are other therapies that might be better than those being studied; they may be too optimistic in talking with patients even when the consent form has all the right details.

On the other hand, mentioning every possible risk leads to very long consent forms that really are barriers. "They are like rental car contracts," one lawyer said. Some subjects don't read forms that run five or six printed pages. Others are frightened by the large number of possible (but unlikely) disasters that might happen and so refuse to participate. Of course, unlikely disasters sometimes happen. When they do, lawsuits follow and the consent forms become yet longer and more detailed.

Confidentiality

Ethical problems do not disappear once a study has been cleared by the review board, has obtained consent from its subjects, and has actually collected data about the subjects. **Confidentiality** means that only the researchers can identify responses of individual subjects. The report of an opinion poll may say what percent of the 1500 respondents felt that legal immigration should be reduced. It may not report what *you* said about this or any other issue.

confidentiality

anonymity

Confidentiality is not the same as **anonymity**. Anonymity means that subjects are anonymous—their names are not known even to the director of the study. Anonymity is rare in statistical studies. Even where anonymity is possible (mainly in surveys conducted by mail), it prevents any follow-up to improve nonresponse or inform subjects of results.

Any breach of confidentiality is a serious violation of data ethics. The best practice is to separate the identity of the subjects from the rest of the data at once. Sample surveys, for example, use the identification only to check on who did or did not respond. In an era of advanced technology, however, it is no longer enough to be sure that each individual set of data protects people's privacy.

The government, for example, maintains a vast amount of information about citizens in many separate databases—census responses, tax returns, Social Security information, data from surveys such as the Current Population Survey, and so on. Many of these databases can be searched by computers for statistical studies.

A clever computer search of several databases might be able, by combining information, to identify you and learn a great deal about you even if your name and other identification have been removed from the data available for search. A colleague from Germany once remarked that "female full professor of statistics with a PhD from the United States" was enough to identify her among all the citizens of Germany. Privacy and confidentiality of data are hot issues among statisticians in the computer age.

EXAMPLE 3.33

Data collected by the government. Citizens are required to give information to the government. Think of tax returns and Social Security contributions. The government needs these data for administrative purposes—to see if we paid the right amount of tax and how large a Social Security benefit we are owed when we retire. Some people feel that individuals should be able to forbid any other use of their data, even with all identification removed. This would prevent using government records to study, say, the ages, incomes, and household sizes of Social Security recipients. Such a study could well be vital to debates on reforming Social Security.

USE YOUR KNOWLEDGE

3.73 How can we obtain informed consent? A researcher suspects that traditional religious beliefs tend to be associated with an authoritarian personality. She prepares a questionnaire that measures authoritarian tendencies and also asks many religious questions. Write a description of the purpose of this research to be read by subjects in order to obtain their informed consent. You must balance the conflicting goals of not deceiving the subjects as to what the questionnaire will tell about them and of not biasing the sample by scaring off religious people.

3.74 Should we allow this personal information to be collected? In which of the following circumstances would you allow collecting personal information without the subjects' consent?

(a) A government agency takes a random sample of income tax returns to obtain information on the average income of people in different occupations. Only the incomes and occupations are recorded from the returns, not the names.

(b) A social psychologist attends public meetings of a religious group to study the behavior patterns of members.

(c) A social psychologist pretends to be converted to membership in a religious group and attends private meetings to study the behavior patterns of members.

Clinical trials

Clinical trials are experiments that study the effectiveness of medical treatments on actual patients. Medical treatments can harm as well as heal, so clinical trials spotlight the ethical problems of experiments with human subjects. Here are the starting points for a discussion:

- Randomized comparative experiments are the only way to see the true effects of new treatments. Without them, risky treatments that are no better than placebos will become common.

- Clinical trials produce great benefits, but most of these benefits go to future patients. The trials also pose risks, and these risks are borne by the subjects of the trial. So we must balance future benefits against present risks.

- Both medical ethics and international human rights standards say that "the interests of the subject must always prevail over the interests of science and society."

The quoted words are from the 1964 Helsinki Declaration of the World Medical Association, the most respected international standard. The most outrageous examples of unethical experiments are those that ignore the interests of the subjects.

EXAMPLE 3.34

National Archives and Records Administration NARA

The Tuskegee study. In the 1930s, syphilis was common among black men in the rural South, a group that had almost no access to medical care. The Public Health Service Tuskegee study recruited 399 poor black sharecroppers with syphilis and 201 others without the disease in order to observe how syphilis progressed when no treatment was given. Beginning in 1943, penicillin became available to treat syphilis. The study subjects were not treated. In fact, the Public Health Service prevented any treatment until word leaked out and forced an end to the study in the 1970s.

The Tuskegee study is an extreme example of investigators following their own interests and ignoring the well-being of their subjects. A 1996 review said, “It has come to symbolize racism in medicine, ethical misconduct in human research, paternalism by physicians, and government abuse of vulnerable people.” In 1997, President Clinton formally apologized to the surviving participants in a White House ceremony.[26]

Because “the interests of the subject must always prevail,” medical treatments can be tested in clinical trials only when there is reason to hope that they will help the patients who are subjects in the trials. Future benefits aren’t enough to justify experiments with human subjects. Of course, if there is already strong evidence that a treatment works and is safe, it is unethical *not* to give it.

Here is the view of Dr. Charles Hennekens of the Harvard Medical School, who directed the large clinical trial that showed that aspirin reduces the risk of heart attacks:[27] A clinical trial is justified if there is some evidence that the treatment will be effective. This evidence, however, is not sufficiently strong, to conclude that we would be harming the subjects who would receive the placebo.

Why is it ethical to give a control group of patients a placebo? Well, we know that placebos often work. What is more, placebos have no harmful side effects. In fact, the placebo group may be getting a better treatment than the drug group. If we *knew* which treatment was better, we would give it to everyone. When we don’t know, it is ethical to try both and compare them.

The idea of using a control or placebo is a fundamental principle to be considered in designing experiments. In many situations, deciding what to use as an appropriate control requires some careful thought.

The choice of the control can have a substantial impact on how the results of an experiment are interpreted. Here is an example.

EXAMPLE 3.35

Attentiveness improves by nearly 20%. The manufacturer of a breakfast cereal designed for children claims that eating this cereal has been clinically shown to improve attentiveness by nearly 20%. The study used two groups of children who were tested before and after breakfast. One group received the cereal for breakfast, while breakfast for the control group was water. The results of the tests taken three hours after breakfast were used in the claim.

The Federal Trade Commission investigated the marketing of this product. It charged that the claim was false and violated federal law. The charges were settled and the company agreed to not use misleading claims in its advertising.[28]

It is not sufficient to obtain appropriate controls. The data must be collected from all groups in the same way. Here is an example of this type of flawed design.

EXAMPLE 3.36

Accurate identification of ovarian cancer. Two scientists published a paper claiming to have developed an exciting new method to detect ovarian cancer using blood samples. When other scientists were unable to reproduce the results in different labs, the original work was examined more carefully. In the original study, there were samples for women with ovarian cancer and for healthy controls. The blood samples were all analyzed using a mass spectrometer. The control samples were analyzed on one day, and the cancer samples were analyzed on the next day. This design was flawed in that it could not control for changes over time in the measuring instrument.[29]

USE YOUR KNOWLEDGE

3.75 Is this study ethical? Researchers on aging proposed to investigate the effect of supplemental health services on the quality of life of older people. Eligible patients on the rolls of a large medical clinic were to be randomly assigned to treatment and control groups. The treatment group would be offered hearing aids, dentures, transportation, and other services not available without charge to the control group. The review board felt that providing these services to some but not other persons in the same institution raised ethical questions. Do you agree?

3.76 Should the treatments be given to everyone? Effective drugs for treating AIDS are very expensive, so most African nations cannot afford to give them to large numbers of people. Yet AIDS is more common in parts of Africa than anywhere else. Several clinical trials are looking at ways to prevent pregnant mothers infected with HIV from passing the infection to their unborn children, a major source of HIV infections in Africa. Some people say these trials are unethical because they do not give effective AIDS drugs to their subjects, as would be required in rich nations. Others reply that the trials are looking for treatments that can work in the real world in Africa and that they promise benefits at least to the children of their subjects. What do you think?

Behavioral and social science experiments

When we move from medicine to the behavioral and social sciences, the direct risks to experimental subjects are less acute, but so are the possible benefits to the subjects. Consider, for example, the experiments conducted by psychologists in their study of human behavior.

EXAMPLE 3.37

Personal space. Psychologists observe that people have a "personal space" and get annoyed if others come too close to them. We don't like strangers to sit at our table in a coffee shop if other tables are available, and we see people move apart in elevators if there is room to do so. Americans tend to require more personal space than people in most other cultures. Can violations of personal space have physical, as well as emotional, effects?

Investigators set up shop in a men's public restroom. They blocked off urinals to force men walking in to use either a urinal next to an experimenter (treatment group) or a urinal separated from the experimenter (control group). Another experimenter, using a periscope from a toilet stall, measured how long the subject took to start urinating and how long he kept at it.[30]

This personal space experiment illustrates the difficulties facing those who plan and review behavioral studies:

- There is no risk of harm to the subjects, although they would certainly object to being watched through a periscope. What should we protect subjects from when physical harm is unlikely? Possible emotional harm? Undignified situations? Invasion of privacy?
- What about informed consent? The subjects in Example 3.37 did not even know they were participating in an experiment. Many behavioral experiments rely on hiding the true purpose of the study. The subjects would change their behavior if told in advance what the investigators were looking for. Subjects are asked to consent on the basis of vague information. They receive full information only after the experiment.

The "Ethical Principles" of the American Psychological Association require consent unless a study merely observes behavior in a public place. They allow deception only when it is necessary to the study, does not hide information that might influence a subject's willingness to participate, and is explained to subjects as soon as possible. The personal space study (from the 1970s) does not meet current ethical standards.

We see that the basic requirement for informed consent is understood differently in medicine and psychology. Here is an example of another setting with yet another interpretation of what is ethical. The subjects get no information and give no consent. They don't even know that an experiment may be sending them to jail for the night.

EXAMPLE 3.38

Domestic violence. How should police respond to domestic violence calls? In the past, the usual practice was to remove the offender and order him to stay out of the household overnight. Police were reluctant to make arrests because the victims rarely pressed charges. Women's groups argued that arresting offenders would help prevent future violence even if no charges were filed. Is there evidence that arrest will reduce future offenses? That's a question that experiments have tried to answer.

A typical domestic violence experiment compares two treatments: arrest the suspect and hold him overnight, or warn the suspect and release him. When police officers reach the scene of a domestic violence call, they calm the participants and investigate. Weapons or death threats require an arrest. If the facts permit an arrest but do not require it, an officer radios headquarters for instructions. The person on duty opens the next envelope in a file prepared in advance by a statistician. The envelopes contain the treatments in random order. The police either arrest the suspect or warn and release him, depending on the contents of the envelope. The researchers then watch police records and visit the victim to see if the domestic violence reoccurs.

The first such experiment appeared to show that arresting domestic violence suspects does reduce their future violent behavior. As a result of this evidence, arrest has become the common police response to domestic violence.

The domestic violence experiments shed light on an important issue of public policy. Because there is no informed consent, the ethical rules that govern clinical trials and most social science studies would forbid these experiments.

They were cleared by review boards because, in the words of one domestic violence researcher, "These people became subjects by committing acts that allow the police to arrest them. You don't need consent to arrest someone."

SECTION 3.4 SUMMARY

- Approval of an **institutional review board** is required for studies that involve humans or animals as subjects.
- Human subjects must give **informed consent** if they are to participate in experiments.
- Data on human subjects must be kept **confidential**.

SECTION 3.4 EXERCISES

For Exercises 3.71 and 3.72, see pages 204–205; for Exercises 3.73 and 3.74, see page 207; and for Exercises 3.75 and 3.76, see page 209.

3.77 Did you give informed consent? You were asked to participate in a study by a friend who is recruiting subjects. You trust your friend and you tell her that you are willing to do whatever is needed for the study. Have you given informed consent? Explain your answer.

3.78 Are the data confidential? You have participated in a study, and the results were published in an article in a very prestigious journal. Only summary information was published. The policy of the journal requires that all data used in the articles it publishes be available to the public, and it archives the data on a website. When you examine the data, you realize that you have a unique set of characteristics that would allow someone who knows you very well to identify which data are from you. Someone who does not know you would not be able to do this. Are the data confidential? Explain your answer.

3.79 Is the IRB responsible? An institutional review board (IRB) approved an experimental cancer vaccine for use in a clinical trial. The subjects were patients who had advanced disease and had received standard treatments with no success. Of the 94 subjects who received the vaccine, 26 died during the study. Their deaths were not due to the vaccine. Some family members of the subjects sued the hospital, the study director, the company that made the vaccine, a university official, individual members of the IRB, and the university bioethicist who consulted with the IRB.[31] Discuss this case from the point of view of ethics. Include any additional information that you would need to form your opinion.

3.80 What is wrong? Explain what is wrong in each of the following scenarios.

(a) Clinical trials are always ethical as long as they randomly assign patients to the treatments.

(b) The job of an institutional review board is complete when it decides to allow a study to be conducted.

(c) A treatment that has no risk of physical harm to subjects is always ethical.

3.81 How should the samples have been analyzed? Refer to the ovarian cancer diagnostic test study in Example 3.36 (page 209). Describe how you would process the samples through the mass spectrometer.

3.82 The Vytorin controversy. Vytorin is a combination pill designed to lower cholesterol. It consists of a relatively inexpensive and widely used drug, Zocor, and a newer drug called Zetia. Early study results suggested that Vytorin was no more effective than Zetia. Critics claimed that the makers of the drugs tried to change the response variable for the study, and two congressional panels investigated why there was a two-year delay in the release of the results. Use the Internet to search for more information about this controversy and write a report about what you find. Include an evaluation in the framework of ethical use of experiments and data. A good place to start your search would be to look for the phrase "Vytorin's shortcomings."

3.83 The General Social Survey. One of the most important nongovernment surveys in the United States is the National Opinion Research Center's General Social Survey. The GSS regularly monitors public opinion on a wide variety of political and social issues. Interviews are conducted in person in the subject's home. Are a subject's responses to GSS questions anonymous, confidential, or both? Explain your answer.

3.84 Anonymity and confidentiality in health screening. Texas A&M, like many universities, offers free screening for HIV, the virus that causes AIDS. The announcement says, "Persons who sign up for the HIV Screening will be assigned a number so that they do not have to give their name." They can learn the results of the test by telephone, still without giving their name. Does this practice offer *anonymity* or just *confidentiality?*

3.85 Anonymity and confidentiality in mail surveys. Some common practices may appear to offer anonymity while actually delivering only confidentiality. Market researchers often use mail surveys that do not ask the respondent's identity but contain hidden codes on the questionnaire that identify the respondent. A false claim of anonymity is clearly unethical. If only confidentiality is promised, is it also unethical to say nothing about the identifying code, perhaps causing respondents to believe their replies are anonymous?

3.86 Use of stored blood. Long ago, doctors drew a blood specimen from you as part of treating minor anemia. Unknown to you, the sample was stored. Now researchers plan to use stored samples from you and many other people to look for genetic factors that may influence anemia. It is no longer possible to ask your consent. Modern technology can read your entire genetic makeup from the blood sample.

(a) Do you think it violates the principle of informed consent to use your blood sample if your name is on it but you were not told that it might be saved and studied later?

(b) Suppose that your identity is not attached. The blood sample is known only to come from (say) "a 20-year-old white female being treated for anemia." Is it now OK to use the sample for research?

(c) Perhaps we should use biological materials such as blood samples only from patients who have agreed to allow the material to be stored for later use in research. It isn't possible to say in advance what kind of research, so this falls short of the usual standard for informed consent. Is it nonetheless acceptable, given complete confidentiality and the fact that using the sample can't physically harm the patient?

3.87 Political polls. The presidential election campaign is in full swing, and the candidates have hired polling organizations to take regular polls to find out what the voters think about the issues. What information should the pollsters be required to give out?

AU: set as (a) and re-letter

(a) What does the standard of informed consent require the pollsters to tell potential respondents?

(b) The standards accepted by polling organizations also require giving respondents the name and address of the organization that carries out the poll. Why do you think this is required?

(c) The polling organization usually has a professional name such as "Samples Incorporated," so respondents don't know that the poll is being paid for by a political party or candidate. Would revealing the sponsor to respondents bias the poll? Should the sponsor always be announced whenever poll results are made public?

3.88 Should poll results be made public? Some people think that the law should require that all political results be made public. Otherwise, the possessors of poll results can use the information to their own advantage. They can act on the information, release only selected parts of it, or time the release for best effect. A candidate's organization replies that it is paying for the poll to gain information for its own use, not to amuse the public. Do you favor requiring complete disclosure of political poll results? What about other private surveys, such as market research surveys of consumer tastes?

3.89 Informed consent to take blood samples. Researchers from Yale, working with medical teams in Tanzania, wanted to know how common infection with the AIDS virus is among pregnant women in that country. To do this, they planned to test blood samples drawn from pregnant women.

Yale's institutional review board insisted that the researchers get the informed consent of each woman and tell her the results of the test. This is the usual procedure in developed nations. The Tanzanian government did not want to tell the women why blood was drawn or tell them the test results. The government feared panic if many people turned out to have an incurable disease for which the country's medical system could not provide care. The study was canceled. Do you think that Yale was right to apply its usual standards for protecting subjects?

CHAPTER 3 EXERCISES

3.90 Visit statistics and the news. STATS is an organization concerned about the appropriate reporting of statistical knowledge in the news media. Visit the website **stats.org/blog.** Some recent postings include discussions of deflategate, health claims for coffee, and the graph that launched a thousand news stories. Read one of the articles posted on this site and then write a short report summarizing the major ideas in the article.

3.91 Online behavioral advertising. The Federal Trade Commission Staff Report, "Self-Regulatory Principles for Online Behavioral Advertising," defines behavioral advertising as "the tracking of a consumer's online activities over time—including the searches the consumer has conducted, the Web pages visited and the content viewed—in order to deliver advertising targeted to the individual consumer's interests." The report suggests four governing concepts for their proposals. These are (1) transparency and control: when companies collect information from consumers for advertising, they should tell consumers how the data will be collected, and consumers should be given a choice about whether to allow the data to be collected; (2) security and data retention: data should be kept secure and should be retained only as long as they are needed; (3) privacy: before data are used in a way that differs from promises made when they were collected, consent should be obtained from the consumer; and (4) sensitive data: affirmative express consent should be obtained before using any sensitive data.[32] Write a report discussing your opinions concerning online behavioral advertising and the four governing concepts. Pay particular attention to issues related to the ethical collection and use of statistical data.

3.92 Confidentiality at NORC. The National Opinion Research Center conducts a large number of surveys and has established procedures for protecting the confidentiality of its survey participants. For its Survey of Consumer Finances, NORC provides a pledge to participants regarding confidentiality. This pledge is available at **scf.norc.org/confidentiality.html**. Review the pledge and summarize its key parts. Do you think that the pledge adequately addresses issues related to the ethical collection and use of data? Explain your answer.

3.93 Make it an experiment! In the following observational studies, describe changes that could be made to the data collection process that would result in an experiment rather than an observational study. Also, offer suggestions about unseen biases or lurking variables that may be present in the studies as they are described here.

(a) A friend of yours likes to play Texas hold 'em. Every time that he tells you about his playing, he says that he won.

(b) In an introductory statistics class, you notice that the students who sit in the first two rows of seats had higher scores on the first exam than the other students in the class.

3.94 Name the designs. What is the name for each of these study designs?

(a) A study to compare two methods of preserving wood started with boards of southern white pine. Each board was ripped from end to end to form two edge-matched specimens. One was assigned to Method A; the other, to Method B.

(b) A survey on youth and smoking contacted by telephone 300 smokers and 300 nonsmokers, all 14 to 22 years of age.

(c) Does air pollution induce DNA mutations in mice? Starting with 40 male and 40 female mice, 20 of each sex were housed in a polluted industrial area downwind from a steel mill. The other 20 of each sex were housed at an unpolluted rural location 30 kilometers away.

3.95 Price promotions and consumer expectations. A researcher studying the effect of price promotions on consumer expectations makes up two different histories of the store price of a hypothetical brand of laundry detergent for the past year. Students in a marketing course view one or the other price history on a computer. Some students see a steady price, while others see regular promotions that temporarily cut the price. The students are then asked what price they would expect to pay for the detergent. Is this study an experiment? Why? What are the explanatory and response variables?

3.96 Calcium and healthy bones. Adults need to eat foods or supplements that contain enough calcium to maintain healthy bones. Calcium intake is generally measured in milligrams per day (mg/d), and one measure of healthy bones is total body bone mineral density measured in grams per centimeter squared (TBBMD, g/cm^2). Suppose that you want to study the relationship between calcium intake and TBBMD.

(a) Design an observational study to study the relationship.

(b) Design an experiment to study the relationship.

(c) Compare the relative merits of your two designs. Which do you prefer? Give reasons for your answer.

3.97 Choose the type of study. Give an example of a question about pets and their owners, their behavior, or their opinions that would best be answered by

(a) a sample survey.

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

(c) an experiment.

3.98 Compare the fries. Do consumers prefer the fries from Burger King or from McDonald's? Design a blind test in which the source of the fries is not identified. Describe briefly the design of a matched pairs experiment to investigate this question. How will you use randomization?

3.99 Bicycle gears. How does the time it takes a bicycle rider to travel 100 meters depend on which gear is used and how steep the course is? It may be, for example, that higher gears are faster on level ground, but lower gears are faster on steep inclines. Discuss the design of a two-factor experiment to investigate this issue, using one bicycle with three gears and one rider. How will you use randomization?

3.100 Design an experiment. The previous two exercises illustrate the use of statistically designed experiments to answer questions that arise in everyday life. Select a question of interest to you that an experiment might answer, and carefully discuss the design of an appropriate experiment.

3.101 Design a survey. You want to investigate the attitudes of students at your school about the faculty's commitment to teaching. The student government will pay the costs of contacting about 500 students.

(a) Specify the exact population for your study; for example, will you include part-time students?

(b) Describe your sample design. Will you use a stratified sample?

(c) Briefly discuss the practical difficulties that you anticipate; for example, how will you contact the students in your sample?

3.102 Compare two doses of a drug. A drug manufacturer is studying how a new drug behaves in patients. Investigators compare two doses: 5 milligrams (mg) and 10 mg. The drug can be administered by injection, by a skin patch, or by intravenous drip. Concentration in the blood after 30 minutes (the response variable) may depend both on the dose and on the method of administration.

(a) Make a sketch that describes the treatments formed by combining dose and method. Then use a diagram to outline a completely randomized design for this two-factor experiment.

(b) "How many subjects?" is a tough issue. We will explain the basic ideas in Chapter 6. What can you say now about the advantage of using larger groups of subjects?

3.103 Would the results be different for men and women? The drug that is the subject of the experiment in Exercise 3.102 may behave differently in men and women. How would you modify your experimental design to take this into account?

3.104 Informed consent. The requirement that human subjects give their informed consent to participate in an experiment can greatly reduce the number of available subjects. For example, a study of new teaching methods asks the consent of parents for their children to be randomly assigned to be taught by either a new method or the standard method. Many parents do not return the forms, so their children must continue to be taught by the standard method. Why is it not correct to consider these children as part of the control group along with children who are randomly assigned to the standard method?

3.105 Two ways to ask sensitive questions. Sample survey questions are usually read from a computer screen. In a Computer-aided personal interview (CAPI), the interviewer reads the questions and enters the responses. In a Computer-aided self interview (CASI), the interviewer stands aside and the respondent reads the questions and enters responses. One method almost always shows a higher percent of subjects admitting use of illegal drugs. Which method? Explain why.

3.106 Your institutional review board. Your college or university has an institutional review board that screens all studies that use human subjects. Get a copy of the document that describes this board (you can probably find it online).

(a) According to this document, what are the duties of the board?

(b) How are members of the board chosen? How many members are not scientists? How many members are not employees of the college? Do these members have some special expertise, or are they simply members of the "general public"?

3.107 Use of data produced by the government. Data produced by the government are often available free or at low cost to private users. For example, satellite weather data produced by the U.S. National Weather Service are available free to TV stations for their weather reports and to anyone on the Internet. *Opinion 1:* Government data should be available to everyone at minimal cost. *Opinion 2:* The satellites are expensive, and the TV stations are making a profit from their weather services, so they should share the cost. European governments, for example, charge TV stations for weather data. Which opinion do you support, and why?

3.108 Should we ask for the consent of the parents? The Centers for Disease Control and Prevention, in a survey of teenagers, asked the subjects if they were sexually active. Those who said Yes were then asked, "How old were you when you had sexual intercourse for the first time?" Should consent of parents be required to ask minors about sex, drugs, and other such issues, or is consent of the minors themselves enough? Give reasons for your opinion.

3.109 A theft experiment. Students sign up to be subjects in a psychology experiment. When they arrive, they are told that interviews are running late and are taken to a waiting room. The experimenters then stage a theft of a valuable object left in the waiting room. Some subjects are alone with the thief, and others are in pairs—these are the treatments being compared. Will the subject report the theft? The students had agreed to take part in an unspecified study, and the true nature of the experiment is explained to them afterward. Do you think this study is ethically OK?

3.110 A cheating experiment. A psychologist conducts the following experiment. She measures the attitude of subjects toward cheating and then has them play a game rigged so that winning without cheating is impossible. The computer that organizes the game also records—unknown to the subjects—whether or not they cheat. Then attitude toward cheating is retested. Subjects who cheat tend to change their attitudes to find cheating more acceptable. Those who resist the temptation to cheat tend to condemn cheating more strongly on the second test of attitude. These results confirm the psychologist's theory. This experiment tempts subjects to cheat. The subjects are led to believe that they can cheat secretly when, in fact, they are observed. Is this experiment ethically objectionable? Explain your position.