

Producing Data



A magazine article says that men need Pilates exercise more than women. Read the Introduction to learn more.

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 we find in data can arise from many sources.*

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. The article describes the benefits that two men received from taking Pilates classes. A newspaper ad states that a particular brand of windows are “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 quantitative data. However, this type of data does not give us a sound basis for drawing conclusions.

- 3.1 Design of Experiments
- 3.2 Sampling Design
- 3.3 Toward Statistical Inference
- 3.4 Ethics

exploratory data analysis



ANECDOTAL EVIDENCE

Anecdotal evidence is based on haphazardly selected individual cases, which often come to our attention because they are striking in some way. These cases need not be representative of any larger group of cases.

USE YOUR KNOWLEDGE

- 3.1 Final Fu.** Your friends are big fans of “Final Fu,” MTV2’s martial arts competition. To what extent do you think you can generalize your preferences for this show to all students at your college?
- 3.2 Describe an example.** Find an example from some recent experience where anecdotal evidence is used to draw a conclusion that is not justified. Describe the example and explain why it cannot be used in this way.
- 3.3 Preference for Jolt Cola.** Jamie is a hard-core computer programmer. He and all his friends prefer Jolt Cola (caffeine equivalent to two cups of coffee) to either Coke or Pepsi (caffeine equivalent to less than one cup of coffee).¹ Explain why Jamie’s experience is not good evidence that most young people prefer Jolt to Coke or Pepsi.
- 3.4 Automobile seat belts.** When the discussion turns to the pros and cons of wearing automobile seat belts, Herman always brings up the case of a friend who survived an accident because he was not wearing a seat belt. The friend was thrown out of the car and landed on a grassy bank, suffering only minor injuries, while the car burst into flames and was destroyed. Explain briefly why this anecdote does not provide good evidence that it is safer not to wear seat belts.

Available 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 in the past for some other purpose but that may help answer a present question.

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. However, the clearest answers to present questions often require that data be produced to answer those specific questions. Here are two examples:

EXAMPLE

3.1 Causes of death. If you visit the National Center for Health Statistics Web site, www.cdc.gov/nchs, you will learn that accidents are the most common cause of death among people aged 20 to 24, accounting for over 40% of all deaths. Homicide is next, followed by suicide. AIDS ranks seventh, behind heart disease and cancer, at 1% of all deaths. The data also show that it is dangerous to be a young man: the overall death rate for men aged 20 to 24 is three times that for women, and the death rate from homicide is more than five times higher among men.

EXAMPLE

3.2 Math skills of children. At the Web site of the National Center for Education Statistics, nces.ed.gov/nationsreportcard/mathematics, you will find full details about the math skills of schoolchildren in the latest National Assessment of Educational Progress (Figure 3.1). Mathematics scores have slowly but steadily increased since 1990. All racial/ethnic groups, both men and women, and students in most states are getting better in math.

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

USE YOUR KNOWLEDGE

3.5 Find some available data. Visit the Internet and find an example of available data that is interesting to you. Explain how the data were collected and what questions the study was designed to answer.

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 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

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.

The screenshot shows the homepage of the National Center for Education Statistics (NCES) for the National Assessment of Educational Progress (NAEP) Mathematics. The browser window title is "NAEP - Scheduled NAEP Mathematics Assessments, Past Results, Trends, Methods - Mozilla Firefox". The URL is "http://nces.ed.gov/nationsreportcard/mathematics/". The page has a blue header with the NCES logo and navigation links. The main content area is titled "Mathematics" and includes a search bar. Below the header, there are several sections: "2005 Mathematics Results" with links to download report cards for grades 4, 8, and 12; "Spotlight on the 2005 State Results" featuring a map of the United States; "How the NAEP Mathematics Assessment Works" with a list of links explaining the assessment process; and "Using NAEP Data" with a link to explore data. The page also includes a "Long-Term Trends" section and a "Spotlight on the Trial Urban District Assessment Results" section.

FIGURE 3.1 The Web sites of government statistical offices are prime sources of data. Here is the home page of the National Assessment of Educational Progress.

EXAMPLE

3.3 The General Social Survey. One of the most important sample surveys is the General Social Survey (GSS) conducted by the NORC, a national organization for research and computing affiliated with the University of Chicago.² The GSS interviews about 3000 adult residents of the United States every second 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 pro-

duced by sampling a population of people or things. Opinion polls, for example, 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 Find a sample survey. Use the Internet or some printed material to find an example of a sample survey that interests you. Describe the population, how the sample was collected, and some of the conclusions.

In all of 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 entire 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 inaccurate. Bored people do not count carefully.

If conclusions based on a sample are to be valid for the entire population, a sound design for selecting the sample is required. Sampling designs are the topic of Section 3.2.

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 treatment on individuals and we observe their responses.

USE YOUR KNOWLEDGE

3.7 Cell phones and brain cancer. One study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same sex, age, and race who did not have brain cancer, then asked about use of cell phones.³ Result: "Our data suggest that use of handheld cellular telephones is not associated with risk of brain cancer." Is this an observational study or an experiment? Why? What are the explanatory and response variables?

3.8 Violent acts on prime-time TV. A typical hour of prime-time television shows three to five violent acts. Linking family interviews and police records shows a clear association between time spent watching TV as a child and later aggressive behavior.⁴

- (a) Explain why this is an observational study rather than an experiment. What are the explanatory and response variables?
- (b) Suggest several lurking variables describing a child's home life that may be related to how much TV he or she watches. Explain why this makes it difficult to conclude that more TV *causes* more aggressive behavior.

intervention

An observational study, even one based on a statistical 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.

EXAMPLE

3.4 Child care and behavior. A study of child care enrolled 1364 infants in 1991 and planned to follow them through their sixth year in school. Twelve years later, the researchers published an article finding that “the more time children spent in child care from birth to age four-and-a-half, the more adults tended to rate them, both at age four-and-a-half and at kindergarten, as less likely to get along with others, as more assertive, as disobedient, and as aggressive.”⁵

What can we conclude from this study? If parents choose to use child care, are they more likely to see these undesirable behaviors in their children?

EXAMPLE

3.5 Is there a cause and effect relationship? Example 3.4 describes an observational study. Parents made all child care decisions and the study did not attempt to influence them. A summary of the study stated, “The study authors noted that their study was not designed to prove a cause and effect relationship. That is, the study cannot prove whether spending more time in child care causes children to have more problem behaviors.”⁶ Perhaps employed parents who use child care are under stress and the children react to their parents' stress. Perhaps single parents are more likely to use child care. Perhaps parents are more likely to place in child care children who already have behavior problems.

We can imagine an experiment that would remove these difficulties. From a large group of young children, choose some to be placed in child care and others to remain at home. This is an experiment because the treatment (child care or not) is imposed on the children. Of course, this particular experiment is neither practical nor ethical.

confounded

In Examples 3.4 and 3.5, we say that the effect of child care on behavior is **confounded** with (mixed up with) other characteristics of families who use child care. 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

3.6 A dietary behavior experiment. An experiment was designed to examine the effect of a 30-minute instructional session in a Food Stamp office on the dietary behavior of low-income women.⁷ A group of women were randomly assigned to either the instructional session or no instruction. Two months later, data were collected on several measures of their behavior.

Experiments usually require some sort of randomization, as in this example. We begin the discussion of statistical designs for data collection in Section 3.1 with the principles underlying the design of experiments.

USE YOUR KNOWLEDGE

3.9 Software for teaching biology. An educational software company wants to compare the effectiveness of its computer animation for teaching cell biology with that of a textbook presentation. The company tests the biological knowledge of each of a group of first-year college students, 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 understanding of cell biology in the two groups. Is this an experiment? Why or why not? What are the explanatory and response variables?

3.10 Find an experiment. Use the Internet or some printed material to find an example of an experiment that interests you. Describe how the experiment was conducted and some of the conclusions.

statistical inference

Statistical techniques for producing data are the foundation for formal **statistical inference**, which answers specific questions with a known degree of confidence. In Section 3.3, we discuss some basic ideas related to inference.

ethics

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.

3.1 Design of Experiments

A study is an experiment when 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, TREATMENT

The individuals on which the experiment is done are the **experimental units**. When the units are human beings, they are called **subjects**. A specific experimental condition applied to the units is called a **treatment**.

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

factors

level of a factor



EXAMPLE

3.7 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 and 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 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 standard tests, were less likely to fail a grade, had better high school grades, and so on. The benefits of small classes were greatest for minority students.⁸

Example 3.7 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 fam-

ily 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

3.8 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 1, 3, or 5 times during the program.

This experiment has two *factors*: length of the commercial, with 2 levels, and repetitions, with 3 levels. The 6 combinations of one level of each factor form 6 *treatments*. Figure 3.2 shows the layout of the treatments. After viewing, all of the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it. These are the *response variables*.⁹

		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, for Example 3.8. Combining the levels of the two factors forms six treatments.

Example 3.8 shows how experiments allow us to study the combined effects of several factors. The interaction of several factors can produce effects that could not 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 3.8 will help us find out.

USE YOUR KNOWLEDGE

3.11 Food for a trip to the moon. Storing food for long periods of time is a major challenge for those planning for human space travel beyond the moon. One problem is that exposure to radiation decreases the length of time that food can be stored. One experiment examined the effects of nine different levels of radiation on a particular type of fat, or lipid.¹⁰ The amount of oxidation of the lipid is the measure of the extent of the damage due to the radiation. Three samples are exposed

to each radiation level. Give the experimental units, the treatments, and the response variable. Describe the factor and its levels. There are many different types of lipids. To what extent do you think the results of this experiment can be generalized to other lipids?

3.12 Learning how to draw. A course in computer graphics technology requires students to learn multiview drawing concepts. This topic is traditionally taught using supplementary material printed on paper. The instructor of the course believes that a Web-based interactive drawing program will be more effective in increasing the drawing skills of the students.¹¹ The 50 students who are enrolled in the course will be randomly assigned to either the paper-based instruction or the Web-based instruction. A standardized drawing test will be given before and after the instruction. Explain why this study is an experiment and give the experimental units, the treatments, and the response variable. Describe the factor and its levels. To what extent do you think the results of this experiment can be generalized to other settings?

Comparative experiments

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

Treatment → **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 in the field 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. A medical example will show what can go wrong.

EXAMPLE

3.9 Gastric freezing. “Gastric freezing” is a clever treatment for ulcers in the upper intestine. The patient swallows a deflated balloon with tubes attached, then a refrigerated liquid is pumped through the balloon for an hour. The idea is that cooling the stomach will reduce its production of acid and so relieve ulcers. An experiment reported in the *Journal of the American Medical Association* showed that gastric freezing did reduce acid production and relieve ulcer pain. The treatment was safe and easy and was widely used for several years. The design of the experiment was

Gastric freezing → **Observe pain relief**

placebo effect

The gastric freezing experiment was poorly designed. The patients’ response may have been due to the **placebo effect**. A placebo is a dummy treatment. Many patients respond favorably to any treatment, even a placebo. This may be due to trust in the doctor and expectations of a cure or simply to the fact that medical conditions often improve without treatment. The response to a dummy treatment is the placebo effect.

A later experiment divided ulcer patients into two groups. One group was treated by gastric freezing as before. The other group received a placebo treatment in which the liquid in the balloon was at body temperature rather than freezing. The results: 34% of the 82 patients in the treatment group improved, but so did 38% of the 78 patients in the placebo group. This and other properly designed experiments showed that gastric freezing was no better than a placebo, and its use was abandoned.¹²

control group



The first gastric freezing experiment gave misleading results because the effects of the explanatory variable were confounded with the placebo effect. We can defeat confounding by *comparing* two groups of patients, as in the second gastric freezing experiment. The placebo effect and other lurking variables now operate on both groups. The only difference between the groups is the actual effect of gastric freezing. The group of patients who received a sham treatment is called a **control group**, because it enables us to control the effects of outside variables on the outcome. Control is the first basic principle of statistical design of experiments. Comparison of several treatments in the same environment is the simplest form of control.

Uncontrolled experiments 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. It should not surprise you to learn that uncontrolled studies in medicine give new therapies a much higher success rate than proper comparative experiments. Well-designed experiments usually compare several treatments.

USE YOUR KNOWLEDGE

3.13 Does using statistical software improve exam scores? An instructor in an elementary statistics course wants to know if using a new statistical software package will improve students' final-exam scores. He asks for volunteers and about half of the class agrees to work with the new software. He compares the final-exam scores of the students who used the new software with the scores of those who did not. Discuss possible sources of bias in this study.

Randomization

experiment design

The **design of an experiment** first describes the response variable or variables, the factors (explanatory variables), and the layout of the treatments, with comparison as the leading principle. Figure 3.2 illustrates this aspect of

the design of a study of response to advertising. The second aspect of design is the rule used to assign the experimental units 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. Systematic differences among the groups of experimental units in a comparative experiment cause bias. How can we assign experimental units to treatments in a way that is fair to all of the treatments?

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 an experimenter might bias the study by, for example, assigning more advanced cancer cases to a promising new treatment in the unconscious hope that it will 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 design creates groups by chance alone. Here is an example.


EXAMPLE

3.10 Cell phones and driving. Does talking on a hands-free cell phone distract drivers? Undergraduate students “drove” in a high-fidelity driving simulator equipped with a hands-free cell phone. The car ahead brakes: how quickly does the subject respond? Twenty students (the control group) simply drove. Another 20 (the experimental group) talked on the cell phone while driving.

This experiment has a single factor (cell phone use) 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 form the experimental group and the remaining 20 make up the control group. Figure 3.3 outlines the design of this experiment.¹³

The use of chance to divide experimental units into groups is called **randomization**. The design in Figure 3.3 combines comparison and ran-

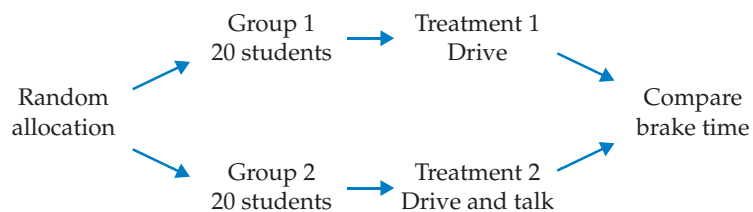


FIGURE 3.3 Outline of a randomized comparative experiment, for Example 3.10.

domization 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 generally using treatment groups about equal in size.

USE YOUR KNOWLEDGE

3.14 Diagram the drawing experiment. Refer to Exercise 3.12 (page 180). Draw a diagram similar to Figure 3.3 that describes the computer graphics drawing experiment.

3.15 Diagram the food for Mars experiment. Refer to Exercise 3.11 (page 179). Draw a diagram similar to Figure 3.3 that describes the food for space travel experiment.

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 cell phone operate equally on both groups.
- Therefore, differences in average brake reaction time must be due either to talking on the cell phone or to the play of chance in the random assignment of subjects to the two groups.

That “either-or” deserves more comment. We cannot say that *any* difference in the average reaction times of the experimental and control groups is caused by talking on the cell phone. There would be some difference even if both groups were treated the same, because the natural variability among people means that some react faster than others. Chance can assign the faster-reacting students to one group or the other, 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 group got lucky and received the subject with quicker reactions. If we assign many students to each group, however, the effects of chance will average out. There will be little difference in the average reaction times in the two groups unless talking on the cell phone 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 impersonal chance to assign experimental units to treatments.

3. Repeat each treatment on many units to reduce chance variation in the results.

We hope to see a difference in the responses so large that it is unlikely to happen just because of chance variation. We can use the laws of probability, which give a mathematical description of chance behavior, to learn if the treatment effects are larger than we would expect to see if only chance were operating. If they are, we call them *statistically significant*.

STATISTICAL SIGNIFICANCE

An observed effect so large that it would rarely occur by chance is called **statistically significant**.

You will often see the phrase “statistically significant” in reports of investigations in many fields of study. It tells you that the investigators found good evidence for the effect they were seeking. The cell phone study, for example, reported statistically significant evidence that talking on a cell phone increases the mean reaction time of drivers when the car in front of them brakes.

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. Most statistical software will choose 20 out of a list of 40 at random, for example. The list might contain the names of 40 human subjects. The 20 chosen form one group, and the 20 that remain form the second group. The *Simple Random Sample* applet on the text CD and Web site makes it particularly easy to choose treatment groups at random.

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 and on the back endpaper 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 saves 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 the properties 1 and 2 described above.

Our goal is to use random digits for experimental randomization. We need the following facts about random digits, which are consequences of the basic 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

3.11 Randomize the students. In the cell phone experiment of Example 3.10, we must divide 40 students at random into two groups of 20 students each.

Step 1: Label. Give each student a numerical label, using as few digits as possible. Two digits are needed to label 40 students, so we use labels

01, 02, 03, ..., 39, 40

It is also correct to use labels 00 to 39 or some other choice of 40 two-digit labels.

Step 2: Table. Start anywhere in Table B and read two-digit groups. Suppose we begin at line 130, which is

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups in this line are

69 05 16 48 17 87 17 40 95 17

Each of these two-digit groups is a label. The labels 00 and 41 to 99 are not used in this example, so we ignore them. The first 20 labels between 01 and 40 that we encounter in the table choose students for the experimental group. Of the first 10 labels in line 130, we ignore four because they are too high (over 40). The others are 05, 16, 17, 17, 40, and 17. The students labeled 05, 16, 17, and 40 go into the experimental group. Ignore the second and third 17s because that student is already in the group. Run your finger across line 130 (and continue to the following lines) until you have chosen 20 students. They are the students labeled

05, 16, 17, 40, 20, 19, 32, 04, 25, 29,
37, 39, 31, 18, 07, 13, 33, 02, 36, 23

You should check at least the first few of these. These students form the experimental group. The remaining 20 are the control group.

As Example 3.11 illustrates, randomization requires two steps: assign labels to the experimental units and then use Table B to select labels at random. Be sure that all labels are the same length so that all have the same chance to be chosen. Use the shortest possible labels—one digit for 9 or fewer individuals, two digits for 10 to 100 individuals, and so on. Don't try to scramble the labels as you assign them. Table B will do the required randomizing, so assign labels in any convenient manner, such as in alphabetical order for human subjects. You can read digits from Table B in any order—along a row, down a column, and so on—because the table has no order. As an easy standard practice, we recommend reading along rows.

It is easy to use statistical software or Excel to randomize. Here are the steps:

Step 1: Label. The first step, assigning labels to the experimental units, is similar to the procedure we described above. One difference, however, is that we are not restricted to using numerical labels. Any system where each experimental unit has a unique label identifier will work.

Step 2: Use the computer. Once we have the labels, we then create a data file with the labels and generate a random number for each label. In Excel, this can be done with the RAND() function. Finally, we sort the entire data set based on the random numbers. Groups are formed by selecting units in order from the sorted list.

This process is essentially the same as writing the labels on a deck of cards, shuffling the cards, and dealing them out one at a time.

EXAMPLE

3.12 Using software for randomization. Let's do a randomization similar to the one we did in Example 3.11, but this time using Excel. Here we will use 10 experimental units. We will assign 5 to the treatment group and 5 to the control group. We 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 5 labels (8, 5, 9, 4, and 6) are assigned to the experimental group. The remaining 5 labels (3, 10, 7, 2, and 1) correspond to the control group.

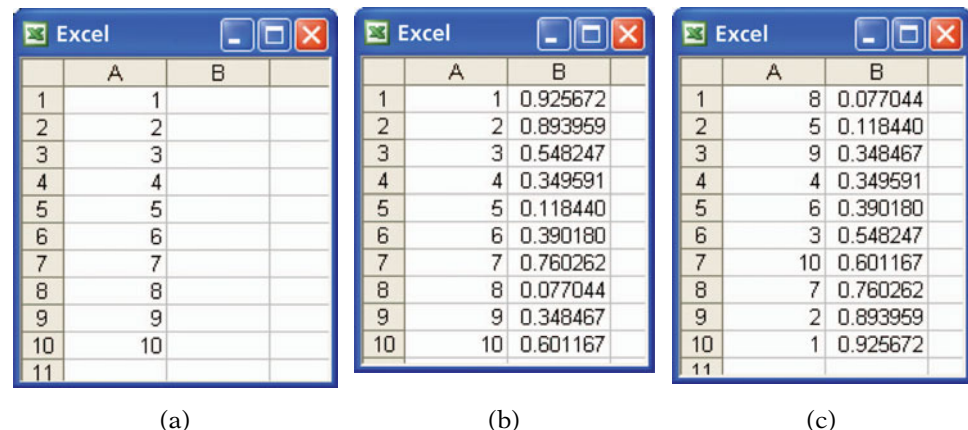


FIGURE 3.4 Randomization of 10 experimental units using a computer, for Example 3.12. (a) Labels. (b) Random numbers. (c) Sorted list of labels.

completely randomized design

When all experimental units are allocated at random among all treatments, as in Example 3.11, 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

3.13 Randomization of the TV commercial experiment. Figure 3.2 (page 179) 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.

To carry out the random assignment, label the 150 students 001 to 150. (Three digits are needed to label 150 subjects.) 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 3 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 hand. We recommend the *Simple Random Sample* applet. Exercise 3.35 shows how to use the applet to do the randomization for this example.

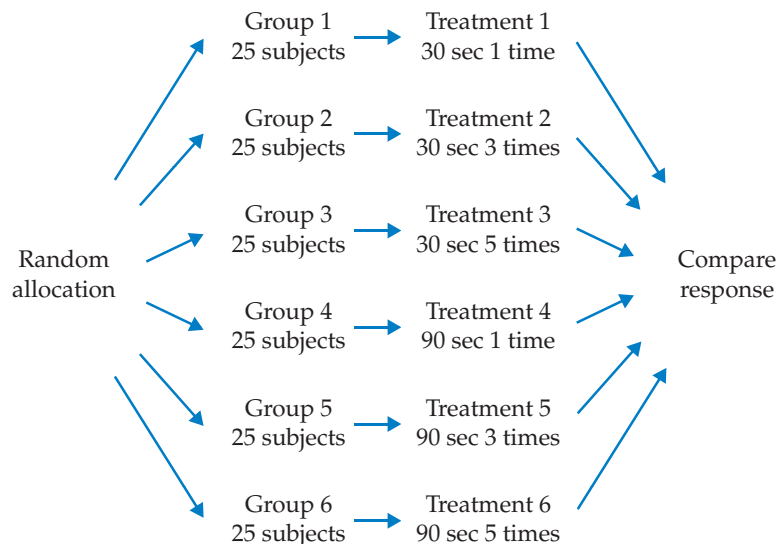


FIGURE 3.5 Outline of a completely randomized design comparing six treatments, for Example 3.13.

USE YOUR KNOWLEDGE

3.16 Do the randomization. Use computer software to carry out the randomization in Example 3.13.

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. For example, the subjects in the second gastric freezing experiment (Example 3.9) all got the same medical attention during the study. Moreover, the study was **double-blind**—neither the subjects themselves nor the medical personnel who worked with them knew which treatment any subject had 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.

double-blind



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

3.14 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.¹⁴

EXAMPLE

3.15 Knock out genes. To study genetic influence on behavior, experimenters “knock out” a gene in one group of mice and compare their behavior with that of a control group of normal mice. The results of these experiments often don't agree as well as hoped, so investigators did exactly the same experiment with the same genetic strain of mice in Oregon, Alberta (Canada), and New York. Many results were very different.¹⁵ It appears that small differences in the lab environments have big effects on the behavior of the mice. Remember this the next time you read that our genes control our behavior.

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? 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 cell phone experiment. Example 3.10 describes an experiment on the effects of talking on a cell phone while driving. The experiment compared two treatments, driving in a simulator and driving in the simulator while talking on a hands-free cell phone. The response variable is the time the driver takes to apply the brake when the car in front brakes suddenly. In Example 3.10, 40 student subjects were assigned at random, 20 students to each treatment. This is a completely randomized design, outlined in Figure 3.3. Subjects differ in driving skill and reaction times. The completely randomized design relies on chance to create two similar groups of subjects.

In fact, the experimenters used a matched pairs design in which all subjects drove both with and without using the cell phone. They compared each individual's reaction times with and without the phone. If all subjects drove first with the phone and then without it, the effect of talking on the cell phone would be confounded with the fact that this is the first run in the simulator. The proper procedure requires that all subjects first be trained in using the simulator, that the *order* in which a subject drives with and without the phone be random, and that the two drives be on separate days to reduce the chance that the results of the second treatment will be influenced by the first treatment.

The completely randomized design uses chance to decide which 20 subjects will drive with the cell phone. The other 20 drive without it. The matched pairs design uses chance to decide which 20 subjects will drive first with and then without the cell phone. The other 20 drive first without and then with the phone.

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 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 therefore 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.



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.

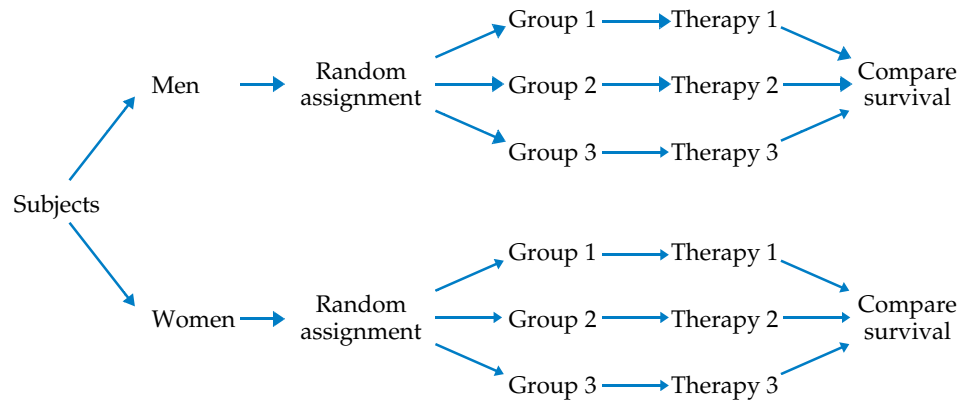


FIGURE 3.6 Outline of a block design, for Example 3.18. The blocks consist of male and female subjects. The treatments are three therapies for cancer.

EXAMPLE

3.20 Blocking in an education experiment. The Tennessee STAR class size experiment (Example 3.7) 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 3.1 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**.

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 **control**, **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 3.1 Exercises

For Exercises 3.1 to 3.4, see page 172; for Exercise 3.5, see page 173; for Exercises 3.6 and 3.7, see page 175; for Exercise 3.8, see page 176; for Exercises 3.9 and 3.10, see page 177; for Exercises 3.11 and 3.12, see pages 179 and 180; for Exercise 3.13, see page 181; for Exercises 3.14 and 3.15, see page 183; and for Exercise 3.16, see page 187.

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

(a) Two forms of a lab exercise are to be compared. There are 10 rows in the classroom. Students who sit in the first 5 rows of the class are given the first form, and students who sit in the last 5 rows are given the second form.

(b) The effectiveness of a leadership program for high school students is evaluated by examining the change in scores on a standardized test of leadership skills.

(c) An innovative method for teaching introductory biology courses is examined by using the traditional method in the fall zoology course and the new method in the spring botany course.

3.18 What is wrong? Explain what is wrong with each of the following randomization procedures and

describe how you would do the randomization correctly.

(a) A list of 50 subjects is entered into a computer file and then sorted by last name. The subjects are assigned to five treatments by taking the first 10 subjects for Treatment 1, the next 10 subjects for Treatment 2, and so forth.

(b) Eight subjects are to be assigned to two treatments, four 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 80 rats to four different treatment conditions. The rats arrive from the supplier in batches of 20 and the treatment lasts two weeks. The first batch of 20 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 20 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.


3.19 Evaluate a new teaching method. A teaching innovation is to be evaluated by randomly assigning students to either the traditional approach or the new approach. The change in a standardized

test score is the response variable. Explain how this experiment should be done in a double-blind fashion.

3.20 Can you change attitudes toward binge drinking?

A experiment designed to change attitudes about binge drinking is to be performed using college students as subjects. Discuss some variables that you might use if you were to use a block design for this experiment.

3.21 Compost tea. Compost tea is rich in micro-organisms that help plants grow. It is made by soaking compost in water.¹⁶ Design a comparative experiment that will provide evidence about whether or not compost tea works for a particular type of plant that interests you. Be sure to provide all details regarding your experiment, including the response variable or variables that you will measure.

3.22  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.¹⁷ 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 of the details that they will need to evaluate your proposal.

For each of the experimental situations described in Exercises 3.23 to 3.25, identify the experimental units or subjects, the factors, the treatments, and the response variables.

3.23 How well do pine trees grow in shade? Ability to grow in shade may help pines in the dry forests of Arizona resist drought. How well do these pines grow in shade? Investigators planted pine seedlings in a greenhouse in either full light or light reduced to 5% of normal by shade cloth. At the end of the study, they dried the young trees and weighed them.

3.24 Will the students do more exercise and eat better? Most American adolescents don't eat well and don't exercise enough. Can middle schools increase physical activity among their students? Can they persuade students to eat better? Investigators designed a "physical activity intervention" to increase activity in physical education classes and during leisure periods throughout the school day.

They also designed a "nutrition intervention" that improved school lunches and offered ideas for healthy home-packed lunches. Each participating school was randomly assigned to one of the interventions, both interventions, or no intervention. The investigators observed physical activity and lunchtime consumption of fat.

3.25 Refusals in telephone surveys. How can we reduce the rate of refusals in telephone surveys? Most people who answer at all listen to the interviewer's introductory remarks and then decide whether to continue. One study made telephone calls to randomly selected households to ask opinions about the next election. In some calls, the interviewer gave her name, in others she identified the university she was representing, and in still others she identified both herself and the university. For each type of call, the interviewer either did or did not offer to send a copy of the final survey results to the person interviewed. Do these differences in the introduction affect whether the interview is completed?

3.26 Does aspirin prevent strokes and heart attacks? The Bayer Aspirin Web site claims that "Nearly five decades of research now link aspirin to the prevention of stroke and heart attacks." The most important evidence for this claim comes from the Physicians' Health Study, a large medical experiment involving 22,000 male physicians. One group of about 11,000 physicians took an aspirin every second day, while the rest took a placebo. After several years the study found that subjects in the aspirin group had significantly fewer heart attacks than subjects in the placebo group.

(a) Identify the experimental subjects, the factor and its levels, and the response variable in the Physicians' Health Study.

(b) Use a diagram to outline a completely randomized design for the Physicians' Health Study.

(c) What does it mean to say that the aspirin group had "significantly fewer heart attacks"?

3.27 Chronic tension headaches. Doctors identify "chronic tension-type headaches" as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone? Investigators compared four treatments: antidepressant alone, placebo alone, antidepressant plus stress management, and placebo plus stress management. Outline the design of the

experiment. The headache sufferers named below have agreed to participate in the study. Use software or Table B at line 151 to randomly assign the subjects to the treatments.

Anderson	Archberger	Bezawada	Cetin	Cheng
Chronopoulou	Codrington	Daggy	Daye	Engelbrecht
Guha	Hatfield	Hua	Kim	Kumar
Leaf	Li	Lipka	Lu	Martin
Mehta	Mi	Nolan	Olbricht	Park
Paul	Rau	Saygin	Shu	Tang
Towers	Tyner	Vassilev	Wang	Watkins
Xu				

3.28 Smoking marijuana and willingness to work.

How does smoking marijuana affect willingness to work? Canadian researchers persuaded people who used marijuana to live for 98 days in a “planned environment.” The subjects earned money by weaving belts. They used their earnings to pay for meals and other consumption and could keep any money left over. One group smoked two potent marijuana cigarettes every evening. The other group smoked two weak marijuana cigarettes. All subjects could buy more cigarettes but were given strong or weak cigarettes, depending on their group. Did the weak and strong groups differ in work output and earnings?¹⁸

- Outline the design of this experiment.
- Here are the names of the 20 subjects. Use software or Table B at line 101 to carry out the randomization your design requires.

Becker	Brifcani	Chen	Crabill	Cunningham
Dicklin	Fein	Gorman	Knapp	Lucas
McCarty	Merkulyeva	Mitchell	Ponder	Roe
Saeed	Seele	Truong	Wayman	Woodley

3.29 Eye cataracts. Eye cataracts are responsible for over 40% of blindness worldwide. Can drinking tea regularly slow the growth of cataracts? We can’t experiment on people, so we use rats as subjects. Researchers injected 21 young rats with a substance that causes cataracts. One group of the rats also received black tea extract; a second group received green tea extract; and a third got a placebo, a substance with no effect on the body. The response variable was the growth of cataracts over the next six weeks. Yes, both tea extracts did slow cataract growth.¹⁹

- Outline the design of this experiment.
- Use software or Table B, starting at line 120, to assign rats to treatments.

3.30 Guilt among workers who survive a layoff.

Workers who survive a layoff of other employees at their location may suffer from “survivor guilt.” A study of survivor guilt and its effects used as subjects 90 students who were offered an opportunity to earn extra course credit by doing proofreading. Each subject worked in the same cubicle as another student, who was an accomplice of the experimenters. At a break midway through the work, one of three things happened:

Treatment 1: The accomplice was told to leave; it was explained that this was because she performed poorly.

Treatment 2: It was explained that unforeseen circumstances meant there was only enough work for one person. By “chance,” the accomplice was chosen to be laid off.

Treatment 3: Both students continued to work after the break.

The subjects’ work performance after the break was compared with performance before the break.²⁰

- Outline the design of this completely randomized experiment.
- If you are using software, randomly assign the 90 students to the treatments. If not, use Table B at line 153 to choose the first four subjects for Treatment 1.

3.31 Diagram the exercise and eating experiment.


Twenty-four public middle schools agree to participate in the experiment described in Exercise 3.24. Use a diagram to outline a completely randomized design for this experiment. Then do the randomization required to assign schools to treatments. If you use Table B, start at line 160.


3.32 Price cuts on athletic shoes. Stores advertise price reductions to attract customers. What type of price cut is most attractive? Market researchers prepared ads for athletic shoes announcing different levels of discounts (20%, 40%, 60%, or 80%). The student subjects who read the ads were also given “inside information” about the fraction of shoes on sale (25%, 50%, 75%, or 100%). Each subject then rated the attractiveness of the sale on a scale of 1 to 7.²¹

- There are two factors. Make a sketch like Figure 3.2 (page 179) that displays the treatments formed by all combinations of levels of the factors.
- Outline a completely randomized design using 96 student subjects. Use software or Table B at line 111 to choose the subjects for the first treatment.


3.33 Treatment of clothing fabrics. A maker of fabric for clothing is setting up a new line to “finish” the raw fabric. The line will use either metal rollers or natural-bristle rollers to raise the surface of the fabric; a dyeing cycle time of either 30 minutes or 40 minutes; and a temperature of either 150° or 175° Celsius. An experiment will compare all combinations of these choices. Four specimens of fabric will be subjected to each treatment and scored for quality.

- (a) What are the factors and the treatments? How many individuals (fabric specimens) does the experiment require?
- (b) Outline a completely randomized design for this experiment. (You need not actually do the randomization.)

3.34  **Use the simple random sample applet.** You can use the *Simple Random Sample* applet to choose a treatment group at random once you have labeled the subjects. Example 3.11 (page 185) uses Table B to choose 20 students from a group of 40 for the treatment group in a study of the effect of cell phones on driving. Use the applet to choose the 20 students for the experimental group. Which students did you choose? The remaining 20 students make up the control group.

3.35  **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 3.13 (page 187) 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 that 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?

3.36  **Effectiveness of price discounts.** Experiments with more than one factor allow insight into interactions between the factors. A study

of the attractiveness of advertised price discounts had two factors: percent of all goods on sale (25%, 50%, 75%, or 100%) and whether the discount was stated precisely as 60% off or as a range, 50% to 70% off. Subjects rated the attractiveness of the sale on a scale of 1 to 7. Figure 3.7 shows the mean ratings for the eight treatments formed from the two factors.²² Based on these results, write a careful description of how percent on sale and precise discount versus range of discounts influence the attractiveness of a sale.

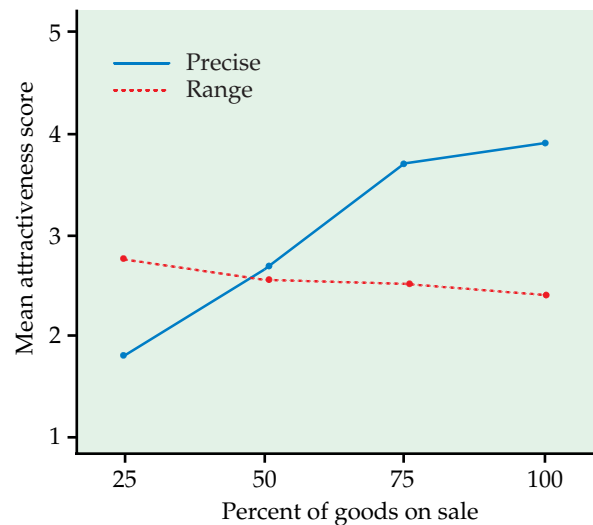



FIGURE 3.7 Mean responses to eight treatments in an experiment with two factors, showing interaction between the factors, for Exercise 3.36.

3.37  **Health benefits of bee pollen.** “Bee pollen is effective for combating fatigue, depression, cancer, and colon disorders.” So says a Web site 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.

1. Find 400 women who do not have colon disorders. Assign 200 to take bee pollen capsules and the other 200 to take placebo capsules that are identical in appearance. Follow both groups for 5 years.
2. 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 5 years.

3.38 Treatment of pain for cancer patients. Health care providers are giving more attention to relieving the pain of cancer patients. An article in the journal

Cancer surveyed a number of studies and concluded that controlled-release morphine tablets, which release the painkiller gradually over time, are more effective than giving standard morphine when the patient needs it.²³ The “methods” section of the article begins: “Only those published studies that were controlled (i.e., randomized, double blind, and comparative), repeated-dose studies with CR morphine tablets in cancer pain patients were considered for this review.” Explain the terms in parentheses to someone who knows nothing about medical trials.

3.39 Saint-John’s-wort and depression. Does the herb Saint-John’s-wort relieve major depression? Here are some excerpts from the report of a study of this issue.²⁴ The study concluded that the herb is no more effective than a placebo.

(a) “Design: Randomized, double-blind, placebo-controlled clinical trial...” Explain the meaning of each of the terms in this description.

(b) “Participants ... were randomly assigned to receive either Saint-John’s-wort extract ($n = 98$) or placebo ($n = 102$). ... The primary outcome measure was the rate of change in the Hamilton Rating Scale for Depression over the treatment period.” Based on this information, use a diagram to outline the design of this clinical trial.

3.40 The Monday effect on stock prices. Puzzling but true: stocks tend to go down on Mondays. There is no convincing explanation for this fact. A recent study looked at this “Monday effect” in more detail, using data on the daily returns of stocks on several U.S. exchanges over a 30-year period. Here are some of the findings:

*To summarize, our results indicate that the well-known Monday effect is caused largely by the Mondays of the last two weeks of the month. The mean Monday return of the first three weeks of the month is, in general, not significantly different from zero and is generally significantly higher than the mean Monday return of the last two weeks. Our finding seems to make it more difficult to explain the Monday effect.*²⁵

A friend thinks that “significantly” in this article has its plain English meaning, roughly “I think this is important.” Explain in simple language what “significantly higher” and “not significantly different from zero” actually tell us here.

3.41 Five-digit zip codes and delivery time of mail. Does adding the five-digit postal zip code to an address really speed up delivery of letters? Does adding the four more digits that make up “zip + 4”

speed delivery yet more? What about mailing a letter on Monday, Thursday, or Saturday? Describe the design of an experiment on the speed of first-class mail delivery. For simplicity, suppose that all letters go from you to a friend, so that the sending and receiving locations are fixed.

3.42  **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 cell phone study in Example 3.11 (page 185) have fast reactions, and that the odd-numbered students have slow reactions. We would like the experimental and control groups to contain similar numbers of the fast reactors. 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.

3.43 Does oxygen help football players? We often see players on the sidelines of a football game inhaling oxygen. Their coaches think this will speed their recovery. We might measure recovery from intense exercise as follows: Have a football player run 100 yards three times in quick succession. Then allow three minutes to rest before running 100 yards again. Time the final run. Because players vary greatly in speed, you plan a matched pairs experiment using 20 football players as subjects. Describe the design of such an experiment to investigate the effect of inhaling oxygen during the rest period. Why should each player’s two trials be on different days? Use Table B at line 140 to decide which players will get oxygen on their first trial.

3.44 Carbon dioxide in the atmosphere. The concentration of carbon dioxide (CO_2) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use CO_2 to fuel photosynthesis, more CO_2 may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra CO_2 to a 30-meter circle of forest. We want to compare the growth in base area of trees in treated and untreated areas to see if extra CO_2 does in fact increase growth. We can afford to treat 3 circular areas.²⁶

(a) Describe the design of a completely randomized experiment using 6 well-separated 30-meter circular areas in a pine forest. Sketch the forest area with the 6 circles and carry out the randomization your design calls for.

(b) Regions within the forest may differ in soil fertility. Describe a matched pairs design using three pairs of circles that will reduce the extra variation due to different fertility. Sketch the forest area with the new arrangement of circles and carry out the randomization your design calls for.


3.45  **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 Hawaiian Punch at either a high or a low level. The camp diet was otherwise the same for all girls. Suppose that there are 50 campers.

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

(b) Describe a matched pairs design in which each girl receives both levels of calcium (with a “washout period” between). 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. Label the subjects 01 to 50. You must choose 25 of the 50. Use Table B at line 110 to choose just the first 5 of the 25. How are the 25 subjects chosen treated in the completely randomized design? How are they treated in the matched pairs design?

3.46  **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.

3.47 **Vitamin C for ultramarathon runners.** An ultramarathon, as you might guess, is a footrace longer than the 26.2 miles of a marathon. Runners commonly develop respiratory infections after an ultramarathon. Will taking 600 milligrams of vitamin C daily reduce these infections? Researchers randomly assigned ultramarathon runners to receive either vitamin C or a placebo. Separately, they also randomly assigned these treatments to a group of nonrunners the same age as the runners. All subjects were watched for 14 days after the big race to see if infections developed.²⁷

(a) What is the name for this experimental design?

(b) Use a diagram to outline the design.

(c) The report of the study said:

Sixty-eight percent of the runners in the placebo group reported the development of symptoms of upper respiratory tract infection after the race; this was significantly more than that reported by the vitamin C-supplemented group (33%).

Explain to someone who knows no statistics why “significantly more” means there is good reason to think that vitamin C works.

3.2 Sampling Design

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 sport 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 survey **Sample surveys** are an important kind of observational study.

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.

sample design

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.



EXAMPLE

3.21 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.²⁸ The National Data Evaluation Center (NDEC) Web site (www.ndec.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. In spite of these difficulties, we still view the list as the population. Also, we do not expect to get a response from every teacher in our random sample. 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 our survey questions.

response rate

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%.

USE YOUR KNOWLEDGE

3.48 Job satisfaction in Mongolian universities. A 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 300 faculty members from a list of 2500 faculty members in these universities. The Job Descriptive Index (JDI) 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. Describe the population and the sample for this study. Can you determine the response rate?

3.49 Taxes and forestland usage. A study was designed to assess the impact of taxes on forestland usage in part of the Upper Wabash River Watershed in Indiana.²⁹ A survey was sent to 772 forest owners from this region and 348 were returned. Consider the population, the sample, and the response rate for this study. Describe these based on the information given and indicate any additional information that you would need to give a complete answer.

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.

Online opinion polls are particularly vulnerable to bias because the sample who respond are not representative of the population at large. Here is an example that also illustrates how the results of such polls can be manipulated.

EXAMPLE

3.23 The American Family Association. The American Family Association (AFA) is a conservative group that claims to stand for “traditional family values.” It regularly posts online poll questions on its Web site—just click on a response to take part. Because the respondents are people who visit this site, the poll results always support AFA’s positions. Well, almost always. In 2004, AFA’s online poll asked about the heated issue of allowing same-sex marriage. Soon, email lists and social-network sites favored mostly by young liberals pointed to the AFA poll. Almost 850,000 people responded, and 60% of them favored legalization of same-sex marriage. AFA claimed that homosexual rights groups had skewed its poll.

As the AFA poll illustrates, you can’t always trust poll results. People who take the trouble to respond to an open invitation are not representative of the entire adult population. That’s true of regular visitors to AFA’s site, of the activists who made a special effort to vote in the marriage poll, and of the people who bother to respond to write-in, call-in, or online polls in general.

In both Examples 3.22 and 3.23, the sample was selected in a manner that guaranteed that it would not be representative of the entire population. These sampling schemes display *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.

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 impersonal chance to do the choosing, so that there is neither favoritism by the sampler (as in Example 3.22) nor voluntary response (as in Example 3.23). 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 subjects.

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 each individual an equal chance to be chosen (thus avoiding bias in the choice) but gives every possible sample 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.

EXAMPLE

3.24 Spring break destinations. A campus newspaper plans a major article on spring break destinations. The authors intend to call a few randomly chosen resorts at each destination to ask about their attitudes toward groups of students as guests. Here are the resorts listed in one city. The first step is to label the members of this population as shown.



01	Aloha Kai	08	Captiva	15	Palm Tree	22	Sea Shell
02	Anchor Down	09	Casa del Mar	16	Radisson	23	Silver Beach
03	Banana Bay	10	Coconuts	17	Ramada	24	Sunset Beach
04	Banyan Tree	11	Diplomat	18	Sandpiper	25	Tradewinds
05	Beach Castle	12	Holiday Inn	19	Sea Castle	26	Tropical Breeze
06	Best Western	13	Lime Tree	20	Sea Club	27	Tropical Shores
07	Cabana	14	Outrigger	21	Sea Grape	28	Veranda

Now enter Table B, and read two-digit groups until you have chosen three resorts. If you enter at line 185, Banana Bay (03), Palm Tree (15), and Cabana (07) will be called.

Most statistical software will select an SRS for you, eliminating the need for Table B. The *Simple Random Sample* applet on the text CD and Web site is a convenient way to automate this task.

Excel can do the job in a way similar to what we used when we randomized experimental units to treatments in designed experiments. There are four steps:

1. Create a data set with all of the elements of the population in the first column.
2. Assign a random number to each element of the population; put these in the second column.
3. Sort the data set by the random number column.
4. The simple random sample is obtained by taking elements in the sorted list until the desired sample size is reached.

We illustrate the procedure with a simplified version of Example 3.24.



EXAMPLE

3.25 Select a random sample. Suppose that the population from Example 3.24 is only the first two rows of the display given there:

Aloha Kai	Captiva	Palm Tree	Sea Shell
Anchor Down	Casa del Mar	Radisson	Silver Beach

Note that we do not need the numerical labels to identify the individuals in the population. Suppose that we want to select a simple random sample of three resorts from this population. Figure 3.8(a) gives the spreadsheet with the population names. The random numbers generated by the `RAND()` function are given in the second column in Figure 3.8(b). The sorted data set is given in Figure 3.8(c). We have added a third column to the spreadsheet to indicate which resorts were selected for our random sample. They are Captiva, Radisson, and Silver Beach.

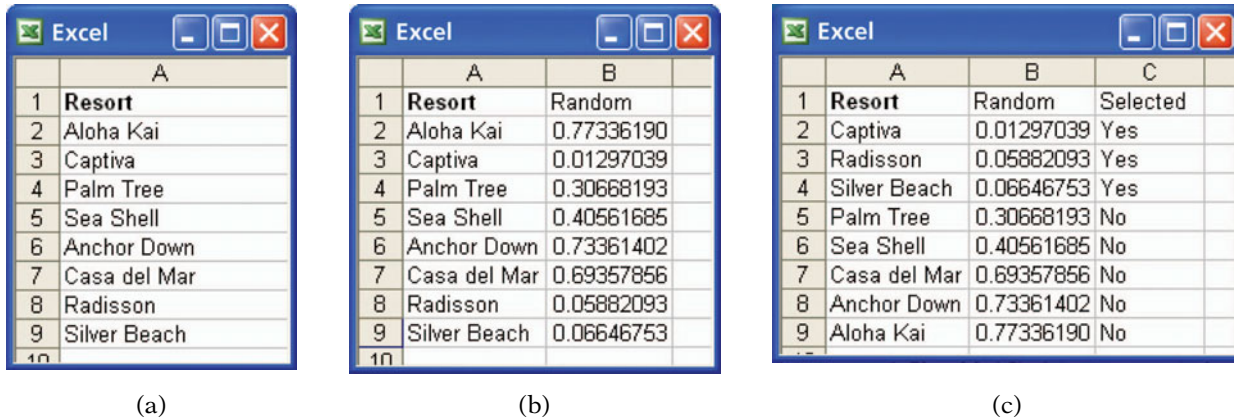


FIGURE 3.8 Selection of a simple random sample of resorts, for Example 3.25.

USE YOUR KNOWLEDGE

3.50 Ringtones for cell phones. You decide to change the ringtones for your cell phone by choosing 2 from a list of the 10 most popular ringtones.³⁰ Here is the list:

Super Mario Brothers Theme	Sexy Love	Ms. New Booty	Ridin' Rims
I Write Sins Not Tragedies	Gasolina	My Humps	The Pink Panther
Down	Agarrala		

Select your two ringtones using a simple random sample.

3.51 Listen to three 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. You decide to take a simple random sample of songs from a Billboard list of Rock Songs.³¹ Here is the list:

Miss Murder	Animal I Have Become	Steady, As She Goes	Dani California
The Kill (Bury Me)	Original Fire	When You Were Young	MakeD—Sure
Vicarious	The Diary of Jane		

Select the three songs for your iPod using a simple random sample.

Stratified 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 dental claims. A dentist is suspected of defrauding insurance companies by describing some dental procedures incorrectly on claim forms and overcharging for them. An investigation begins by examining a sample of his bills for the past three years. Because there are five suspicious types of procedures, the investigators take a stratified sample. That is, they randomly select bills for each of the five types of procedures separately.

Multistage samples

Another common means of restricting random selection is to choose the sample in stages. This is common practice for 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. The Current Population Survey therefore uses a **multistage sampling design**. The final sample consists of clusters of nearby households that an interviewer can easily visit.

multistage sample

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:³²

- 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.³³ 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 prison inmates and students in dormitories. An opinion poll conducted by telephone will miss the 6% 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 4% of the households in the CPS sample refuse to take part and another 3% or 4% can't be contacted. 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.

What about polls done by the media and by market research and opinion-polling firms? We don't know their rates of nonresponse, because they won't say. That itself is a bad sign. The Pew Research Center for People and the Press designed a careful telephone survey and published the results: out of 2879 households called, 1658 were never at home, refused, or would not finish the interview. That's a nonresponse rate of 58%.³⁴

The screenshot shows the 'General Social Survey Codebook' website. At the top, it says 'THE NATIONAL OPINION RESEARCH CENTER AT THE UNIVERSITY OF CHICAGO'. There are navigation buttons for 'Pick List', 'Extract', 'Analyze', 'Homepage', 'Site Map', 'e-Mail GSS', and 'Search'. A 'Site Help' button is in the top right. On the left, there is a sidebar with links for 'Introduction', 'About GSSDirs', 'GSS News', 'Credits', 'Codebook Indexes', 'Mnemonic', 'Sequential', 'Subject', 'Collections', 'GSS Publications', 'Questionnaires', and 'Appendix'. The main content area is titled 'Subject Index: E' and has a 'Help' button. Below the title are buttons for 'Previous', 'Pick Page', and 'Next'. A horizontal bar contains the letters A through Z, with 'E' highlighted. Below this bar, a list of subjects is shown, including 'Economy', 'Education', 'Egypt, see Countries', 'Elderly, see Aged', 'Elections, see Political', 'Employment', 'England Countries', 'English, see Citizen Obligation', 'Environment', 'Equal Rights Amendment', 'Equality, see Income, Inequality', 'Esp, see Religion', 'Ethnicity', 'Euthanasia', 'Evenings, see Sociability', 'Experiments, see Split Ballots', and 'Extramarital sex, see Sex'.

FIGURE 3.9 Part of the subject index 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.

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.

EXAMPLE

3.28 How should the data be collected? A Pew Research Center Poll has asked about belief in God for many years. In response to the statement “I never doubt the existence of God,” subjects are asked to choose from the responses

completely agree mostly agree mostly disagree completely disagree

In 1990, subjects were interviewed in person and were handed a card with the four responses on it. In 1991, the poll switched to telephone interviews. In 1990, 60% said “completely agree,” in line with earlier years. In 1991, 71% completely agreed. The increase is probably explained by the effect of hearing “completely agree” read first by the interviewer.³⁵

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 6 months?” will often elicit a “Yes” from someone who last visited a dentist 8 months ago.³⁶

EXAMPLE

3.29 Overreporting of voter behavior. “One of the most frequently observed survey measurement errors is the overreporting of voting behavior.”³⁷ People know they should vote, so those who didn’t vote tend to save face by saying that they did. Here are the data from a typical sample of 663 people after an election:

		What they said:	
		I voted	I didn’t
What they did:	Voted	358	13
	Didn’t vote	120	172

You can see that 478 people (72%) said that they voted, but only 371 people (56%) actually did vote.

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

3.30 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.³⁸



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.*

SECTION 3.2 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 by labeling the members of the population and using a **table of random digits** to select the sample. Software can automate this process.

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 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 3.2 Exercises

For Exercises 3.48 and 3.49, see pages 198 and 199; and for Exercises 3.50 and 3.51, see page 202.

3.52 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 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.

3.53 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 of 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.

3.54 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 150 businesses at random. Of these, 73 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?

3.55 Popularity of news personalities. A Gallup Poll conducted telephone interviews with 1001 U.S.

adults aged 18 and over on July 24–27, 2006. One of the questions asked whether the respondents had a favorable or an unfavorable opinion of 17 news personalities. Diane Sawyer received the highest rating, with 80% of the respondents giving her a favorable rating.³⁹

(a) What is the population for this sample survey? What was the sample size?

(b) The report on the survey states that 8% of the respondents either never heard of Sawyer or had no opinion about her. When they included only those who provided an opinion, Sawyer's approval percent rose to 88% and she was still at the top of the list. Charles Gibson, on the other hand, was ranked eighth on the original list, with a 55% favorable rating. When only those providing an opinion were counted, his rank rose to second, with 87% approving. Discuss the advantages and disadvantages of the two different ways of reporting the approval percent. State which one you prefer and why.

3.56 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 5 students will be selected to be interviewed.

(b) The American Community Survey (ACS) will replace the census "long form" starting with the 2010 census. The main part of 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?"


3.57 Interview residents of apartment complexes. You are planning a report on apartment living in


a college town. You decide to select 5 apartment complexes at random for in-depth interviews with residents. Select a simple random sample of 5 of the following apartment complexes. If you use Table B, start at line 137.

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

3.58 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 25 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 112 and choose only the first 5 pixels in the sample.

3.59  **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.

3.60  **Use the simple random sample applet.** There are approximately 371 active telephone area codes covering Canada, the United States, and some Caribbean areas. (More are created regularly.) You want to choose an SRS of 25 of these area codes for a study of available telephone numbers. Label the codes 001 to 371 and use the *Simple Random Sample* applet to choose your sample. (If you use Table B, start at line 120 and choose only the first 5 codes in the sample.)

3.61 Census tracts. The Census Bureau divides the entire country into “census tracts” that contain about 4000 people. Each tract is in turn divided into small “blocks,” which in urban areas are bounded by local streets. An SRS of blocks from a census tract is often the next-to-last stage in a multistage sample. Figure 3.10 shows part of census tract 8051.12, in Cook County, Illinois, west of Chicago. The 44 blocks in this tract are divided into three “block groups.” Group 1 contains 6 blocks numbered 1000 to 1005;

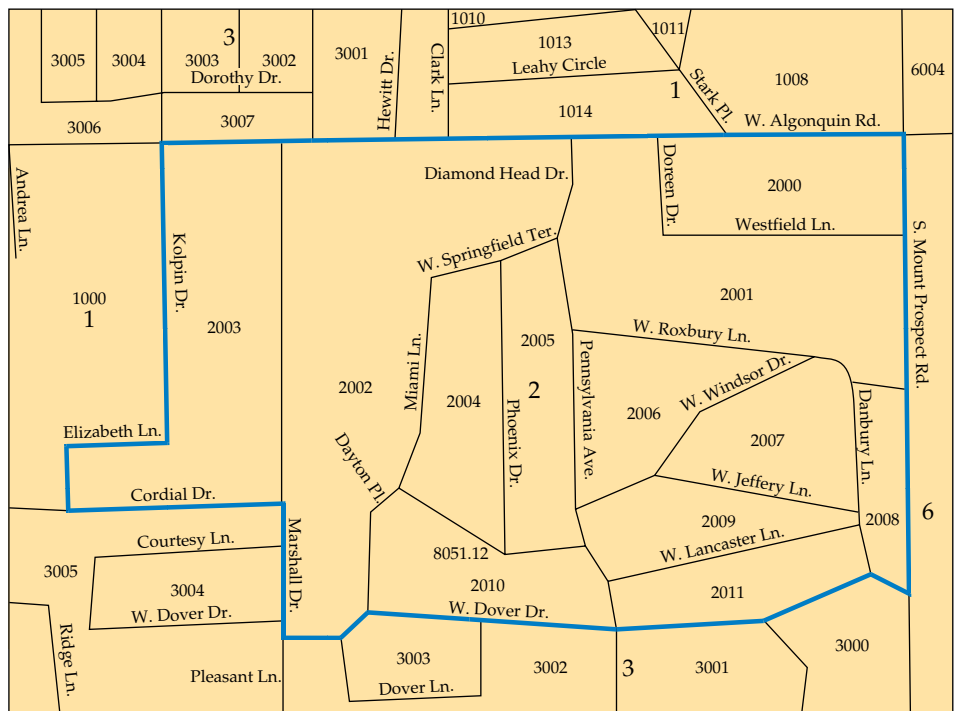


FIGURE 3.10 Census blocks in Cook County, Illinois, for Exercises 3.61 and 3.63. The outlined area is a block group.

Group 2 (outlined in Figure 3.10) contains 12 blocks numbered 2000 to 2011; Group 3 contains 26 blocks numbered 3000 to 3025. Use Table B, beginning at line 135, to choose an SRS of 5 of the 44 blocks in this census tract. Explain carefully how you labeled the blocks.

3.62 Repeated use of Table B. In using Table B repeatedly to choose samples or do randomization for experiments, you should not always begin at the same place, such as line 101. Why not?

3.63 A stratified sample. Exercise 3.61 asks you to choose an SRS of blocks from the census tract pictured in Figure 3.10. You might instead choose a stratified sample of one block from the 6 blocks in Group 1, two from the 12 blocks in Group 2, and three from the 26 blocks in Group 3. Choose such a sample, explaining carefully how you labeled blocks and used Table B.

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 4 addresses out of 100. Because $100/4 = 25$, we can think of the list as four lists of 25 addresses. Choose 1 of the first 25 at random, using Table B. The sample contains this address and the addresses 25, 50, and 75 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, and 88.

(a) A study of dating among college students wanted a sample of 200 of the 9000 single male students on campus. The sample consisted of every 45th name from a list of the 9000 students. Explain why the survey chooses every 45th name.

(b) Use Table B at line 125 to choose the starting point for this systematic sample.

3.65 CHALLENGE 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.

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

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

530 707 916 209 415 925 510 650 408 831 805 559 760
661 818 213 626 323 562 709 310 949 909 858 619

Another California survey does not call numbers in all area codes but starts with an SRS of 10 area codes. Choose such an SRS. If you use Table B, start at line 122.

3.67 Stratified samples of forest areas. Stratified samples are widely used to study large areas of forest. Based on satellite images, a forest area in the Amazon basin is divided into 14 types. Foresters studied the four most commercially valuable types: alluvial climax forests of quality levels 1, 2, and 3, and mature secondary forest. They divided the area of each type into large parcels, chose parcels of each type at random, and counted tree species in a 20-by 25-meter rectangle randomly placed within each parcel selected. Here is some detail:

Forest type	Total parcels	Sample size
Climax 1	36	4
Climax 2	72	7
Climax 3	31	3
Secondary	42	4

Choose the stratified sample of 18 parcels. Be sure to explain how you assigned labels to parcels. If you use Table B, start at line 140.


3.68 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 5 students and 3 faculty members to a convention and decides to choose those who will go by random selection. Select a stratified random sample of 5 students and 3 faculty members.

- 3.69**  **Stratified samples for alcohol attitudes.** At a party there are 30 students over age 21 and 20 students under age 21. You choose at random 3 of those over 21 and separately choose at random 2 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?

- 3.70** **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 Table B, starting at line 115, to select the first 5 accounts from each of these strata.

- 3.71** **Nonresponse in telephone surveys.** A common form of nonresponse in telephone surveys is “ring-no-answer.” That is, a call is made to an active number but no one answers. The Italian National Statistical Institute looked at nonresponse to a government survey of households in Italy during the periods January 1 to Easter and July 1 to August 31. All calls were made between 7 and 10 P.M., but 21.4% gave “ring-no-answer” in one period versus 41.5% “ring-no-answer” in the other period.⁴⁰ Which period do you think had the higher rate of no answers? Why? Explain why a high rate of nonresponse makes sample results less reliable.

- 3.72** **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 first three digits). Which of the households that you mentioned in your answer to (a) will be included in the sampling frame by random digit dialing?

- 3.73** **The Excite Poll.** The Excite Poll can be found online at poll.excite.com. The question appears on the screen, and you simply click buttons to vote “Yes,” “No,” “Not sure,” or “Don’t care.” On July 22, 2006, the question was “Do you agree or disagree with proposed legislation that would discontinue the U.S. penny coin?” In all, 631 said “Yes,” another 564 said “No,” and the remaining 65 indicated that they were not sure.

- (a) What is the sample size for this poll?
 (b) Compute the percent of responses in each of the possible response categories.
 (c) Discuss the poll in terms of the population and sample framework that we have studied in this chapter.

- 3.74** **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?”


- 3.75** **Use of a budget surplus.** In 2000, when the federal budget showed a large surplus, the Pew Research Center asked two questions of random samples of adults. Both questions stated that Social Security would be “fixed.” Here are the uses suggested for the remaining surplus:

Should the money be used for a tax cut, or should it be used to fund new government programs?

Should the money be used for a tax cut, or should it be spent on programs for education, the environment, health care, crime-fighting and military defense?


One of these questions drew 60% favoring a tax cut; the other, only 22%. Which wording pulls respondents toward a tax cut? Why?

3.76  **How many children are in your family?** A teacher asks her class, “How many children are there in your family, including yourself?” The mean response is about 3 children. According to the 2000 census, families that have children average 1.86 children. Why is a sample like this biased toward higher outcomes?

3.77  **Bad survey questions.** Write your own examples of bad sample survey questions.

(a) Write a biased question designed to get one answer rather than another.

(b) Write a question that is confusing, so that it is hard to answer.

3.78  **Economic attitudes of Spaniards.** Spain’s Centro de Investigaciones Sociológicas carried out a sample survey on the economic attitudes of Spaniards.⁴¹ Of the 2496 adults interviewed, 72% agreed that “Employees with higher performance must get higher pay.” On the other hand, 71% agreed that “Everything a society produces should be distributed among its members as equally as possible and there should be no major differences.” Use these conflicting results as an example in a short explanation of why opinion polls often fail to reveal public attitudes clearly.

3.3 Toward Statistical Inference

A market research firm interviews a random sample of 2500 adults. Result: 66% find shopping for clothes frustrating and time-consuming. That’s the truth about the 2500 people in the sample. What is the truth about the almost 220 million American adults who make up the population? Because the sample was chosen at random, it’s reasonable to think that these 2500 people represent the entire population fairly well. So the market researchers turn the *fact* that 66% of the *sample* find shopping frustrating into an *estimate* that about 66% of *all adults* feel this way. That’s a basic move in statistics: use a fact about a sample to estimate the truth about the whole population. We call this **statistical inference** because we infer conclusions about the wider population from data on selected individuals. To think about inference, we must keep straight whether a number describes a sample or a population. Here is the vocabulary we use.

statistical inference

PARAMETERS AND STATISTICS

A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we do not know its value.

A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.

EXAMPLE

3.31 Attitudes toward shopping. Are attitudes toward shopping changing? Sample surveys show that fewer people enjoy shopping than in the past. A survey by the market research firm Yankelovich Clancy Shulman asked a nationwide random sample of 2500 adults if they agreed or disagreed that “I like buying new clothes, but shopping is often frustrating and time-consuming.” Of the respondents, 1650, or 66%, said they agreed.⁴² The



proportion of the sample who agree is

$$\hat{p} = \frac{1650}{2500} = 0.66 = 66\%$$

The number $\hat{p} = 0.66$ is a *statistic*. The corresponding *parameter* is the proportion (call it p) of all adult U.S. residents who would have said “Agree” if asked the same question. We don’t know the value of the parameter p , so we use the statistic \hat{p} to estimate it.

USE YOUR KNOWLEDGE

3.79 Sexual harassment of college students. A recent survey of 2036 undergraduate college students aged 18 to 24 reports that 62% of college students say they have encountered some type of sexual harassment while at college.⁴³ Describe the sample and the population for this setting.

3.80 Web polls. If you connect to the Web site worldnetdaily.com/polls/, you will be given the opportunity to give your opinion about a different question of public interest each day. Can you apply the ideas about populations and samples that we have just discussed to this poll? Explain why or why not.

Sampling variability

sampling variability

If Yankelovich took a second random sample of 2500 adults, the new sample would have different people in it. It is almost certain that there would not be exactly 1650 positive responses. That is, the value of the statistic \hat{p} will vary from sample to sample. This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling. Could it happen that one random sample finds that 66% of adults find clothes shopping frustrating and a second random sample finds that only 42% feel this way? Random samples eliminate *bias* from the act of choosing a sample, but they can still be wrong because of the *variability* that results when we choose at random. If the variation when we take repeat samples from the same population is too great, we can’t trust the results of any one sample.

We are saved by the second great advantage of random samples. The first advantage is that choosing at random eliminates favoritism. That is, random sampling attacks bias. The second advantage is that if we take lots of random samples of the same size from the same population, the variation from sample to sample will follow a predictable pattern. **All of statistical inference is based on one idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.**

To understand why sampling variability is not fatal, we ask, “What would happen if we took many samples?” Here’s how to answer that question:

- Take a large number of samples from the same population.
- Calculate the sample proportion \hat{p} for each sample.
- Make a histogram of the values of \hat{p} .

- Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

In practice it is too expensive to take many samples from a large population such as all adult U.S. residents. But we can imitate many samples by using random digits. Using random digits from a table or computer software to imitate chance behavior is called **simulation**.

simulation

EXAMPLE

3.32 Simulate a random sample. We will simulate drawing simple random samples (SRSs) of size 100 from the population of all adult U.S. residents. Suppose that in fact 60% of the population find clothes shopping time-consuming and frustrating. Then the true value of the parameter we want to estimate is $p = 0.6$. (Of course, we would not sample in practice if we already knew that $p = 0.6$. We are sampling here to understand how sampling behaves.)

We can imitate the population by a table of random digits, with each entry standing for a person. Six of the ten digits (say 0 to 5) stand for people who find shopping frustrating. The remaining four digits, 6 to 9, stand for those who do not. Because all digits in a random number table are equally likely, this assignment produces a population proportion of frustrated shoppers equal to $p = 0.6$. We then imitate an SRS of 100 people from the population by taking 100 consecutive digits from Table B. The statistic \hat{p} is the proportion of 0s to 5s in the sample.

Here are the first 100 entries in Table B with digits 0 to 5 highlighted:

19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	00095				

There are 64 digits between 0 and 5, so $\hat{p} = 64/100 = 0.64$. A second SRS based on the second 100 entries in Table B gives a different result, $\hat{p} = 0.55$. The two sample results are different, and neither is equal to the true population value $p = 0.6$. That's sampling variability.

Sampling distributions

Simulation is a powerful tool for studying chance. Now that we see how simulation works, it is faster to abandon Table B and to use a computer programmed to generate random numbers.

EXAMPLE

3.33 Take many random samples. Figure 3.11 illustrates the process of choosing many samples and finding the sample proportion \hat{p} for each one. Follow the flow of the figure from the population at the left, to choosing an SRS and finding the \hat{p} for this sample, to collecting together the \hat{p} 's from many samples. The histogram at the right of the figure shows the distribution of the values of \hat{p} from 1000 separate SRSs of size 100 drawn from a population with $p = 0.6$.

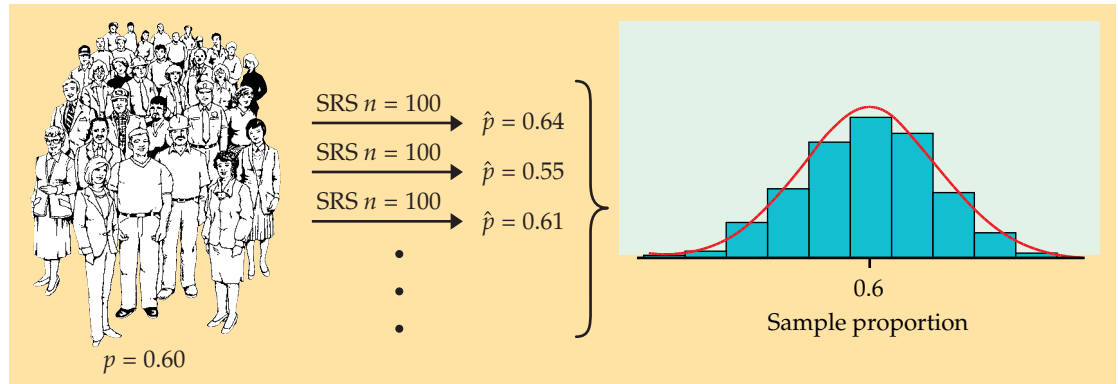


FIGURE 3.11 The results of many SRSs have a regular pattern. Here, we draw 1000 SRSs of size 100 from the same population. The population proportion is $p = 0.60$. The histogram shows the distribution of the 1000 sample proportions.

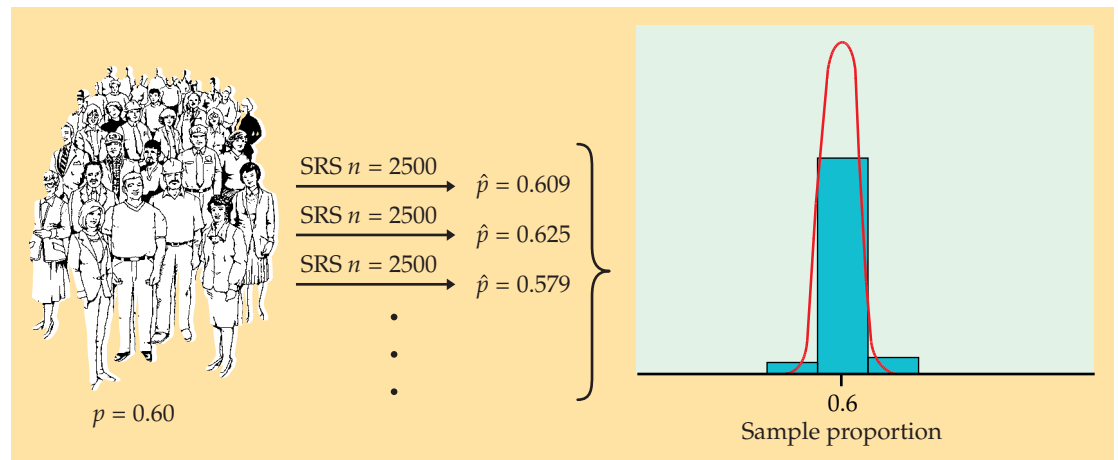


FIGURE 3.12 The distribution of sample proportions for 1000 SRSs of size 2500 drawn from the same population as in Figure 3.11. The two histograms have the same scale. The statistic from the larger sample is less variable.

Of course, Yankelovich interviewed 2500 people, not just 100. Figure 3.12 is parallel to Figure 3.11. It shows the process of choosing 1000 SRSs, each of size 2500, from a population in which the true proportion is $p = 0.6$. The 1000 values of \hat{p} from these samples form the histogram at the right of the figure. Figures 3.11 and 3.12 are drawn on the same scale. Comparing them shows what happens when we increase the size of our samples from 100 to 2500. These histograms display the *sampling distribution* of the statistic \hat{p} for two sample sizes.

SAMPLING DISTRIBUTION

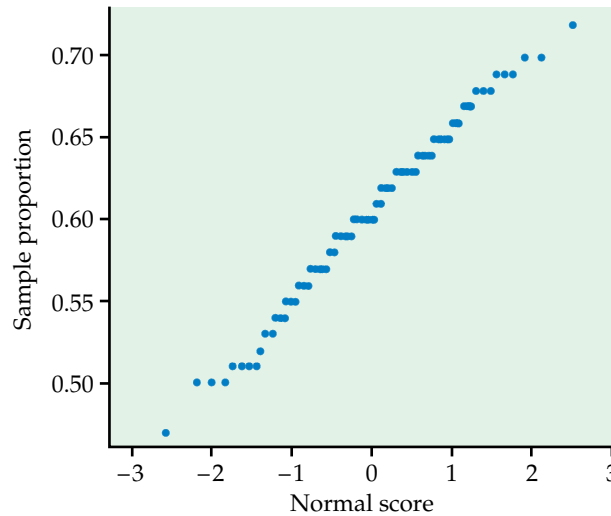
The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Strictly speaking, the sampling distribution is the ideal pattern that would emerge if we looked at all possible samples of size 100 from our population. A distribution obtained from a fixed number of trials, like the 1000 trials in Figure 3.11, is only an approximation to the sampling distribution. We will see that probability theory, the mathematics of chance behavior, can sometimes describe sampling distributions exactly. The interpretation of a sampling distribution is the same, however, whether we obtain it by simulation or by the mathematics of probability.

We can use the tools of data analysis to describe any distribution. Let's apply those tools to Figures 3.11 and 3.12.

- **Shape:** The histograms look Normal. Figure 3.13 is a Normal quantile plot of the values of \hat{p} for our samples of size 100. It confirms that the distribution in Figure 3.11 is close to Normal. The 1000 values for samples of size 2500 in Figure 3.12 are even closer to Normal. The Normal curves drawn through the histograms describe the overall shape quite well.

FIGURE 3.13 Normal quantile plot of the sample proportions in Figure 3.11. The distribution is close to Normal except for some granularity due to the fact that sample proportions from a sample of size 100 can take only values that are multiples of 0.01. Because a plot of 1000 points is hard to read, this plot presents only every 10th value.



- **Center:** In both cases, the values of the sample proportion \hat{p} vary from sample to sample, but the values are centered at 0.6. Recall that $p = 0.6$ is the true population parameter. Some samples have a \hat{p} less than 0.6 and some greater, but there is no tendency to be always low or always high. That is, \hat{p} has no **bias** as an estimator of p . This is true for both large and small samples. (Want the details? The mean of the 1000 values of \hat{p} is 0.598 for samples of size 100 and 0.6002 for samples of size 2500. The median value of \hat{p} is exactly 0.6 for samples of both sizes.)
- **Spread:** The values of \hat{p} from samples of size 2500 are much less spread out than the values from samples of size 100. In fact, the standard deviations are 0.051 for Figure 3.11 and 0.0097, or about 0.01, for Figure 3.12.

Although these results describe just two sets of simulations, they reflect facts that are true whenever we use random sampling.

USE YOUR KNOWLEDGE

3.81 Effect of sample size on the sampling distribution. You are planning a study and are considering taking an SRS of either 200 or 400 observations. Explain how the sampling distribution would differ for these two scenarios.

Bias and variability

Our simulations show that a sample of size 2500 will almost always give an estimate \hat{p} that is close to the truth about the population. Figure 3.12 illustrates this fact for just one value of the population proportion, but it is true for any population. Samples of size 100, on the other hand, might give an estimate of 50% or 70% when the truth is 60%.

Thinking about Figures 3.11 and 3.12 helps us restate the idea of bias when we use a statistic like \hat{p} to estimate a parameter like p . It also reminds us that variability matters as much as bias.

BIAS AND VARIABILITY

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size n . Statistics from larger probability samples have smaller spreads.

We can think of the true value of the population parameter as the bull's-eye on a target, and of the sample statistic as an arrow fired at the bull's-eye. Bias and variability describe what happens when an archer fires many arrows at the target. *Bias* means that the aim is off, and the arrows land consistently off the bull's-eye in the same direction. The sample values do not center about the population value. Large *variability* means that repeated shots are widely scattered on the target. Repeated samples do not give similar results but differ widely among themselves. Figure 3.14 shows this target illustration of the two types of error.

Notice that small variability (repeated shots are close together) can accompany large bias (the arrows are consistently away from the bull's-eye in one direction). And small bias (the arrows center on the bull's-eye) can accompany large variability (repeated shots are widely scattered). A good sampling scheme, like a good archer, must have both small bias and small variability. Here's how we do this.

MANAGING BIAS AND VARIABILITY

To reduce bias, use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased

estimates—the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.

To reduce the variability of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

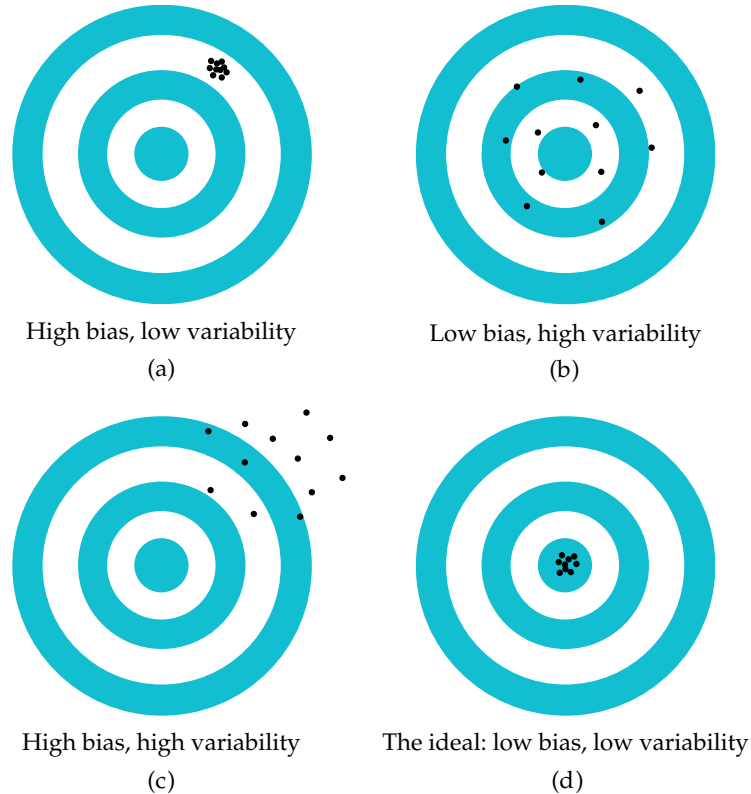


FIGURE 3.14 Bias and variability in shooting arrows at a target. Bias means the archer systematically misses in the same direction. Variability means that the arrows are scattered.

In practice, Yankelovich takes only one sample. We don't know how close to the truth an estimate from this one sample is because we don't know what the truth about the population is. But *large random samples almost always give an estimate that is close to the truth*. Looking at the pattern of many samples shows that we can trust the result of one sample. The Current Population Survey's sample of 60,000 households estimates the national unemployment rate very accurately. Of course, only probability samples carry this guarantee. The American Family Association's voluntary response sample (Example 3.23, page 199) is worthless even though 850,000 people responded. Using a probability sampling design and taking care to deal with practical difficulties reduce bias in a sample. The size of the sample then determines how close to the population truth the sample result is likely to fall. Results from a sample survey usually come with a **margin of error** that sets bounds on the size of the likely error. The margin of error directly reflects the variability of the sample statistic, so it is smaller for larger samples. We will describe the details in later chapters.

margin of error

Sampling from large populations

Yankelovich's sample of 2500 adults is only about 1 out of every 90,000 adults in the United States. Does it matter whether we sample 1-in-100 individuals in the population or 1-in-90,000?

POPULATION SIZE DOESN'T MATTER

The variability of a statistic from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

Why does the size of the population have little influence on the behavior of statistics from random samples? To see why this is plausible, imagine sampling harvested corn by thrusting a scoop into a lot of corn kernels. The scoop doesn't know whether it is surrounded by a bag of corn or by an entire truckload. As long as the corn is well mixed (so that the scoop selects a random sample), the variability of the result depends only on the size of the scoop.

The fact that the variability of sample results is controlled by the size of the sample has important consequences for sampling design. An SRS of size 2500 from the 220 million adult residents of the United States gives results as precise as an SRS of size 2500 from the 665,000 adult inhabitants of San Francisco. This is good news for designers of national samples but bad news for those who want accurate information about the citizens of San Francisco. If both use an SRS, both must use the same size sample to obtain equally trustworthy results.

Why randomize?

Why randomize? The act of randomizing guarantees that the results of analyzing our data are subject to the laws of probability. The behavior of statistics is described by a sampling distribution. The form of the distribution is known, and in many cases is approximately Normal. Often the center of the distribution lies at the true parameter value, so that the notion that randomization eliminates bias is made more precise. The spread of the distribution describes the variability of the statistic and can be made as small as we wish by choosing a large enough sample. In a randomized experiment, we can reduce variability by choosing larger groups of subjects for each treatment.

These facts are at the heart of formal statistical inference. Later chapters will have much to say in more technical language about sampling distributions and the way statistical conclusions are based on them. What any user of statistics must understand is that all the technical talk has its basis in a simple question: *What would happen if the sample or the experiment were repeated many times?* The reasoning applies not only to an SRS but also to the complex sampling designs actually used by opinion polls and other national sample surveys. The same conclusions hold as well for randomized experimental designs. The details vary with the design but the basic facts are true whenever randomization is used to produce data.

Remember that proper statistical design is not the only aspect of a good sample or experiment. *The sampling distribution shows only how a statistic*



varies due to the operation of chance in randomization. It reveals nothing about possible bias due to undercoverage or nonresponse in a sample, or to lack of realism in an experiment. The actual error in estimating a parameter by a statistic can be much larger than the sampling distribution suggests. What is worse, there is no way to say how large the added error is. The real world is less orderly than statistics textbooks imply.

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

3.34 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

proportion banded in sample = 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 number that describes a population is a **parameter**. A number that can be computed from the data is a **statistic**. The purpose of sampling or experimentation is usually **inference**: use sample statistics to make statements about unknown population parameters.

A statistic from a probability sample or randomized experiment has a **sampling distribution** that describes how the statistic varies in repeated data production. The sampling distribution answers the question “What would happen if we repeated the sample or experiment many times?” Formal statistical inference is based on the sampling distributions of statistics.

A statistic as an estimator of a parameter may suffer from **bias** or from high **variability**. Bias means that the center of the sampling distribution is not equal to the true value of the parameter. The variability of the statistic is described by the spread of its sampling distribution. Variability is usually reported by giving a **margin of error** for conclusions based on sample results.

Properly chosen statistics from randomized data production designs have no bias resulting from the way the sample is selected or the way the experimental units are assigned to treatments. We can reduce the variability of the statistic by increasing the size of the sample or the size of the experimental groups.

SECTION 3.3 Exercises

For Exercises 3.79 and 3.80, see page 213; and for Exercise 3.81, see page 217.

3.82 What’s wrong? State what is wrong in each of the following scenarios.

- (a) A sampling distribution describes the distribution of some characteristic in a population.
- (b) A statistic will have a large amount of bias whenever it has high variability.
- (c) The variability of a statistic based on a small sample from a population will be the same as the variability of a large sample from the same population.

3.83 Describe the population and the sample. For each of the following situations, describe the population and the sample.

- (a) A survey of 17,096 students in U.S. four-year colleges reported that 19.4% were binge drinkers.

- (b) In a study of work stress, 100 restaurant workers were asked about the impact of work stress on their personal lives.

- (c) A tract of forest has 584 longleaf pine trees. The diameters of 40 of these trees were measured.

3.84 Bias and variability. Figure 3.15 (on page 222) shows histograms of four sampling distributions of statistics intended to estimate the same parameter. Label each distribution relative to the others as high or low bias and as high or low variability.

3.85 Opinions of Hispanics. A New York Times News Service article on a poll concerned with the opinions of Hispanics includes this paragraph:

The poll was conducted by telephone from July 13 to 27, with 3,092 adults nationwide, 1,074 of whom described themselves as Hispanic. It has a margin of sampling error of plus or minus three percentage points for the entire poll and plus or minus four

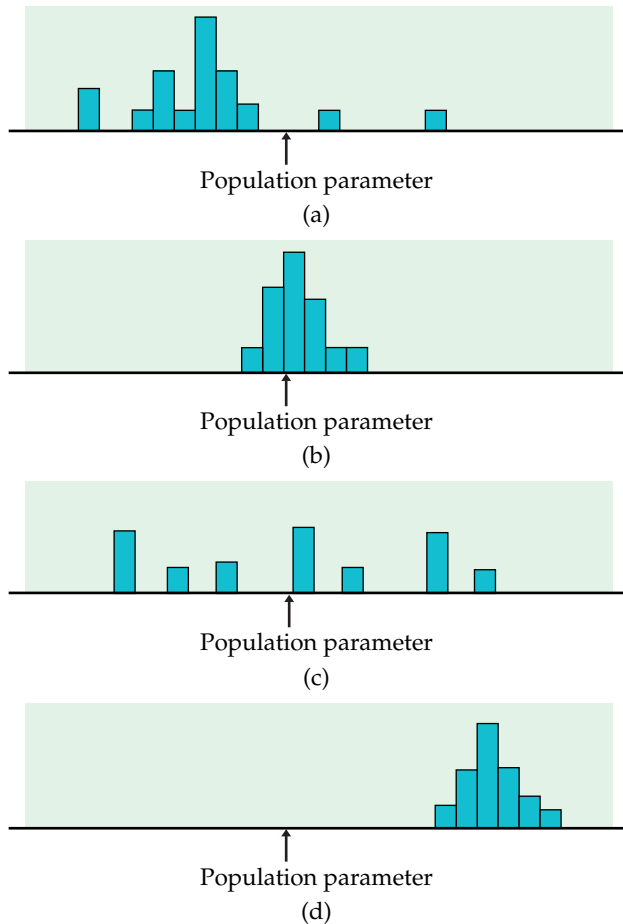


FIGURE 3.15 Determine which of these sampling distributions displays high or low bias and high or low variability, for Exercise 3.84.

percentage points for Hispanics. Sample sizes for most Hispanic nationalities, like Cubans or Dominicans, were too small to break out the results separately.⁴⁴

(a) Why is the “margin of sampling error” larger for Hispanics than for all 3092 respondents?

(b) Why would a very small sample size prevent a responsible news organization from breaking out results for Cubans?

3.86 Gallup Canada polls. Gallup Canada bases its polls of Canadian public opinion on telephone samples of about 1000 adults, the same sample size as Gallup uses in the United States. Canada’s population is about one-ninth as large as that of the United States, so the percent of adults that Gallup interviews in Canada is nine times as large as in the United States. Does this mean that the margin of error for a Gallup Canada poll is smaller? Explain your answer.

3.87 Real estate ownership. An agency of the federal government plans to take an SRS of residents in each state to estimate the proportion of owners of real estate in each state’s population. The populations of the states range from less than 500,000 people in Wyoming to about 35 million in California.

(a) Will the variability of the sample proportion vary from state to state if an SRS of size 2000 is taken in each state? Explain your answer.

(b) Will the variability of the sample proportion change from state to state if an SRS of 1/10 of 1% (0.001) of the state’s population is taken in each state? Explain your answer.

3.88 The health care system in Ontario. The Ministry of Health in the Canadian province of Ontario wants to know whether the national health care system is achieving its goals in the province. The ministry conducted the Ontario Health Survey, which interviewed a probability sample of 61,239 adults who live in Ontario.⁴⁵

(a) What is the population for this sample survey? What is the sample?

(b) The survey found that 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. Do you think these estimates are close to the truth about the entire population? Why?

The remaining exercises demonstrate the idea of a sampling distribution. Sampling distributions are the basis for statistical inference. We strongly recommend doing some of these exercises.

3.89



Use the probability applet. The Probability

applet simulates tossing a coin, with the advantage that you can choose the true long-term proportion, or probability, of a head. Example 3.33 discusses sampling from a population in which proportion $p = 0.6$ (the parameter) find shopping frustrating. Tossing a coin with probability $p = 0.6$ of a head simulates this situation: each head is a person who finds shopping frustrating, and each tail is a person who does not. Set the “Probability of heads” in the applet to 0.6 and the number of tosses to 25. This simulates an SRS of size 25 from this population. By alternating between “Toss” and “Reset” you can take many samples quickly.

(a) Take 50 samples, recording the number of heads in each sample. Make a histogram of the 50 sample proportions (count of heads divided by 25). You are constructing the sampling distribution of this statistic.

(b) Another population contains only 20% who approve of legal gambling. Take 50 samples of size 25 from this population, record the number in each sample who approve, and make a histogram of the 50 sample proportions. How do the centers of your two histograms reflect the differing truths about the two populations?

3.90



Use the statistical software for simulations.

Statistical software can speed simulations. We are interested in the sampling distribution of the proportion \hat{p} of people who find shopping frustrating in an SRS from a population in which proportion p find shopping frustrating. Here, p is a parameter and \hat{p} is a statistic used to estimate p . We will see in Chapter 5 that “binomial” is the key word to look for in the software menus. For example, in CrunchIt! go to “Simulate data” in the “Data” menu, and choose “Binomial.”

(a) Set $n = 50$ and $p = 0.6$ and generate 100 binomial observations. These are the counts for 100 SRSs of size 50 when 60% of the population finds shopping frustrating. Save these counts and divide them by 50 to get values of \hat{p} from 100 SRSs. Make a stemplot of the 100 values of \hat{p} .

(b) Repeat this process with $p = 0.3$, representing a population in which only 30% of people find shopping frustrating. Compare your two stemplots. How does changing the parameter p affect the center and spread of the sampling distribution?

(c) Now generate 100 binomial observations with $n = 200$ and $p = 0.6$. This simulates 100 SRSs, each of size 200. Obtain the 100 sample proportions \hat{p} and make a stemplot. Compare this with your stemplot from (a). How does changing the sample size n affect the center and spread of the sampling distribution?

3.91 Use Table B for a simulation. We can construct a sampling distribution by hand in the case of a very small sample from a very small population. The population contains 10 students. Here are their scores on an exam:

Student	0	1	2	3	4	5	6	7	8	9
Score	82	62	80	58	72	73	65	66	74	62

The parameter of interest is the mean score, which is 69.4. The sample is an SRS of $n = 4$ students drawn from this population. The students are labeled 0 to 9 so that a single random digit from Table B chooses one student for the sample.

(a) Use Table B to draw an SRS of size 4 from this population. Write the four scores in your sample and calculate the mean \bar{x} of the sample scores. This statistic is an estimate of the population parameter.

(b) Repeat this process 9 more times. Make a histogram of the 10 values of \bar{x} . You are constructing the sampling distribution of \bar{x} . Is the center of your histogram close to 69.4? (Ten repetitions give only a crude approximation to the sampling distribution. If possible, pool your work with that of other students—using different parts of Table B—to obtain several hundred repetitions and make a histogram of the values of \bar{x} . This histogram is a better approximation to the sampling distribution.)

3.92



Use the simple random sample applet.

The *Simple Random Sample* applet can illustrate the idea of a sampling distribution. Form a population labeled 1 to 100. We will choose an SRS of 10 of these numbers. That is, in this exercise, the numbers themselves are the population, not just labels for 100 individuals. The mean of the whole numbers 1 to 100 is 50.5. This is the parameter, the mean of the population.

(a) Use the applet to choose an SRS of size 10. Which 10 numbers were chosen? What is their mean? This is a statistic, the sample mean \bar{x} .

(b) Although the population and its mean 50.5 remain fixed, the sample mean changes as we take more samples. Take another SRS of size 10. (Use the “Reset” button to return to the original population before taking the second sample.) What are the 10 numbers in your sample? What is their mean? This is another value of \bar{x} .

(c) Take 8 more SRSs from this same population and record their means. You now have 10 values of the sample mean \bar{x} from 10 SRSs of the same size from the same population. Make a histogram of the 10 values and mark the population mean 50.5 on the horizontal axis. Are your 10 sample values roughly centered at the population value? (If you kept going forever, your \bar{x} -values would form the sampling distribution of the sample mean; the population mean would indeed be the center of this distribution.)

3.93 Analyze simple random samples. The CSDATA data set contains the college grade point averages (GPAs) of all 224 students in a university entering class who planned to major in computer science. This is our population. Statistical software can take repeated samples to illustrate sampling variability.

(a) Using software, describe this population with a histogram and with numerical summaries. In particular, what is the mean GPA in the population? This is a parameter.

(b) Choose an SRS of 20 members from this population. Make a histogram of the GPAs in the sample and find their mean. The sample mean is a statistic. Briefly compare the distributions of GPA in the sample and in the population.

(c) Repeat the process of choosing an SRS of size 20 four more times (five in all). Record the five histograms of your sample GPAs. Does it seem reasonable to you from this small trial that an SRS will usually produce a sample that is generally representative of the population?

3.94 Simulate the sampling distribution of the mean.

Continue the previous exercise, using software to illustrate the idea of a sampling distribution.

(a) Choose 20 more SRSs of size 20 in addition to the 5 you have already chosen. Don't make histograms of these latest samples—just record the mean GPA for each sample. Make a histogram of the 25 sample means. This histogram is a rough approximation to the sampling distribution of the mean.

(b) One sign of bias would be that the distribution of the sample means was systematically on one side of the true population mean. Mark the population

mean GPA on your histogram of the 25 sample means. Is there a clear bias?

(c) Find the mean and standard deviation of your 25 sample means. We expect that the mean will be close to the true mean of the population. Is it? We also expect that the standard deviation of the sampling distribution will be smaller than the standard deviation of the population. Is it?

3.95 Toss a coin. Coin tossing can illustrate the idea of a sampling distribution. The population is all outcomes (heads or tails) we would get if we tossed a coin forever. The parameter p is the proportion of heads in this population. We suspect that p is close to 0.5. That is, we think the coin will show about one-half heads in the long run. The sample is the outcomes of 20 tosses, and the statistic \hat{p} is the proportion of heads in these 20 tosses (count of heads divided by 20).

(a) Toss a coin 20 times and record the value of \hat{p} .

(b) Repeat this sampling process 9 more times. Make a stemplot of the 10 values of \hat{p} . You are constructing the sampling distribution of \hat{p} . Is the center of this distribution close to 0.5? (Ten repetitions give only a crude approximation to the sampling distribution. If possible, pool your work with that of other students to obtain several hundred repetitions and make a histogram of the values of \hat{p} .)

3.4 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

3.35 Provide all of the critical information. Papers reporting scientific research are supposed to be short, with no extra baggage. 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.”⁴⁶ 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.” 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. When the government temporarily stopped human-subject research at Duke University Medical Center in 1999 due to inadequate protection of subjects, more than 2000 studies were going on. That’s a lot of review work. 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.

3.96 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 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"?

- (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.

3.97 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

3.36 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 most medical experiments in prisons.

Very 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.⁴⁷

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. It is important to protect the subjects' privacy by keeping all data about individuals confidential. 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.

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 data bases—census responses, tax returns, Social Security information, data from surveys such as the Current Population Survey, and so on. Many of these data bases can be searched by computers for statistical studies. A clever computer search of several data bases 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 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

3.37 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

- 3.98 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.
- 3.99 Should we allow this personal information to be collected?** In which of the circumstances below would you allow collecting personal information without the subjects' consent?
- 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.
 - A social psychologist attends public meetings of a religious group to study the behavior patterns of members.
 - 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

3.38 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.⁴⁸

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 are the words 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:

*There’s a delicate balance between when to do or not do a randomized trial. On the one hand, there must be sufficient belief in the agent’s potential to justify exposing half the subjects to it. On the other hand, there must be sufficient doubt about its efficacy to justify withholding it from the other half of subjects who might be assigned to placebos.*⁴⁹

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. So in the state of balanced doubt described by Dr. Hennekens, 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