3.1 Sources of Data

ANECDOTAL DATA

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

AVAILABLE DATA

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

Sample surveys and experiments

sample surveys

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

sample

population

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

census

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

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

OBSERVATION VERSUS EXPERIMENT

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

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

intervention

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

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

SECTION 3.1 SUMMARY

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

3.2 Design of Experiments

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

EXPERIMENTAL UNITS, SUBJECTS, TREATMENTS, AND OUTCOMES

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

factors

level of a factor

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

response variables is important. The explanatory variables 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.

Comparative experiments

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

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 outside the laboratory or with living subjects, such simple designs often yield invalid data. That is, we cannot tell whether the response was due to the treatment or to lurking variables.

placebo effect

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

comparative experiment

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

control group

treatment group

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

Uncontrolled experiments (that is, experiments that don't include a control group) in medicine and the behavioral sciences can be dominated by such influences as the

details of the experimental arrangement, the selection of subjects, and the placebo effect. The result is often *bias.*

BIAS

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

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

Randomization

experimental design

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

matching

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

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

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

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

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

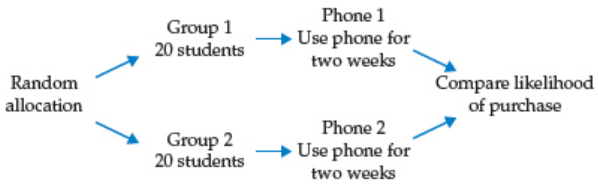

FIGURE 3.3 Outline of a randomized comparative experiment, Example 3.11.

randomization

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

Randomized comparative experiments

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

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

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

PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles of statistical design of experiments are

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

How to randomize

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

Randomization using random digits

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

RANDOM DIGITS

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

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

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

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

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

EXAMPLE 3.13

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

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

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

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

80011 09937 57195 33906 94831 10056 42211 65491

The first 10 two-digit groups in this line are

80 01 10 99 37 57 19 53 39 06

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

completely randomized design

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

EXAMPLE 3.14

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

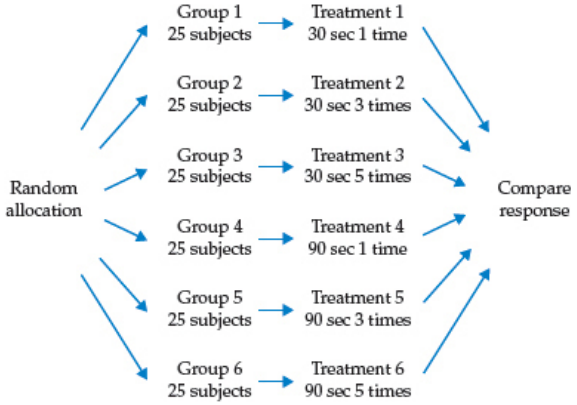

FIGURE 3.5 Outline of a completely randomized design comparing six treatments, Example 3.14.

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

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

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

Cautions about experimentation

double-blind

The logic of a randomized comparative experiment depends on our ability to treat all the experimental units identically in every way except for the actual treatments being compared. Good experiments, therefore, require careful attention to details. The ideal situation is where a study is **double-blind**—neither the subjects themselves nor the experimenters know which treatment any subject has received. The double-blind method avoids unconscious bias by, for example, a doctor who doesn't think that "just a placebo" can benefit a patient.

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

EXAMPLE 3.15

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

lack of realism

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

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

Matched pairs designs

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

matched pairs design

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

EXAMPLE 3.17

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

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

cross-over

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

Block designs

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

BLOCK DESIGN

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

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

EXAMPLE 3.18

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

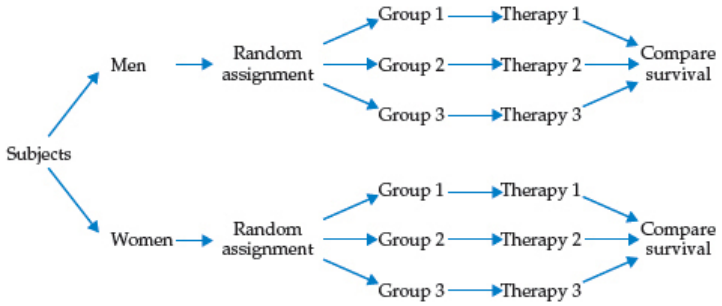

FIGURE 3.6 Outline of a block design, Example 3.18. The blocks consist of male and female subjects. The treatments are the three therapies for cancer.

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

EXAMPLE 3.20

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

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

SECTION 3.2 SUMMARY

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

- The basic principles of statistical design of experiments are **compare**, **randomization**, and **repetition**.
- The simplest form of control is **comparison**. Experiments should compare two or more treatments in order to prevent **confounding** the effect of a treatment with other influences, such as lurking variables.
- **Randomization** uses chance to assign subjects to the treatments. Randomization creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias**, or systematic favoritism, in experiments.
- You can carry out randomization by giving numerical labels to the experimental units and using a **table of random digits** to choose treatment groups.
- **Repetition** of the treatments on many units reduces the role of chance variation and makes the experiment more sensitive to differences among the treatments.
- Good experiments require attention to detail as well as good statistical design. Many behavioral and medical experiments are **double-blind**. **Lack of realism** in an experiment can prevent us from generalizing its results.
- In addition to comparison, a second form of control is to restrict randomization by forming **blocks** of experimental units that are similar in some way that is important to the response. Randomization is then carried out separately within each block.
- **Matched pairs** are a common form of blocking for comparing just two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, the subjects are matched in pairs as closely as possible, and one subject in each pair receives each treatment.

3.3 Sampling Design

sample survey

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

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.

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%.

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

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

VOLUNTARY RESPONSE SAMPLE

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

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

Simple random samples

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

SIMPLE RANDOM SAMPLE

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

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

How to select a simple random sample

The basic ideas needed to select a simple random sample are very similar to those that we discussed when we randomized subjects to treatments (page 177). We first assign a

label to each case in our population. Then we perform the randomization using software or random digits from Table B.

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

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

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

EXAMPLE 3.24

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

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

68417 35013 15529 72765 85089 57067 50211 47487

If we arrange these into sets of three, we have

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

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

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

Stratified random samples

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

PROBABILITY SAMPLE

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

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

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

STRATIFIED RANDOM SAMPLE

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

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

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

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

Multistage random samples

multistage random sample

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

clusters

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

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

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

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

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

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

Cautions about sample surveys

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

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

UNDERCOVERAGE AND NONRESPONSE

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

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

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

response bias

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

wording of questions

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

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.*

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*.

SECTION 3.3 SUMMARY

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

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.

Institutional review boards

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

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

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

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.

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

confidentiality

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

anonymity

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

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

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

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

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.

Behavioral and social science experiments

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

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

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

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

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

SECTION 3.4 SUMMARY

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