PART II

Producing Data:

Samples, Experiments, and Simulations

RONALD A. FISHER

The Father of Statistics

The ideas and methods that we study as "statistics" were invented in the nineteenth and twentieth centuries by people working on problems that required analysis of data. Astronomy, biology, social science, and even surveying can claim a role in the birth of statistics. But if anyone can claim to be "the father of statistics," that honor belongs to *Sir Ronald A. Fisher (1890–1962)*.

Fisher's writings helped organize statistics as a distinct field of study whose methods apply to practical problems across many disciplines. He systematized the mathematical theory of statistics and invented many new techniques. The randomized comparative experiment is perhaps Fisher's greatest contribution.

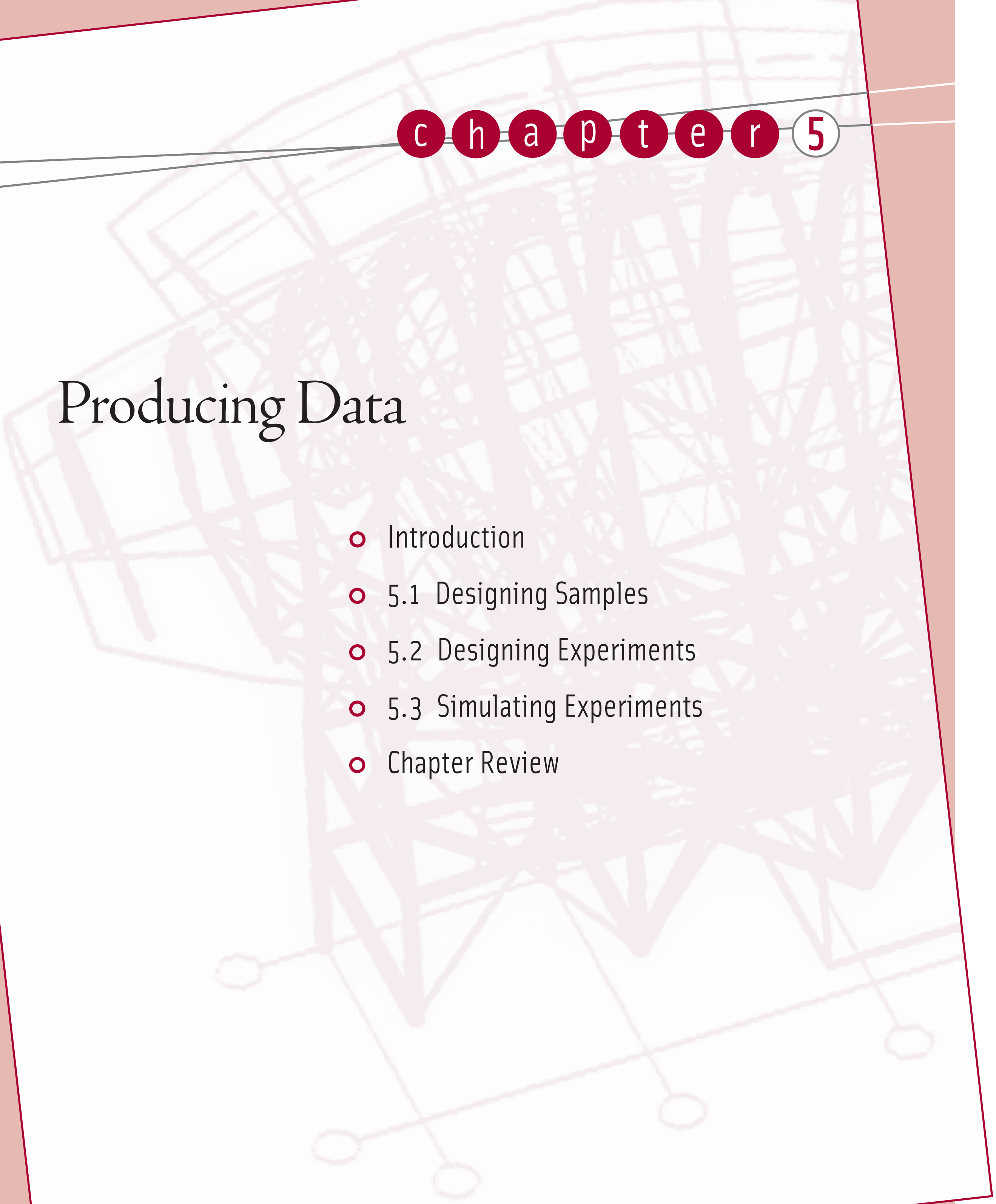
Like other statistical pioneers, Fisher was driven by the demands of practical problems. Beginning in 1919, he worked on agricultural field experiments at Rothamsted in England. How should we arrange the planting of different crop varieties or the application of different fertilizers to get a fair comparison among them? Because fertility and other variables change as we move across a field, experiments used elaborate checkerboard planting arrangements to obtain fair comparisons. Fisher had a better idea: "arrange the plots deliberately at random."

This chapter explores statistical design for producing data to answer specific questions like "Which crop variety has the highest mean yield?" Fisher's innovation, the deliberate use of chance in producing data, is the central theme of the chapter and one of the most important ideas in statistics.

Like other statistical pioneers, Fisher was driven by the demands of practical problems.

chapter 5

Producing Data



- Introduction
- 5.1 Designing Samples
- 5.2 Designing Experiments
- 5.3 Simulating Experiments
- Chapter Review

ACTIVITY 5A A Class Survey

A class survey is a quick way to collect interesting data. Certainly there are things about the class as a group that you would like to know. Your task here is to construct a *draft* of a class survey, a questionnaire that would be used to gather data about the members of your class. Here are the steps to take:

1. As a class, discuss the questions you would like to include on the survey. In addition to *what* you want to ask, you should also consider *how many questions* you want to ask. Have one student serve as recorder and make a list on the blackboard or overhead projector of topics to include.
2. Once you have identified the topics, then work on the wording of the questions. Try to achieve as much consensus as possible. If there is a computer in the room, a student could use a word-processing program to enter the questions as they are developed.
3. Make one copy of the final draft of the survey for each student, but do not distribute the surveys at this time. The surveys are to be put aside for the time being. As you complete this chapter, you will return to take another look at the survey you have constructed, make final adjustments, and then administer the survey to all of the members of your class. This survey should provide some interesting data that can be analyzed during the remainder of the course.

As a starting point, here is a sample of a short survey:

CLASS SURVEY

Your answers to the questions below will help describe your class. DO NOT PUT YOUR NAME ON THIS PAPER. Your answers are completely private. They just help us describe the entire class.

1. Are you MALE or FEMALE? (Circle one.)
2. How many brothers and sisters do you have? ________
3. How tall are you in inches, to the nearest inch? ________
4. Estimate the number of pairs of shoes you own. ________
5. How much money in coins are you carrying right now? (Don't count any paper money, just coins.) ________
6. On a typical school night, how much time do you spend doing homework? (Answer in minutes. For example, 2 hours is 120 minutes.) ________
7. On a typical school night, how much time do you spend watching television? (Answer in minutes.) ________

INTRODUCTION

Exploratory data analysis seeks to discover and describe what data say by using graphs and numerical summaries. The conclusions we draw from data analysis apply to the specific data that we examine. Often, however, we want to answer questions about some large group of individuals. To get sound answers, we must produce data in a way that is designed to answer our questions.

Suppose our question is "What percent of American adults agree that the United Nations should continue to have its headquarters in the United States?" To answer the question, we interview American adults. We can't afford to ask all adults, so we put the question to a ***sample*** chosen to represent the entire adult population. How shall we choose a sample that truly represents the opinions of the entire population? Statistical designs for choosing samples are the topic of Section 5.1.

sample

Our goal in choosing a sample 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*. In other settings, we gather data from an *experiment*. In doing an experiment, we don't just observe individuals or ask them questions. We actively impose some treatment in order to observe the response. Experiments can answer questions such as "Does aspirin reduce the chance of a heart attack?" and "Does a majority of college students prefer Pepsi to Coke when they taste both without knowing which they are drinking?" Experiments, like samples, provide useful data only when properly designed. We will discuss statistical design of experiments in Section 5.2. The distinction between experiments and observational studies is one of the most important ideas in statistics.

OBSERVATION VERSUS EXPERIMENT

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.

An **experiment,** on the other hand, deliberately imposes some treatment on individuals in order to observe their responses.

Observational studies are essential sources of data about topics from the opinions of voters to the behavior of animals in the wild. But an observational study, even one based on a statistical sample, is a poor way to gauge the effect of an intervention. To see the response to a change, we must actually impose the change. When our goal is to understand cause and effect, experiments are the only source of fully convincing data.

EXAMPLE 5.1 HELPING WELFARE MOTHERS FIND JOBS

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

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

When we simply observe welfare mothers, the effect of job-training programs on success in finding work is *confounded* with (mixed up with) the characteristics of mothers who seek out training on their own. Recall that two variables (explanatory variables or lurking variables) are said to be **confounded** when their effects on a response variable cannot be distinguished from each other.

Observational studies of the effect of one variable on another often fail because the explanatory variable is confounded with lurking variables. We will see that well-designed experiments take steps to defeat confounding. Because experiments allow us to pin down the effects of specific variables of interest to us, they are the preferred method of gaining knowledge in science, medicine, and industry.

simulation

In some situations, it may not be possible to observe individuals directly or to perform an experiment. In other cases, it may be logistically difficult or simply inconvenient to obtain a sample or to impose a treatment. ***Simulations*** provide an alternative method for producing data in such circumstances. Section 5.3 introduces techniques for simulating experiments.

statistical inference

Statistical techniques for producing data open the door to formal ***statistical inference***, which answers specific questions with a known degree of confidence. The later chapters of this book are devoted to inference. We will see that careful design of data production is the most important prerequisite for trustworthy inference.

5.1 DESIGNING SAMPLES

A political scientist wants to know what percent of the voting-age population 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. Time, cost, and inconvenience forbid contacting every individual. In such cases, we gather information about only part of the group in order to draw conclusions about the whole.

POPULATION AND SAMPLE

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

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

Notice that "population" is defined in terms of our desire for knowledge. If we wish to draw conclusions about all U.S. college students, that group is our population even if only local students are available for questioning. The sample is the part from which we draw conclusions about the whole. *Sampling* and conducting a *census* are two distinct ways of collecting data.

SAMPLING VERSUS A CENSUS

Sampling involves studying a part in order to gain information about the whole.

A **census** attempts 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.

sample design

If conclusions based on a sample are to be valid for the entire population, a sound design for selecting the sample is required. The ***design*** of a sample refers to the method used to choose the sample from the population. Poor sample designs can produce misleading conclusions, as the following examples illustrate.

EXAMPLE 5.2 CALL-IN OPINION POLLS

Television news programs like to conduct call-in polls of public opinion. The program announces a question and asks viewers to call one telephone number to respond "Yes" and another for "No." Telephone companies charge for these calls. The ABC network program *Nightline* once asked whether the United Nations should continue to have its headquarters in the United States. More than 186,000 callers responded, and 67% said "No."

People who spend the time and money to respond to call-in polls are not representative of the entire adult population. In fact, they tend to be the same people who call radio talk shows. People who feel strongly, especially those with strong negative opinions, are more likely to call. It is not surprising that a properly designed sample showed that 72% of adults want the UN to stay.[1]

Call-in opinion polls are an example of *voluntary response sampling*. A voluntary response sample can easily produce 67% "No" when the truth about the population is close to 72% "Yes."

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.

convenience sampling

Voluntary response is one common type of bad sample design. Another is ***convenience sampling***, which chooses the individuals easiest to reach. Here is an example of convenience sampling.

EXAMPLE 5.3 INTERVIEWING AT THE MALL

Manufacturers and advertising agencies often use interviews at shopping malls to gather information about the habits of consumers and the effectiveness of ads. A sample of mall shoppers is fast and cheap. "Mall interviewing is being propelled primarily as a budget issue," one expert told the *New York Times*. But people contacted at shopping malls are not representative of the entire U.S. population. They are richer, for example, and more likely to be teenagers or retired. Moreover, mall interviewers tend to select neat, safe-looking individuals from the stream of customers. Decisions based on mall interviews may not reflect the preferences of all consumers.[2]

Both voluntary response samples and convenience samples choose a sample that is almost guaranteed not to represent the entire population. These sampling methods display *bias*, or systematic error, in favoring some parts of the population over others.

BIAS

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

EXERCISES

5.1 FUNDING FOR DAY CARE A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of the 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Only 48 questionnaires are returned. What is the population in this study? What is the sample?

5.2 WHAT IS THE POPULATION? For each of the following sampling situations, identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not complete, complete the description of the population in a reasonable way.

(a) Each week, the Gallup Poll questions a sample of about 1500 adult U.S. residents to determine national opinion on a wide variety of issues.

(b) The 2000 census tried to gather basic information from every household in the United States. But a "long form" requesting much additional information was sent to a sample of about 17% of households.

(c) A machinery manufacturer purchases voltage regulators from a supplier. There are reports that variation in the output voltage of the regulators is affecting the performance of the finished products. To assess the quality of the supplier's production, the manufacturer sends a sample of 5 regulators from the last shipment to a laboratory for study.

5.3 TEACHING READING An educator wants to compare the effectiveness of computer software that teaches reading with that of a standard reading curriculum. He tests the reading ability of each student in a class of fourth graders, then divides them into two groups. One group uses the computer regularly, while the other studies a standard curriculum. At the end of the year, he retests all the students and compares the increase in reading ability in the two groups. Is this an experiment? Why or why not? What are the explanatory and response variables?

5.4 THE EFFECTS OF PROPAGANDA In 1940, a psychologist conducted an experiment to study the effect of propaganda on attitude toward a foreign government. He administered a test of attitude toward the German government to a group of American students. After the students read German propaganda for several months, he tested them again to see if their attitudes had changed.

Unfortunately, Germany attacked and conquered France while the experiment was in progress. Explain clearly why confounding makes it impossible to determine the effect of reading the propaganda.

5.5 ALCOHOL AND HEART ATTACKS Many studies have found that people who drink alcohol in moderation have lower risk of heart attacks than either nondrinkers or heavy drinkers. Does alcohol consumption also improve survival after a heart attack? One study followed 1913 people who were hospitalized after severe heart attacks. In the year before their heart attack, 47% of these people did not drink, 36% drank moderately, and 17% drank heavily. After four years, fewer of the moderate drinkers had died.[3] Is this an observational study or an experiment? Why? What are the explanatory and response variables?

5.6 ARE ANESTHETICS SAFE? The National Halothane Study was a major investigation of the safety of anesthetics used in surgery. Records of over 850,000 operations performed in 34 major hospitals showed the following death rates for four common anesthetics:[4]

Anesthetic:	A	B	C	D
Death rate:	1.7%	1.7%	3.4%	1.9%

There is a clear association between the anesthetic used and the death rate of patients. Anesthetic C appears to be dangerous.

(a) Explain why we call the National Halothane Study an observational study rather than an experiment, even though it compared the results of using different anesthetics in actual surgery.

(b) When the study looked at other variables that are confounded with a doctor's choice of anesthetic, it found that Anesthetic C was not causing extra deaths. Suggest several variables that are mixed up with what anesthetic a patient receives.

5.7 CALL THE SHOTS A newspaper advertisement for *USA Today: The Television Show* once said:

Should handgun control be tougher? You call the shots in a special call-in poll tonight. If yes, call 1-900-720-6181. If no, call 1-900-720-6182. Charge is 50 cents for the first minute.

Explain why this opinion poll is almost certainly biased.

5.8 EXPLAIN IT TO THE CONGRESSWOMAN You are on the staff of a member of Congress who is considering a bill that would provide government-sponsored insurance for nursing home care. You report that 1128 letters have been received on the issue, of which 871 oppose the legislation. "I'm surprised that most of my constituents oppose the bill. I thought it would be quite popular," says the congresswoman. Are you convinced that a majority of the voters oppose the bill? How would you explain the statistical issue to the congresswoman?

Simple random samples

In a voluntary response sample, people choose whether to respond. In a convenience sample, the interviewer makes the choice. In both cases, personal choice produces bias. The statistician's remedy is to allow impersonal chance to choose the sample. A sample chosen by chance allows neither favoritism by the sampler nor self-selection by respondents. Choosing a sample by chance attacks bias by giving all individuals an equal chance to be chosen. Rich and poor, young and old, black and white, all have the same chance to be in the sample.

The simplest way to use chance to select a sample is to place names in a hat (the population) and draw out a handful (the sample). This is the idea of *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.

An SRS not only gives each individual an equal chance to be chosen (thus avoiding bias in the choice) but also 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. Exercise 5.30 describes one such design, called systematic random sampling.

The idea of an SRS is to choose our sample by drawing names from a hat. In practice, computer software can choose an SRS almost instantly from a list of the individuals in the population. If you don't use software, you can randomize by using a *table of random digits.*

RANDOM DIGITS

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these two properties:

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.

2. The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Table B at the back of the book is a table of random digits. 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 randomly chosen digits. Because the digits in Table B are random:

- Each entry is equally likely to be any of the 10 possibilities 0, 1, . . . , 9.
- Each pair of entries is equally likely to be any of the 100 possible pairs 00, 01, . . . , 99.
- Each triple of entries is equally likely to be any of the 1000 possibilities 000, 001, . . . , 999, and so on.

These “equally likely” facts make it easy to use Table B to choose an SRS. Here is an example that shows how.

EXAMPLE 5.4 HOW TO CHOOSE AN SRS

Joan’s small accounting firm serves 30 business clients. Joan wants to interview a sample of 5 clients in detail to find ways to improve client satisfaction. To avoid bias, she chooses an SRS of size 5.

Step 1: **Label.** Give each client a numerical label, using as few digits as possible. Two digits are needed to label 30 clients, so we use labels

01, 02, 03, . . . , 29, 30

It is also correct to use labels 00 to 29 or even another choice of 30 two-digit labels. Here is the list of clients, with labels attached:

01	A-1 Plumbing	16	JL Records
02	Accent Printing	17	Johnson Commodities
03	Action Sport Shop	18	Keiser Construction
04	Anderson Construction	19	Liu’s Chinese Restaurant
05	Bailey Trucking	20	MagicTan
06	Balloons Inc.	21	Peerless Machine
07	Bennett Hardware	22	Photo Arts
08	Best’s Camera Shop	23	River City Books
09	Blue Print Specialties	24	Riverside Tavern
10	Central Tree Service	25	Rustic Boutique
11	Classic Flowers	26	Satellite Services
12	Computer Answers	27	Scotch Wash
13	Darlene’s Dolls	28	Sewer’s Center
14	Fleisch Realty	29	Tire Specialties
15	Hernandez Electronics	30	Von’s Video Store

Step 2: **Table.** Enter Table B anywhere and read two-digit groups. Suppose we enter 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 successive two-digit group is a label. The labels 00 and 31 to 99 are not used in this example, so we ignore them. The first 5 labels between 01 and 30 that we encounter in the table choose our sample. Of the first 10 labels in line 130, we ignore 5 because they are too high (over 30). The others are 05, 16, 17, 17, and 17. The clients labeled 05, 16, and 17 go into the sample. Ignore the second and third 17s because that client is already in the sample. Now run your finger across line 130 (and continue to line 131 if needed) until 5 clients are chosen.

The sample is the clients labeled 05, 16, 17, 20, 19. These are Bailey Trucking, JL Records, Johnson Commodities, MagicTan, and Liu's Chinese Restaurant.

CHOOSING AN SRS

Choose an SRS in two steps:

Step 1: **Label.** Assign a numerical label to every individual in the population.

Step 2: **Table.** Use Table B to select labels at random.

You can assign labels in any convenient manner, such as alphabetical order for names of people. Be certain that all labels have the same number of digits. Only then will all individuals have the same chance to be chosen. Use the shortest possible labels: one digit for a population of up to 10 members, 2 digits for 11 to 100 members, three digits for 101 to 1000 members, and so on. As standard practice, we recommend that you begin with label 1 (or 01 or 001, as needed). You can read digits from Table B in any order—across a row, down a column, and so on—because the table has no order. As standard practice, we recommend reading across rows.

Other sampling designs

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. If all individuals in each stratum are identical, for example, just one individual from each stratum is enough to completely describe the population.

EXAMPLE 5.5 WHO WROTE THAT SONG?

A radio station that broadcasts a piece of music owes a royalty to the composer. The organization of composers (called ASCAP) collects these royalties for all its members by charging stations a license fee for the right to play members' songs. ASCAP has four million songs in its catalog and collects $435 million in fees each year. How should ASCAP distribute this income among its members? By sampling: ASCAP tapes about 60,000 hours from the 53 million hours of local radio programs across the country each year.

Radio stations are stratified by type of community (metropolitan, rural), geographic location (New England, Pacific, etc.), and the size of the license fee paid to ASCAP, which reflects the size of the audience. In all, there are 432 strata. Tapes are made at random hours for randomly selected members of each stratum. The tapes are reviewed by experts who can recognize almost every piece of music ever written, and the composers are then paid according to their popularity.[5]

Another common means of restricting random selection is to choose the sample in stages. This is usual 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 55,000 households each month. It is not practical to maintain a list of all U.S. households from which to select an SRS. Moreover, the cost of sending interviewers to the widely scattered households in an SRS would be too high. The Current Population Survey therefore uses a ***multistage sampling design.*** The final sample consists of clusters of near-

multistage sample

by households 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:[6]

Stage 1: Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. Select a sample of 756 PSUs. This sample includes the 428 PSUs with the largest population and a stratified sample of 328 of the others.

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

Stage 3: Sort the housing units in each neighborhood into clusters of four nearby units. Interview the households in a random 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.

EXERCISES

5.9 CHOOSE YOUR SAMPLE You must choose an SRS of 10 of the 440 retail outlets in New York that sell your company's products. How would you label this population? Use Table B, starting at line 105, to choose your sample.

5.10 WHO SHOULD BE INTERVIEWED? A firm wants to understand the attitudes of its minority managers toward its system for assessing management performance. Below is a list of all the firm's managers who are members of minority groups. Use Table B at line 139 to choose 6 to be interviewed in detail about the performance appraisal system.

Agarwal
Anderson
Baxter
Bonds
Bowman
Castillo
Cross
Dewald
Fernandez
Fleming
Gates
Goel
Gomez
Hernandez
Huang
Kim
Liao
Mourning
Naber
Peters
Pliego
Puri
Richards
Rodriguez
Santiago
Shen
Vega
Wang

5.11 WHO GOES TO THE CONVENTION? A club has 30 student members and 10 faculty members. The students are

Abel
Carson
Chen
David
Deming
Elashoff
Fisher
Ghosh
Griswold
Hein
Hernandez
Holland
Huber
Jimenez
Jones
Kim
Klotz
Liu
Miranda
Moskowitz
Neyman
O'Brien
Pearl
Potter
Reinmann
Santos
Shaw
Thompson
Utts
Varga

The faculty members are

Andrews
Besicovitch
Fernandez
Gupta
Kim
Lightman
Moore
Phillips
West
Yang

The club can send 4 students and 2 faculty members to a convention. It decides to choose those who will go by random selection. Use Table B, beginning at line 106, to choose a stratified random sample of 4 students and 2 faculty members.

5.12 SAMPLING BY ACCOUNTANTS Accountants often 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 *only the first* 5 accounts from each of these strata.

Cautions about sample surveys

Random selection eliminates bias in the choice of a sample from a list of the population. When the population consists of human beings, however, accurate information from a sample requires much more than a good sampling design.[7] 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 7% to 8% 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 30% 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 5.6 THE CENSUS UNDERCOUNT

Even the U.S. census, backed by the resources of the federal government, suffers from undercoverage and nonresponse. The census begins by mailing forms to every household in the country. The Census Bureau's list of addresses is incomplete, resulting in undercoverage. Despite special efforts to count homeless people (who can't be reached at any address), homelessness causes more undercoverage.

In 1990, about 35% of households that were mailed census forms did not mail them back. In New York City, 47% did not return the form. That's nonresponse. The Census Bureau sent interviewers to these households. In inner-city areas, the interviewers could not contact about one in five of the nonresponders, even after six tries.

The Census Bureau estimates that the 1990 census missed about 1.8% of the total population due to undercoverage and nonresponse. Because the undercount was greater in the poorer sections of large cities, the Census Bureau estimates that it failed to count 4.4% of blacks and 5.0% of Hispanics.[8]

For the 2000 census, the Bureau planned to replace follow-up of all nonresponders with more intense pursuit of a probability sample of nonresponding households plus a national sample of 750,000 households. The final counts would be based on comparing the national sample with the original responses. This idea was politically controversial. The Supreme Court ruled that the sampling could be used for most purposes, but not for dividing seats in Congress among the states.

response bias

In addition, 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 sample then underestimates the presence of such behavior in the population. An interviewer whose attitude suggests that some answers are more desirable than others will get these answers more often. 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 draw a "Yes" from someone who last visited a dentist 8 months ago.[9] Careful training of interviewers and careful supervision to avoid variation among the interviewers can greatly reduce response bias. Good interviewing technique is another aspect of a well-done sample survey.

wording effects

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 two examples.

EXAMPLE 5.7 SHOULD WE BAN DISPOSABLE DIAPERS?

A survey paid for by makers of disposable diapers found that 84% of the sample opposed banning disposable diapers. Here is the actual question:

> *It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?*[10]

This question gives information on only one side of an issue, then asks an opinion. That's a sure way to bias the responses. A different question that described how long disposable diapers take to decay and how many tons they contribute to landfills each year would draw a quite different response.

EXAMPLE 5.8 DOUBTING THE HOLOCAUST

An opinion poll conducted in 1992 for the American Jewish Committee asked: "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" When 22% of the sample said "possible," the news media wondered how so many Americans could be uncertain that the Holocaust happened. Then a second poll asked the question in different words: "Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?" Now only 1% of the sample said "possible." The complicated wording of the first question confused many respondents.[11]

Never trust the results of a sample survey until you have read the exact questions posed. The sampling design, the amount of nonresponse, and the date of the survey are also important. Good statistical design is a part, but only a part, of a trustworthy survey.

Inference about the population

Despite the many practical difficulties in carrying out a sample survey, using chance to choose a sample does eliminate bias in the actual selection of the sample from the list of available individuals. But it is unlikely that results from a sample are exactly the same as for the entire population. Sample results, like the official unemployment rate obtained from the monthly Current Population Survey, are only estimates of the truth about the population. If we select two samples at random from the same population, we will draw different individuals. So the sample results will almost certainly differ somewhat. Two runs of the Current Population Survey would produce somewhat different unemployment rates. Properly designed samples avoid systematic bias, but their results are rarely exactly correct and they vary from sample to sample.

How accurate is a sample result like the monthly unemployment rate? We can't say for sure, because the result would be different if we took another sample. But the results of random sampling don't change haphazardly from sample to sample. Because we deliberately use chance, the results obey the laws of ***probability*** that govern chance behavior. We can say how large an error we are likely to make in drawing conclusions about the population from a sample. Results from a sample survey usually come with a margin of error that sets bounds on the size of the likely error. How to do this is part of the business of statistical inference. We will describe the reasoning in Chapter 10.

probability

One point is worth making now: **larger random samples give more accurate results than smaller samples**. By taking a very large sample, you can be confident that the sample result is very close to the truth about the population. The Current Population Survey's sample of 50,000 households estimates the national unemployment rate very accurately. Of course, only probability samples carry this guarantee. *Nightline's* voluntary response sample is worthless even though 186,000 people called in. 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.

EXERCISES

5.13 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.

sampling frame

(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 more common 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 exchange (the first three digits). Which of the households you mentioned in your answer to (a) will be included in the sampling frame by random digit dialing?

5.14 RING-NO-ANSWER 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.[12] 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.

5.15 QUESTION WORDING During the 2000 presidential campaign, the candidates debated what to do with the large government 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?

5.16 GRADING THE PRESIDENT A newspaper article about an opinion poll says that "43% of Americans approve of the president's overall job performance." Toward the end of the article, you read: "The poll is based on telephone interviews with 1210 adults from around the United States, excluding Alaska and Hawaii." What variable did this poll measure? What population do you think the newspaper wants information about? What was the sample? Are there any sources of bias in the sampling method used?

5.17 EQUAL PAY FOR MALE AND FEMALE ATHLETES? The Excite Poll can be found online at http://lite.excite.com. The question appears on the screen, and you simply click buttons to vote "Yes," "No," or "Not sure." On January 25, 2000, the question was "Should female athletes be paid the same as men for the work they do?" In all, 13,147 (44%) said "Yes," another 15,182 (50%) said "No," and the remaining 1448 said "Not sure."

(a) What is the sample size for this poll?

(b) That's a much larger sample than standard sample surveys. In spite of this, we can't trust the result to give good information about any clearly defined population. Why?

(c) More men than women use the Web. How might this fact affect the poll results?

5.18 WORDING BIAS 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?"

SUMMARY

Data analysis is sometimes **exploratory** in nature. Exploratory analysis asks what the data tell us about the variables and their relations to each other. The conclusions of an exploratory analysis may not generalize beyond the specific data studied.

Statistical inference produces answers to specific questions, along with a statement of how confident we can be that the answer is correct. The conclusions of statistical inference are usually intended to apply beyond the individuals actually studied. Successful statistical inference usually requires **production of data** intended to answer the specific questions posed.

We can produce data intended to answer specific questions by sampling or experimentation. **Sampling** selects a part of a population of interest to represent the whole. **Experiments** are distinguished from **observational studies** such as sample surveys by the active imposition of some treatment on the subjects of the experiment.

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**.

Larger samples give more accurate results than smaller samples.

SECTION 5.1 EXERCISES

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

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

(b) A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of the 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random.

(c) The American Community Survey will contact 3 million households, including some in every county in the United States. This new Census Bureau survey will ask each household questions about their housing, economic, and social status.

5.20 THE REAGAN-CARTER ELECTION DEBATE Some television stations take quick polls of public opinion by announcing a question on the air and asking viewers to call one of two telephone numbers to register their opinion as "Yes" or "No." Telephone companies make available "900" numbers for this purpose. Dialing a 900 number results in a small charge to your telephone bill. The first major use of call-in polling was by the ABC television network in October 1980. At the end of the first Reagan-Carter presidential election debate, ABC asked its viewers which candidate won. The call-in poll proclaimed that Reagan had won the debate by a 2 to 1 margin. But a random survey by CBS News showed only a 44% to 36% margin for Reagan, with the rest undecided. Why are call-in polls likely to be biased? Can you suggest why this bias might have favored the Republican Reagan over the Democrat Carter?

5.21 TESTING CHEMICALS A manufacturer of chemicals chooses 3 from each lot of 25 containers of a reagent to test for purity and potency. Below are the control numbers stamped on the bottles in the current lot. Use Table B at line 111 to choose an SRS of 3 of these bottles.

A1096	A1097	A1098	A1101	A1108
A1112	A1113	A1117	A2109	A2211
A2220	B0986	B1011	B1096	B1101
B1102	B1103	B1110	B1119	B1137
B1189	B1223	B1277	B1286	B1299

5.22 INCREASING SAMPLE SIZE Just before a presidential election, a national opinion polling firm increases the size of its weekly sample from the usual 1500 people to 4000 people. Why do you think the firm does this?

5.23 CENSUS TRACT Figure 5.1 is a map of a census tract in a fictitious town. Census tracts are small, homogeneous areas averaging 4000 in population. On the map, each block is marked with a Census Bureau identification number. An SRS of blocks from

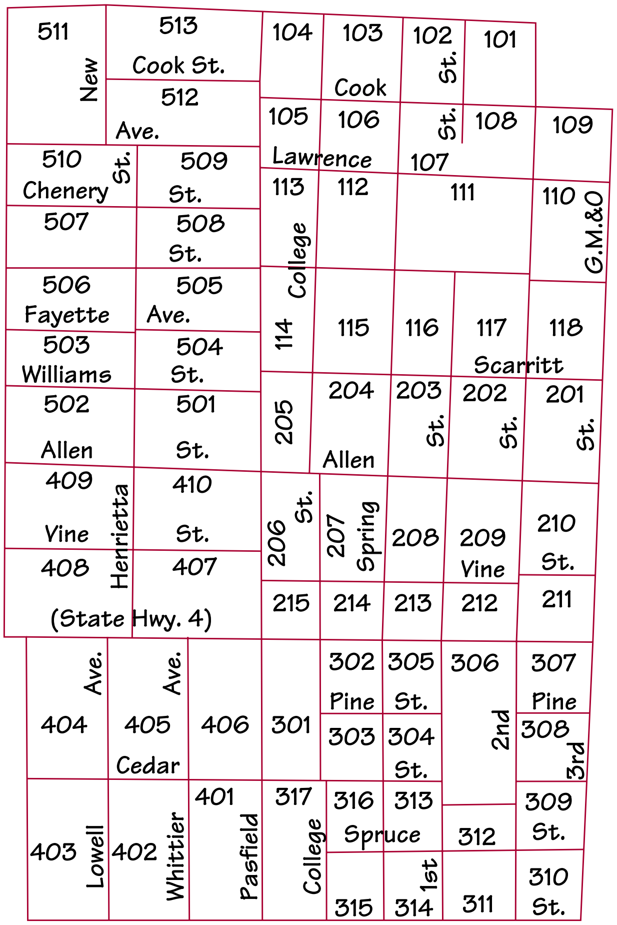


FIGURE 5.1 Map of a census tract.

a census tract is often the next-to-last stage in a multistage sample. Use Table B, beginning at line 125, to choose an SRS of 5 blocks from this census tract.

5.24 RANDOM DIGITS Which of the following statements are true of a table of random digits, and which are false? Briefly 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.

5.25 IS IT AN SRS? A corporation employs 2000 male and 500 female engineers. A stratified random sample of 200 male and 50 female engineers gives each engineer 1 chance in 10 to be chosen. This sample design gives every individual in the population the same chance to be chosen for the sample. Is it an SRS? Explain your answer.

5.26 CHECKING FOR BIAS Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

(a) Which of the following best represents your opinion on gun control?

1. The government should confiscate our guns.
2. We have the right to keep and bear arms.

(b) A freeze in nuclear weapons should be favored because it would begin a much-needed process to stop everyone in the world from building nuclear weapons now and reduce the possibility of nuclear war in the future. Do you agree or disagree?

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

5.27 SAMPLING ERROR A *New York Times* opinion poll on women's issues contacted a sample of 1025 women and 472 men by randomly selecting telephone numbers. The *Times* publishes complete descriptions of its polling methods. Here is part of the description for this poll:[13]

> *In theory, in 19 cases out of 20 the results based on the entire sample will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all adult Americans.*
>
> *The potential sampling error for smaller subgroups is larger. For example, for men it is plus or minus five percentage points.*

Explain why the margin of error is larger for conclusions about men alone than for conclusions about all adults.

5.28 ATTITUDES TOWARD ALCOHOL 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 the chance? Why is your sample not an SRS?

5.29 WHAT DO SCHOOLKIDS WANT? What are the most important goals of schoolchildren? Do girls and boys have different goals? Are goals different in urban, suburban, and rural areas? To find out, researchers wanted to ask children in the fourth, fifth, and sixth grades this question:

What would you most like to do at school?

A. *Make good grades.*

B. *Be good at sports.*

C. *Be popular.*

Because most children live in heavily populated urban and suburban areas, an SRS might contain few rural children. Moreover, it is too expensive to choose children at random from a large region---we must start by choosing schools rather than children. Describe a suitable sample design for this study and explain the reasoning behind your choice of design.

5.30 SYSTEMATIC RANDOM SAMPLE Sample surveys often use a ***systematic random sample*** 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.

systematic random 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 addresses at random using Table B. The sample contains this address and the addresses 25, 50, and 75 places down the list from it. If the table gives 13, for example, then the systematic random sample consists of the addresses numbered 13, 38, 63, and 88.

(a) Use Table B to choose a systematic random sample of 5 addresses from a list of 200. Enter the table at line 120.

(b) Like an SRS, a systematic random sample gives all individuals the same chance to be chosen. Explain why this is true. Then explain carefully why a systematic sample is nonetheless *not* an SRS.

Activity 5B The Class Survey Revisited

Each student should have a copy of the survey that the class constructed in Activity 5A at the beginning of the chapter. Now that you are experts on good and bad characteristics of survey questions, do the following:

1. Consider the questions in order. As you look at each item, see if the question contains bias. Does it advocate a position? Does the question contain any complicated words or phrasing that might be misinterpreted? Will any questions evoke response bias?

2. Make any changes that the group feels are needed. Remember that the survey should be *anonymous* (no names on the papers) so that students are assured that the class *as a whole* rather than themselves as individuals will be described.

3. Print the final version of the survey. Make one copy for each member of the class and an extra copy on which to tally the results.

4. Each student should complete the survey.

5. Place the completed surveys, upside down, in a pile. The last student finished should shuffle the pile of surveys to ensure anonymity.

6. Designate someone (the teacher?) to tally the responses as homework and prepare a cumulative summary. Give a copy of the results to each student in the class for later analysis.

5.2 DESIGNING 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**.

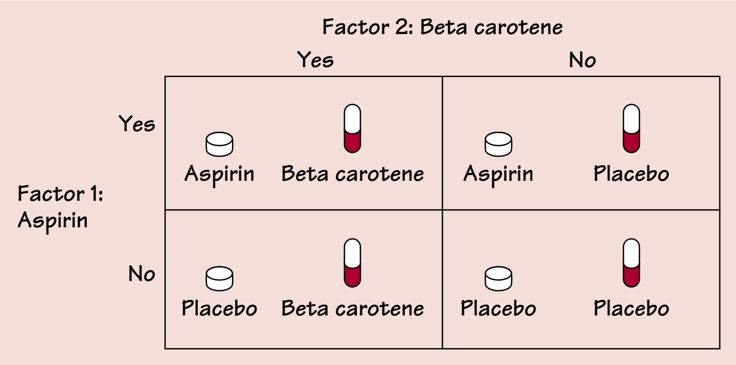
factor

level

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.

EXAMPLE 5.9 THE PHYSICIANS' HEALTH STUDY

Does regularly taking aspirin help protect people against heart attacks? The Physicians' Health Study was a medical experiment that helped answer this question. In fact, the Physicians' Health Study looked at the effects of two drugs: aspirin and beta carotene. The body converts beta carotene into vitamin A, which may help prevent some forms of cancer. The *subjects* were 21,996 male physicians. There were two *factors*, each having two levels: aspirin (yes or no) and beta carotene (yes or no). Combinations of the levels of these factors form the four *treatments* shown in Figure 5.2. One-fourth of the subjects were assigned to each of these treatments.



		Factor 2: Beta carotene	
		Yes	No
Factor 1: Aspirin	Yes	Aspirin Beta carotene	Aspirin Placebo
	No	Placebo Beta carotene	Placebo Placebo

FIGURE 5.2 The treatments in the Physicians' Health Study.

On odd-numbered days, the subjects took a white tablet that contained either aspirin or a ***placebo,*** a dummy pill that looked and tasted like the aspirin but had no active ingredient. On even-numbered days, they took a red capsule containing either beta carotene or a placebo. There were several *response variables*—the study looked for heart attacks, several kinds of cancer, and other medical outcomes. After several years, 239 of the placebo group but only 139 of the aspirin group had suffered heart attacks. This difference is large enough to give good evidence that taking aspirin does reduce heart attacks.[14] It did not appear, however, that beta carotene had any effect.

placebo

EXAMPLE 5.10 DOES STUDYING A FOREIGN LANGUAGE IN HIGH SCHOOL INCREASE VERBAL ABILITY IN ENGLISH?

Julie obtains lists of all seniors in her high school who did and did not study a foreign language. Then she compares their scores on a standard test of English reading and grammar given to all seniors. The average score of the students who studied a foreign language is much higher than the average score of those who did not.

This observational study gives no evidence that studying another language builds skill in English. Students decide for themselves whether or not to elect a foreign language. Those who choose to study a language are mostly students who are already better at English than most students who avoid foreign languages. The difference in average test scores just shows that students who choose to study a language differ (on the average) from those who do not. We can't say whether studying languages *causes* this difference.

Examples 5.9 and 5.10 illustrate the big advantage of experiments over observational studies. **In principle, experiments can give good evidence for causation.** All the doctors in the Physicians' Health Study took a pill every other day, and all got the same schedule of checkups and information. The only difference was the content of the pill. When one group had many fewer heart attacks, we conclude that it was the content of the pill that made the difference. Julie's observational study—a *census* of all seniors in her high school—does a good job of describing differences between seniors who have studied foreign languages and those who have not. But she can say nothing about cause and effect.

Another advantage of experiments is that they allow us to study the specific factors we are interested in, while controlling the effects of lurking variables. The subjects in the Physicians' Health Study were all middle-aged male doctors and all followed the same schedule of medical checkups. These similarities reduce variation among the subjects and make any effects of aspirin or beta carotene easier to see. Experiments also 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. The Physicians' Health Study tells us that aspirin helps prevent heart attacks, at least in middle-aged men, and that beta carotene taken with the aspirin neither helps nor hinders aspirin's protective powers.

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

Units → 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. Another medical example will show what can go wrong.

EXAMPLE 5.11 TREATING ULCERS

"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

Subjects → 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.[15]

control group

The first gastric freezing experiment gave misleading results because the effects of the explanatory variable were *confounded* with (mixed up 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.

Without control, experimental results 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*, systematic favoritism toward one outcome. 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, like the Physicians' Health Study and the second gastric freezing study, usually compare several treatments.

EXERCISES

For each of the experimental situations described in Exercises 5.31 to 5.34, identify the experimental units or subjects, the factors, the treatments, and the response variables.

5.31 RESISTING DROUGHT The ability to grow in shade may help pines found in the dry forests of Arizona to 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.

5.32 PACKAGE LINERS A manufacturer of food products uses package liners that are sealed at the top by applying heated jaws after the package is filled. The customer peels the sealed pieces apart to open the package. What effect does the temperature of the jaws have on the force required to peel the liner? To answer this question, the engineers prepare 20 pairs of pieces of package liner. They seal five pairs at each of 250° F, 275° F, 300° F, and 325° F. Then they measure the strength needed to peel each seal.

5.33 IMPROVING RESPONSE RATE 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?

5.34 SICKLE-CELL DISEASE Sickle-cell disease is an inherited disorder of the red blood cells that in the United States affects mostly blacks. It can cause severe pain and many complications. Can the drug hydroxyurea reduce the severe pain caused by sickle-cell disease? A study by the National Institutes of Health gave the drug to 150 sickle-cell sufferers and a placebo (a dummy medication) to another 150. The researchers then counted the episodes of pain reported by each subject.

5.35 COMPARING LEARNING METHODS An educator wants to compare the effectiveness of computer software that teaches reading with that of a standard reading curriculum.

She tests the reading ability of each student in a class of fourth graders, then divides them into two groups. One group uses the computer regularly, while the other studies a standard curriculum. At the end of the year, she retests all the students and compares the increase in reading ability in the two groups.

(a) Is this an experiment? Why or why not?

(b) What are the explanatory and response variables?

5.36 OPTIMIZING A PRODUCTION PROCESS A chemical engineer is designing the production process for a new product. The chemical reaction that produces the product may have higher or lower yield, depending on the temperature and the stirring rate in the vessel in which the reaction takes place. The engineer decides to investigate the effects of combinations of two temperatures (50° C and 60° C) and three stirring rates (60 rpm, 90 rpm, and 120 rpm) on the yield of the process. She will process two batches of the product at each combination of temperature and stirring rate.

(a) What are the experimental units and the response variable in this experiment?

(b) How many factors are there? How many treatments? Use a diagram like that in Figure 5.2 (page 290) to lay out the treatments.

(c) How many experimental units are required for the experiment?

Randomization

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 5.2 illustrates this aspect of the design of the Physicians' Health Study. 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 5.12 TESTING A BREAKFAST FOOD

A food company assesses the nutritional quality of a new "instant breakfast" product by feeding it to newly weaned male white rats. The response variable is a rat's weight gain over a 28-day period. A control group of rats eats a standard diet but otherwise receives exactly the same treatment as the experimental group.

This experiment has one factor (the diet) with two levels. The researchers use 30 rats for the experiment and so must divide them into two groups of 15. To do this in an unbiased fashion, put the cage numbers of the 30 rats in a hat, mix them up, and draw 15. These rats form the experimental group and the remaining 15 make up the control group. That is, *each group is an SRS of the available rats.* Figure 5.3 outlines the design of this experiment.

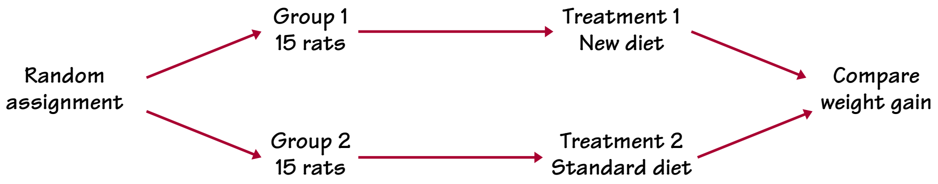


FIGURE 5.3 Outline of a randomized comparative experiment.

We can use software or the table of random digits to randomize. Label the rats 01 to 30. Enter Table B at (say) line 130. Run your finger along this line (and continue to lines 131 and 132 as needed) until 15 rats are chosen. They are the rats labeled

05, 16, 17, 20, 19, 04, 25, 29, 18, 07, 13, 02, 23, 27, 21

These rats form the experimental group; the remaining 15 are the control group.

Randomization, the use of chance to divide experimental units into groups, is an essential ingredient for a good experimental design. The design in Figure 5.3 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 generally using treatment groups about equal in size.

Randomized comparative experiments

The logic behind the randomized comparative design in Figure 5.3 is as follows:

- Randomization produces groups of rats that should be similar in all respects before the treatments are applied.

- Comparative design ensures that influences other than the diets operate equally on both groups.
- Therefore, differences in average weight gain must be due either to the diets or to the play of chance in the random assignment of rats to the two diets.

That "either-or" deserves more thought. We cannot say that *any* difference in the average weight gains of rats fed the two diets must be caused by a difference between the diets. There would be some difference even if both groups received the same diet, because the natural variability among rats means that some grow faster than others. Chance assigns the faster-growing rats to one group or the other, and this creates a chance difference between the groups. We would not trust an experiment with just one rat in each group, for example. The results would depend too much on which group got lucky and received the faster-growing rat. If we assign many rats to each diet, however, **the effects of chance will average out** and there will be little difference in the average weight gains in the two groups unless the diets themselves cause a difference. **"Use enough experimental units 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. **Control** the effects of lurking variables on the response, most simply by comparing two or more treatments.
2. **Randomize**—use impersonal chance to assign experimental units to treatments.
3. **Replicate** 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 Physicians' Health Study, for example, reported statistically significant evidence that aspirin reduces the number of heart attacks compared with a placebo.

EXAMPLE 5.13 ENCOURAGING ENERGY CONSERVATION

Many utility companies have programs to encourage their customers to conserve energy. An electric company is considering placing electronic meters in households to show what the cost would be if the electricity use at that moment continued for a month. Will meters reduce electricity use? Would cheaper methods work almost as well? The company decides to design an experiment.

One cheaper approach is to give customers a chart and information about monitoring their electricity use. The experiment compares these two approaches (meter, chart) with each other and also with a control group of customers who receive no help in monitoring electricity use. The response variable is total electricity used in a year. The company finds 60 single-family residences in the same city willing to participate, so it assigns 20 residences at random to each of the three treatments. The outline of the design appears in Figure 5.4.

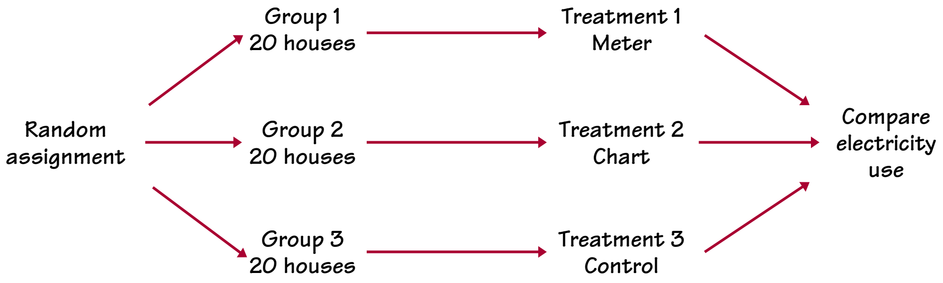


FIGURE 5.4 Outline of a completely randomized design comparing three treatments.

To carry out the random assignment, label the 60 houses 01 to 60. Then enter Table B and read two-digit groups until you have selected 20 houses to receive the meters. Continue in Table B to select 20 more to receive charts. The remaining 20 form the control group. The process is simple but tedious.

When all experimental units are allocated at random among all treatments, the experimental design is ***completely randomized.*** The designs in Figures 5.3 (page 295) and 5.4 are both completely randomized. Completely randomized designs can compare any number of treatments. In Example 5.13, we compared the three levels of a single factor: the method used to encourage energy conservation. The treatments can be formed by more than one factor. The Physicians' Health Study had two factors, which combine to form the four treatments shown in Figure 5.2 (page 290). The study

completely randomized design

used a completely randomized design that assigned 5499 of the 21,996 subjects to each of the four treatments.

EXERCISES

5.37 TREATING PROSTATE DISEASE A large study used records from Canada's national health care system to compare the effectiveness of two ways to treat prostate disease. The two treatments are traditional surgery and a new method that does not require surgery. The records described many patients whose doctors had chosen each method. The study found that patients treated by the new method were significantly more likely to die within 8 years.[16]

(a) Further study of the data showed that this conclusion was wrong. The extra deaths among patients who got the new method could be explained by lurking variables. What lurking variables might be confounded with a doctor's choice of surgical or nonsurgical treatment?

(b) You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. Use a diagram to outline the design of a randomized comparative experiment. (When using a diagram to outline the design of an experiment, be sure to indicate the size of the treatment groups and the response variable. The diagrams in Examples 5.12 (page 295) and 5.13 (page 297) are models.)

5.38 PACKAGE LINERS

(a) Use a diagram to describe a completely randomized experimental design for the package liner experiment of Exercise 5.32. (When using a diagram to outline the design of an experiment, be sure to indicate the size of the treatment groups and the response variable. The diagrams in Examples 5.12 (page 295) and 5.13 (page 297) are models.)

(b) Use Table B, starting at line 120, to do the randomization required by your design.

5.39 RECRUITING FEMALE EMPLOYEES Will providing child care for employees make a company more attractive to women, even those who are unmarried? You are designing an experiment to answer this question. You prepare recruiting material for two fictitious companies, both in similar businesses in the same location. Company A's brochure does not mention child care. There are two versions of Company B's material, identical except that one describes the company's on-site child-care facility. Your subjects are 40 unmarried women who are college seniors seeking employment. Each subject will read recruiting material for both companies and choose the one she would prefer to work for. You will give each version of Company B's brochure to half the women. You expect that a higher percentage of those who read the description that includes child care will choose Company B.

(a) Outline an appropriate design for the experiment.

(b) The names of the subjects appear below. Use Table B, beginning at line 131, to do the randomization required by your design. List the subjects who will read the version that mentions child care.

Abrams
Adamson
Afifi
Brown
Cansico
Chen
Cortez
Curzakis
Danielson
Durr
Edwards
Fluharty
Garcia
Gerson
Green
Gupta
Gutierrez
Howard
Hwang
Iselin
Janle
Kaplan
Kim
Lattimore
Lippman
Martinez
McNeill
Morse
Ng
Quinones
Rivera
Roberts
Rosen
Sugiwara
Thompson
Travers
Turing
Ullmann
Williams
Wong

5.40 ENCOURAGING ENERGY CONSERVATION Example 5.13 (page 297) describes an experiment to learn whether providing households with electronic indicators or charts will reduce their electricity consumption. An executive of the electric company objects to including a control group. He says, “It would be simpler to just compare electricity use last year (before the indicator or chart was provided) with consumption in the same period this year. If households use less electricity this year, the indicator or chart must be working.” Explain clearly why this design is inferior to that in Example 5.13.

5.41 EXERCISE AND HEART ATTACKS Does regular exercise reduce the risk of a heart attack? Here are two ways to study this question. Explain clearly why the second design will produce more trustworthy data.

1. A researcher finds 2000 men over 40 who exercise regularly and have not had heart attacks. She matches each with a similar man who does not exercise regularly, and she follows both groups for 5 years.
2. Another researcher finds 4000 men over 40 who have not had heart attacks and are willing to participate in a study. She assigns 2000 of the men to a regular program of supervised exercise. The other 2000 continue their usual habits. The researcher follows both groups for 5 years.

5.42 STOCKS DECLINE ON MONDAYS 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 of 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.*[17]

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.

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 both the Physicians' Health Study (Example 5.9, page 290) and the second gastric freezing experiment (Example 5.11, page 292) all got the same medical attention over the several years the studies continued. Moreover, these studies were ***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

DOUBLE-BLIND EXPERIMENT

In a double-blind experiment, neither the subjects nor the people who have contact with them know which treatment a subject received.

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 are some examples.

EXAMPLE 5.14 RESPONSE TO ADVERTISING

A study compares two television advertisements by showing TV programs to student subjects. The students know it's "just an experiment." We can't be sure that the results apply to everyday television viewers. Many behavioral science experiments use as subjects students who know they are subjects in an experiment. That's not a realistic setting.

EXAMPLE 5.15 CENTER BRAKE LIGHTS

Do those high center brake lights, required on all cars sold in the United States since 1986, really reduce rear-end collisions? Randomized comparative experiments with fleets of rental and business cars, done before the lights were required, showed that the third brake light reduced rear-end collisions by as much as 50%. Alas, requiring the third light in all cars led to only a 5% drop.

What happened? Most cars did not have the extra brake light when the experiments were carried out, so it caught the eye of following drivers. Now that almost all cars have the third light, they no longer capture attention.

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 the original experiment cannot tell us how far the results will generalize. 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 replication. 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.

EXAMPLE 5.16 CEREAL LEAF BEETLES

Are cereal leaf beetles more strongly attracted by the color yellow or by the color green? Agriculture researchers want to know, because they detect the presence of the pests in farm fields by mounting sticky boards to trap insects that land on them. The board color should attract beetles as strongly as possible. We must design an experiment to compare yellow and green by mounting boards on poles in a large field of oats.

The experimental units are locations within the field far enough apart to represent independent observations. We erect a pole at each location to hold the boards. We might employ a completely randomized design in which we randomly select half the poles to receive a yellow board while the remaining poles receive green. The locations vary widely in the number of beetles present. For example, the alfalfa that borders the oats on one side is a natural host of the beetles, so locations near the alfalfa will have extra beetles. This variation among experimental units can hide the systematic effect of the board color.

matched pairs design

It is more efficient to use a ***matched pairs design*** in which we mount boards of both colors on each pole. The observations (numbers of beetles trapped) are matched in pairs from the same poles. We compare the number of trapped beetles on a yellow board with the number trapped by the green board on the same pole. Because the boards are mounted one above the other, we select the color of the top board at random. Just toss a coin for each board---if the coin falls heads, the yellow board is mounted above the green board.

Matched pairs designs compare just two treatments. We choose *blocks* of two units that are as closely matched as possible. In Example 5.16, two boards on the same pole form a block. We assign one of the treatments to each unit by tossing a coin or reading odd and even digits from Table B. Alternatively, each block in a matched pairs design may consist of just one subject, who gets both treatments one after the other. Each subject serves as his or her own control. The *order* of the treatments can influence the subject's response, so we randomize the order for each subject, again by a coin toss.

Block designs

The matched pairs design of Example 5.16 uses the principles of comparison of treatments, randomization, and replication on several experimental units. However, the randomization is not complete (all locations randomly assigned to treatment groups) but restricted to assigning the order of the boards at each location. The matched pairs design reduces the effect of variation among locations in the field by comparing the pair of boards at each location. Matched pairs are an example of *block designs*.

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 5.17 COMPARING CANCER THERAPIES

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 5.5 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.

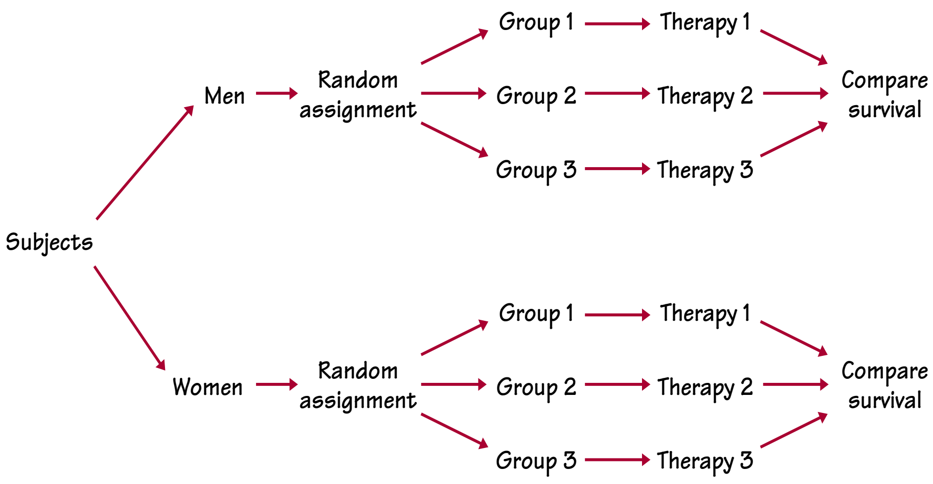


FIGURE 5.5 Outline of a block design. The blocks consist of male and female subjects. The treatments are three therapies for cancer.

EXAMPLE 5.18 SOYBEANS

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 5.19 STUDYING WELFARE SYSTEMS

A social policy experiment will assess the effect on family income of several proposed new welfare systems and compare them with the present welfare system. Because the income of a family under any welfare system is strongly related to its present income, the families who agree to participate are divided into blocks of similar income levels. The families in each block are then allocated at random among the welfare systems.

Blocks allow us to draw separate conclusions about each block, for example, about men and women in the cancer study in Example 5.17. 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.

EXERCISES

5.43 MEDITATION FOR ANXIETY An experiment that claimed to show that meditation lowers anxiety proceeded as follows. The experimenter interviewed the subjects and rated their level of anxiety. Then the subjects were randomly assigned to two groups. The experimenter taught one group how to meditate and they meditated daily for a month. The other group was simply told to relax more. At the end of the month, the experimenter interviewed all the subjects again and rated their anxiety level. The meditation group now had less anxiety. Psychologists said that the results were suspect because the ratings were not blind. Explain what this means and how lack of blindness could bias the reported results.

5.44 PAIN RELIEF STUDY Fizz Laboratories, a pharmaceutical company, has developed a new pain-relief medication. Sixty patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, “About what percentage of pain relief did you experience?”

(a) Why should Fizz not simply administer the new drug and record the patients’ responses?

(b) Outline the design of an experiment to compare the drug’s effectiveness with that of aspirin and of a placebo.

(c) Should patients be told which drug they are receiving? How would this knowledge probably affect their reactions?

(d) If patients are not told which treatment they are receiving, the experiment is single-blind. Should this experiment be double-blind also? Explain.

5.45 COMPARING WEIGHT-LOSS TREATMENTS Twenty overweight females have agreed to participate in a study of the effectiveness of four weight-loss treatments: A, B, C, and D. The researcher first calculates how overweight each subject is by comparing the subject's actual weight with her "ideal" weight. The subjects and their excess weights in pounds are

Birnbaum	35	Hernandez	25	Moses	25	Smith	29
Brown	34	Jackson	33	Nevesky	39	Stall	33
Brunk	30	Kendall	28	Obrach	30	Tran	35
Cruz	34	Loren	32	Rodriguez	30	Wilansky	42
Deng	24	Mann	28	Santiago	27	Williams	22

The response variable is the weight lost after 8 weeks of treatment. Because a subject's excess weight will influence the response, a block design is appropriate.

(a) Arrange the subjects in order of increasing excess weight. Form 5 blocks of 4 subjects each by grouping the 4 least overweight, then the next 4, and so on.

(b) Use Table B to randomly assign the 4 subjects in each block to the 4 weight-loss treatments. Be sure to explain exactly how you used the table.

5.46 CARBON DIOXIDE AND TREE GROWTH 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 three circular areas.[18]

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

(b) Areas 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 circles and carry out the randomization your design calls for.

5.47 DOES ROOM TEMPERATURE AFFECT MANUAL DEXTERITY? An expert on worker performance is interested in the effect of room temperature on the performance of tasks requiring manual dexterity. She chooses temperatures of 70° F and 90° F as treatments. The response variable is the number of correct insertions, during a 30-minute period, in a peg-and-hole apparatus that requires the use of both hands simultaneously. Each subject is trained on the apparatus and then asked to make as many insertions as possible in 30 minutes of continuous effort.

(a) Outline a completely randomized design to compare dexterity at 70° and 90°. Twenty subjects are available.

(b) Because individuals differ greatly in dexterity, the wide variation in individual scores may hide the systematic effect of temperature unless there are many subjects in each group. Describe in detail the design of a matched pairs experiment in which each subject serves as his or her own control.

5.48 CHARTING AS AN INVESTMENT STRATEGY Some investment advisors believe that charts of past trends in the prices of securities can help predict future prices. Most economists disagree. In an experiment to examine the effects of using charts, business students trade (hypothetically) a foreign currency at computer screens. There are 20 student subjects available, named for convenience A, B, C, . . . , T. Their goal is to make as much money as possible, and the best performances are rewarded with small prizes. The student traders have the price history of the foreign currency in dollars in their computers. They may or may not also have software that highlights trends. Describe *two* designs for this experiment, a completely randomized design and a matched pairs design in which each student serves as his or her own control. In both cases, carry out the randomization required by the design.

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 **replication**.

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.

Replication 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 5.2 EXERCISES

5.49 DOES SAINT-JOHN'S-WORT RELIEVE MAJOR DEPRESSION? Here are some excerpts from the report of a study of this issue.[19] 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.

5.50 MARKETING TO CHILDREN, I If children are given more choices within a class of products, will they tend to prefer that product to a competing product that offers fewer choices? Marketers want to know. An experiment prepared three "choice sets" of beverages. The first contained two milk drinks and two fruit drinks. The second had the same two fruit drinks but four milk drinks. The third contained four fruit drinks but only the original two milk drinks. The researchers divided 210 children aged 4 to 12 years into 3 groups at random. They offered each group one of the choice sets. As each child chose a beverage to drink from the choice set presented, the researchers noted whether the choice was a milk drink or a fruit drink.

(a) What are the experimental units or subjects?

(b) What is the factor, and what are its levels?

(c) What is the response variable?

5.51 BODY TEMPERATURE AND SURGERY Surgery patients are often cold because the operating room is kept cool and the body's temperature regulation is disturbed by anesthetics. Will warming patients to maintain normal body temperature reduce infections after surgery? In one experiment, patients undergoing colon surgery received intravenous fluids from a warming machine and were covered with a blanket through which air circulated. For some patients, the fluid and the air were warmed; for others, they were not. The patients received identical treatment in all other respects.[20]

(a) Identify the experimental subjects, the factor and its levels, and the response variables.

(b) Draw a diagram to outline the design of a randomized comparative experiment for this study.

(c) The following subjects have given consent to participate in this study. Do the random assignment required by your design. (If you use Table B, begin at line 121.)

Abbott
Adamson
Afifi
Brown
Cansico
Chen
Cordoba
Curzakis
Decker
Devlin
Engel
Fluharty
Garcia
Gerson
Green
Gupta
Gutierrez
Howard
Hwang
Iselin
Janle
Kaplan
Kim
Lattimore
Lucero
Masters
McNeill
Morse
Ng
Quinones
Rivera
Roberts
Rosen
Sugiwara
Thompson
Travers
Turing
Ullmann
Williams
Wong

(d) To simplify the setup of the study, we might warm the fluids and air blanket for one operating team and not for another doing the same kind of surgery. Why might this design result in bias?

(e) The operating team did not know whether fluids and air blanket were heated, nor did the doctors who followed the patients after surgery. What is this practice called? Why was it used here?

5.52 MARKETING TO CHILDREN, II Use a diagram to outline a completely randomized design for the children's choice study of Exercise 5.50.

5.53 DOES CALCIUM REDUCE BLOOD PRESSURE? You are participating in the design of a medical experiment to investigate whether a calcium supplement in the diet will reduce the blood pressure of middle-aged men. Preliminary work suggests that calcium may be effective and that the effect may be greater for black men than for white men. You have available 40 men with high blood pressure who are willing to serve as subjects.

(a) Outline an appropriate design for the experiment.

(b) The names of the subjects appear below. Use Table B, beginning at line 119, to do the randomization required by your design, and list the subjects to whom you will give the drug.

Alomar
Asihiro
Bennett
Bikalis
Chen
Clemente
Cranston
Curtis
Denman
Durr
Edwards
Farouk
Fratianna
George
Green
Guillen
Han
Howard
Hruska
Imrani
James
Kaplan
Krushchev
Lawless
Liang
Maldonado
Marsden
Moore
O'Brian
Ogle
Plochman
Rodriguez
Rosen
Solomon
Tompkins
Townsend
Tullock
Underwood
Willis
Zhang

(c) Choosing the sizes of the treatment groups requires more statistical expertise. We will learn more about this aspect of design in later chapters. Explain in plain language the advantage of using larger groups of subjects.

5.54 MARKETING TO CHILDREN, III The children's choice experiment in Exercise 5.50 has 210 subjects. Explain how you would assign labels to the 210 children in the actual experiment. Then use Table B at line 125 to choose *only the first* 5 children assigned to the first treatment.

5.55 PLACEBO EFFECT A survey of physicians found that some doctors give a placebo to a patient who complains of pain for which the physician can find no cause. If the patient's pain improves, these doctors conclude that it had no physical basis. The medical school researchers who conducted the survey claimed that these doctors do not understand the placebo effect. Why?

5.56 WILL TAKING ANTIOXIDANTS HELP PREVENT COLON CANCER? People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods. Fruits and vegetables are rich in "antioxidants" such as vitamins A, C, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied 864 people who were at risk of colon cancer. The subjects were divided into four groups: daily beta carotene, daily vitamins C and E, all three vitamins every day, and daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups.[21]

(a) What are the explanatory and response variables in this experiment?

(b) Outline the design of the experiment. Use your judgment in choosing the group sizes.

(c) Assign labels to the 864 subjects and use Table B, starting at line 118, to choose the first 5 subjects for the beta carotene group.

(d) The study was double-blind. What does this mean?

(e) What does "no significant difference" mean in describing the outcome of the study?

(f) Suggest some lurking variables that could explain why people who eat lots of fruits and vegetables have lower rates of colon cancer. The experiment suggests that these variables, rather than the antioxidants, may be responsible for the observed benefits of fruits and vegetables.

5.57 TREATING DRUNK DRIVERS Once a person has been convicted of drunk driving, one purpose of court-mandated treatment or punishment is to prevent future offenses of the same kind. Suggest three different treatments that a court might require. Then outline the design of an experiment to compare their effectiveness. Be sure to specify the response variables you will measure.

5.58 ACCULTURATION RATING There are several psychological tests that measure the extent to which Mexican Americans are oriented toward Mexican/Spanish or Anglo/English culture. Two such tests are the Bicultural Inventory (BI) and the Acculturation Rating Scale for Mexican Americans (ARSMA). To study the correlation between the scores on these two tests, researchers will give both tests to a group of 22 Mexican Americans.

(a) Briefly describe a matched pairs design for this study. In particular, how will you use randomization in your design?

(b) You have an alphabetized list of the subjects (numbered 1 to 22). Carry out the randomization required by your design and report the result.

5.3 SIMULATING EXPERIMENTS

Toss a coin 10 times. What is the likelihood of a run of 3 or more consecutive heads or tails? A couple plans to have children until they have a girl or until they have four children, whichever comes first. What are the chances that they will have a girl among their children? An airline knows from past experience that a certain percentage of customers who have purchased tickets will not show up to board the airplane. If the airline "overbooks" a particular flight (i.e., sells more tickets than they have seats), what are the chances that the airline will encounter more ticketed passengers than they have seats for? There are three methods we can use to answer questions involving chance like these:

1. Try to estimate the likelihood of a result of interest by actually carrying out the experiment many times and calculating the result's relative frequency. That's slow, sometimes costly, and often impractical or logistically difficult.
2. Develop a ***probability model*** and use it to calculate a theoretical answer. This requires that we know something about the rules of probability and therefore may not be feasible. (We will develop a probability model in the next chapter.)
3. Start with a model that, in some fashion, reflects the truth about the experiment, and then develop a procedure for imitating—or simulating—a number of repetitions of the experiment. This is quicker than repeating the real experiment, especially if we can use the TI-83/89 or a computer, and it allows us to do problems that are hard when done with formal mathematical analysis.

probability model

Here is an example of a simulation.

EXAMPLE 5.20 A GIRL IN THE FAMILY

Suppose we are interested in estimating the likelihood of a couple's having a girl among their first four children. Let a flip of a fair coin represent a birth, with heads corresponding to a girl and tails a boy. Since girls and boys are equally likely to occur on any birth, the coin flip is an accurate imitation of the situation. Flip the coin until a head appears or until the coin has been flipped 4 times, whichever comes first. The appearance of a head within the first 4 flips corresponds to the couple's having a girl among their first four children.

If this coin-flipping procedure is repeated many times, to represent the births in a large number of families, then the proportion of times that a head appears within the first 4 flips should be a good estimate of the true likelihood of the couple's having a girl.

A single die (one of a pair of dice) could also be used to simulate the birth of a son or daughter. Let an even number of spots (called pips) represent a girl, and let an odd number of spots represent a boy.

SIMULATION

The imitation of chance behavior, based on a model that accurately reflects the experiment under consideration, is called a **simulation**.

Simulation is an effective tool for finding likelihoods of complex results once we have a trustworthy model. In particular, we can use random digits from a table, graphing calculator, or computer software to simulate many repetitions quickly. The proportion of repetitions on which a result occurs will eventually be close to its true likelihood, so simulation can give good estimates of probabilities. The art of random digit simulation can be illustrated by a series of examples.

EXAMPLE 5.21 SIMULATION STEPS

Step 1: **State the problem or describe the experiment.** Toss a coin 10 times. What is the likelihood of a run of at least 3 consecutive heads or 3 consecutive tails?

Step 2: **State the assumptions.** There are two:

- A head or a tail is equally likely to occur on each toss.
- Tosses are independent of each other (i.e., what happens on one toss will not influence the next toss).

Step 3: **Assign digits to represent outcomes.** In a random number table, such as Table B in the back of the book, the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 occur with the same long-term relative frequency (1/10). We also know that the successive digits in the table are independent. It follows that even digits and odd digits occur with the same long-term relative frequency, 50%. Here is one assignment of digits for coin tossing:

- One digit simulates one toss of the coin.
- Odd digits represent heads; even digits represent tails.

Successive digits in the table simulate independent tosses.

Step 4: **Simulate many repetitions.** Looking at 10 consecutive digits in Table B simulates one repetition. Read many groups of 10 digits from the table to simulate many repetitions. Be sure to keep track of whether or not the event we want (a run of 3 heads or 3 tails) occurs on each repetition.

Here are the first three repetitions, starting at line 101 in Table B. Runs of 3 or more heads or tails have been underlined.

Digits	1 9 2 2 3 9 5 0 3 4	0 5 7 5 6 2 8 7 1 3	9 6 4 0 9 1 2 5 3 1
Heads/tails	H H T T $\underline{\text{H H H}}$ T H T	T $\underline{\text{H H H}}$ $\underline{\text{T T T}}$ $\underline{\text{H H H}}$	H $\underline{\text{T T T}}$ H H T $\underline{\text{H H H}}$
Run of 3	YES	YES	YES

Twenty-two additional repetitions were done for a total of 25 repetitions; 23 of them did have a run of 3 or more heads or tails.

Step 5: **State your conclusions.** We estimate the probability of a run by the proportion

$$\text{estimated probability} = \frac{23}{25} = 0.92$$

Of course, 25 repetitions are not enough to be confident that our estimate is accurate. Now that we understand how to do the simulation, we can tell a computer to do many thousands of repetitions. A long simulation (or mathematical analysis) finds that the true probability is about 0.826.

Once you have gained some experience in simulation, establishing a correspondence between random numbers and outcomes in the experiment is usually the hardest part, and must be done carefully. Although coin tossing may not fascinate you, the model in Example 5.21 is typical of many probability problems because it consists of independent trials (the tosses) all having the same possible outcomes and probabilities. The coin tosses are said to be ***independent*** because the result of one toss has no effect or influence over the next coin toss. Shooting 10 free throws and observing the sexes of 10 children have similar models and are simulated in much the same way.

independent

The idea is to state the basic structure of the random phenomenon and then use simulation to move from this model to the probabilities of more complicated events. The model is based on opinion and past experience. If it does not correctly describe the random phenomenon, the probabilities derived from it by simulation will also be incorrect.

Step 3 (assigning digits) can usually be done in several different ways, but some assignments are more efficient than others. Here are some examples of this step.

EXAMPLE 5.22 ASSIGNING DIGITS

(a) Choose a person at random from a group of which 70% are employed. One digit simulates one person:

0, 1, 2, 3, 4, 5, 6 = employed
7, 8, 9 = not employed

The following correspondence is also satisfactory:

00, 01, . . . , 69 = employed
70, 71, . . . , 99 = not employed

This assignment is less efficient, however, because it requires twice as many digits and ten times as many numbers.

(b) Choose one person at random from a group of which 73% are employed. Now *two* digits simulate one person:

00, 01, 02, . . . , 72 = employed
73, 74, 75, . . . , 99 = not employed

We assigned 73 of the 100 two-digit pairs to “employed” to get probability 0.73. Representing “employed” by 01, 02, . . . , 73 would also be correct.

(c) Choose one person at random from a group of which 50% are employed, 20% are unemployed, and 30% are not in the labor force. There are now three possible outcomes, but the principle is the same. One digit simulates one person:

0, 1, 2, 3, 4 = employed
5, 6 = unemployed
7, 8, 9 = not in the labor force

Another valid assignment of digits might be

0, 1 = unemployed
2, 3, 4 = not in the labor force
5, 6, 7, 8, 9 = employed

What is important is the number of digits assigned to each outcome, not the order of the digits.

As the last example shows, simulation methods work just as easily when outcomes are not equally likely. Consider the following slightly more complicated example.

EXAMPLE 5.23 FROZEN YOGURT SALES

Orders of frozen yogurt flavors (based on sales) have the following relative frequencies: 38% chocolate, 42% vanilla, and 20% strawberry. The experiment consists of customers entering the store and ordering yogurt. The task is to simulate 10 frozen yogurt sales based on this recent history. Instead of considering the random number table to be made up of single digits, we now consider it to be made up of pairs of digits. This is because the relative frequencies of interest have a maximum of *two* significant digits. The range of the pairs of digits is 00 to 99, and since all the pairs are equally likely to occur, the pairs 00, 01, 02, . . . , 99 all have relative frequency 0.01.

Thus we may assign the numbers in the random number table as follows:

- 00 to 37 to correspond to the outcome chocolate (C)
- 38 to 79 to correspond to the outcome vanilla (V)
- 80 to 99 to correspond to the outcome strawberry (S)

The sequence of random numbers (starting at the 21st column of row 112 in Table B) is as follows:

19352 73089 84898 45785

This yields the following two-digit numbers:

19 35 27 30 89 84 89 84 57 85

which correspond to the outcomes

C C C C S S S S V S

EXAMPLE 5.24 A GIRL OR FOUR

A couple plans to have children until they have a girl or until they have four children, whichever comes first. We will show how to use random digits to estimate the likelihood that they will have a girl.

The model is the same as for coin tossing. We will assume that each child has probability 0.5 of being a girl and 0.5 of being a boy, and the sexes of successive children are independent.

Assigning digits is also easy. One digit simulates the sex of one child:

0, 1, 2, 3, 4 = girl
5, 6, 7, 8, 9 = boy

To simulate one repetition of this child-bearing strategy, read digits from Table B until the couple has either a girl or four children. Notice that the number of digits needed to simulate one repetition depends on how quickly the couple gets a girl. Here is the simulation, using line 130 of Table B. To interpret the digits, G for girl and B for boy are written under them, space separates repetitions, and under each repetition "+" indicates if a girl was born and "–" indicates one was not.

690	51	64	81	7871	74	0
BBG	BG	BG	BG	BBBG	BG	G
+	+	+	+	+	+	+
951	784	53	4	0	64	8987
BBG	BBG	BG	G	G	BG	BBBB
+	+	+	+	+	+	–

In these 14 repetitions, a girl was born 13 times. Our estimate of the probability that this strategy will produce a girl is therefore

$$\text{estimated probability} = \frac{13}{14} = 0.93$$

Some mathematics shows that if our probability model is correct, the true likelihood of having a girl is 0.938. Our simulated answer came quite close. Unless the couple is unlucky, they will succeed in having a girl.

EXERCISES

5.59 ESTABLISHING A CORRESPONDENCE State how you would use the following aids to establish a correspondence in a simulation that involves a 75% chance:

(a) a coin

(b) a six-sided die

(c) a random digit table (Table B)

(d) a standard deck of playing cards

5.60 THE CLEVER COINS Suppose you left your statistics textbook and calculator in your locker, and you need to simulate a random phenomenon that has a 25% chance of a desired outcome. You discover two nickels in your pocket that are left over from your lunch money. Describe how you could use the two coins to set up your simulation.

5.61 ABOLISH EVENING EXAMS? Suppose that 84% of a university's students favor abolishing evening exams. You ask 10 students chosen at random. What is the likelihood that all 10 favor abolishing evening exams?

(a) Describe how you would pose this question to 10 students independently of each other. How would you model the procedure?

(b) Assign digits to represent the answers "Yes" and "No."

(c) Simulate 5 repetitions, starting at line 129 of Table B. Then combine your results with those of the rest of your class. What is your estimate of the likelihood of the desired result?

5.62 SHOOTING FREE THROWS A basketball player makes 70% of her free throws in a long season. In a tournament game she shoots 5 free throws late in the game and misses 3 of them. The fans think she was nervous, but the misses may simply be chance. You will shed some light by estimating a probability.

(a) Describe how to simulate a single shot if the probability of making each shot is 0.7. Then describe how to simulate 5 independent shots.

(b) Simulate 50 repetitions of the 5 shots and record the number missed on each repetition. Use Table B starting at line 125. What is the approximate likelihood that the player will miss 3 or more of the 5 shots?

5.63 A POLITICAL POLL, I An opinion poll selects adult Americans at random and asks them, "Which political party, Democratic or Republican, do you think is better able to manage the economy?" Explain carefully how you would assign digits from Table B to simulate the response of one person in each of the following situations.

(a) Of all adult Americans, 50% would choose the Democrats and 50% the Republicans.

(b) Of all adult Americans, 60% would choose the Democrats and 40% the Republicans.

(c) Of all adult Americans, 40% would choose the Democrats, 40% would choose the Republicans, and 20% would be undecided.

(d) Of all adult Americans, 53% would choose the Democrats and 47% the Republicans.

5.64 A POLITICAL POLL, II Use Table B to simulate the responses of 10 independently chosen adults in each of the four situations of Exercise 5.63.

(a) For situation (a), use line 110.

(b) For situation (b), use line 111.

(c) For situation (c), use line 112.

(d) For situation (d), use line 113.

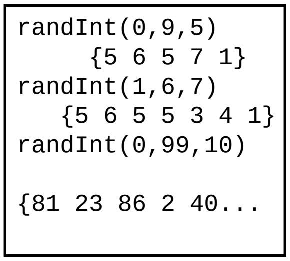
Simulations with the calculator or computer

The calculator and computer can be extremely useful in conducting simulations because they can be easily programmed to quickly perform a large number of repetitions. Study the reasoning and the steps involved in the following example so that you may become adept at using the capabilities of the TI-83/89 to design and carry out simulations.

EXAMPLE 5.25 RANDOMIZING WITH THE CALCULATOR

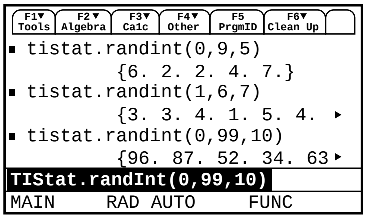
The command `randInt` (found under MATH/PRB/5:randInt on the TI-83, and under CATALOG F3 (Flash Apps) on the TI-89) can be used to generate random digits between any two specified values. Here are three applications.

The command `randInt(0,9,5)` generates 5 random integers between 0 and 9. This could serve as a block of 5 random digits in the random number table. The command `randInt(1,6,7)` could be used to simulate rolling a die 7 times. Generating 10 two-digit numbers between 00 and 99 from Example 5.23 could be done with the command `randInt(0,99,10)`.



```
randInt(0,9,5)
     {5 6 5 7 1}
randInt(1,6,7)
   {5 6 5 5 3 4 1}
randInt(0,99,10)

{81 23 86 2 40...
```



```
■ tistat.randint(0,9,5)
            {6. 2. 2. 4. 7.}
■ tistat.randint(1,6,7)
            {3. 3. 4. 1. 5. 4. ▸
■ tistat.randint(0,99,10)
            {96. 87. 52. 34. 63▸
TIStat.randInt(0,99,10)
MAIN      RAD AUTO      FUNC
```

Using the statistical software package Minitab, the following set of commands will generate a set of 10 random numbers in the range 00 to 99 and store these numbers in column C1.

```
MTB > random 10 c1;
SUBC> integer 0 99.
MTB > Print C1

C1
   38   93   14   30   50   92   16   18   84   20
```

When you combine the power and simplicity of simulations with the power of technology, you have formidable tools for answering questions involving chance behavior.

EXERCISES

5.65 A GIRL OR FOUR Use your calculator to simulate a couple's having children until they have a girl or until they have four children, whichever comes first. (See Example 5.24.) Use the simulation to estimate the probability that they will have a girl among their children. Compare your calculator results with those of Example 5.24.

5.66 WORLD SERIES Suppose that in a particular year the American League baseball team is considered to have a 60% chance of beating the National League team in any given World Series game. (This assumption ignores any possible home-field advantage, which is probably not very realistic.) To win the World Series, a team must win 4 out of 7 games in the series. Further assume that the outcome of each game is not influenced by the outcome of any other game (that is, who wins one game is independent of who wins any other game).

(a) Use simulation methods to approximate the number of games that would have to be played in order to determine the world champion.

(b) The so-called home-field advantage is one factor that might be an explanatory variable in determining the winner of a game. What are some other possible factors?

5.67 TENNIS RACQUETS Professional tennis players bring multiple racquets to each match. They know that high string tension, the force with which they hit the ball, and occasional "racquet abuse" are all reasons why racquets break during a match. Brian Lob's coach tells him that he has a 15% chance of breaking a racquet in any given match. How many matches, on average, can Brian expect to play until he breaks a racquet and needs to use a backup? Use simulation methods to answer this question.

SUMMARY

There are times when actually carrying out an experiment is too costly, too slow, or simply impractical. In situations like these, a carefully designed **simulation** can provide approximate answers to our questions.

A simulation is an imitation of chance behavior, most often carried out with random numbers. The **steps of a simulation** are:

1. State the problem or describe the experiment.
2. State the assumptions.
3. Assign digits to represent outcomes.
4. Simulate many repetitions.
5. State your conclusions.

Programmable calculators, like the TI-83/89, and computers are particularly useful for conducting simulations because they can perform many repetitions quickly.

SECTION 5.3 EXERCISES

5.68 GAME OF CHANCE, I Amarillo Slim is a cardsharp who likes to play the following game. Draw 2 cards from the deck of 52 cards. If at least one of the cards is a heart, then you win $1. If neither card is a heart, then you lose $1.

(a) Describe a correspondence between random numbers and possible outcomes in this game.

(b) Simulate playing the game for 25 rounds. Shuffle the cards after each round. See if you can beat Amarillo Slim at his own game. Remember to write down the results of each game. When you finish, combine your results with those of 3 other students to obtain a total of 100 trials. Report your cumulative proportion of wins. Do you think this is a "fair" game? That is, do both you and Slim have an equal chance of winning?

5.69 GAME OF CHANCE, II A certain game of chance is based on randomly selecting three numbers from 00 to 99, inclusive (allowing repetitions), and adding the numbers. A person wins the game if the resulting sum is a multiple of 5.

(a) Describe your scheme for assigning random numbers to outcomes in this game.

(b) Use simulation to estimate the proportion of times a person wins the game.

5.70 THE BIRTHDAY PROBLEM Use your calculator and the simulation method to show that in a class of 23 unrelated students, the chances of at least 2 students with the same birthday are about 50%. Show that in a room of 41 people, the chances of at least 2 people having the same birthday are about 90%. What assumptions are you using in your simulations?

5.71 BATTER UP! Suppose a major league baseball player has a current batting average of .320. Note that the batting average = (number of hits)/(number of at-bats).

(a) Describe an assignment of random numbers to possible results in order to simulate the player's next 20 at-bats.

(b) Carry out the simulation for 20 repetitions, and report your results. What is the relative frequency of at-bats in which the player gets a hit?

(c) Compare your simulated experimental results with the player's actual batting average of .320.

5.72 NUCLEAR SAFTEY A nuclear reactor is equipped with two independent automatic shutdown systems to shut down the reactor when the core temperature reaches the danger level. Neither system is perfect. System A shuts down the reactor 90% of the time when the danger level is reached. System B does so 80% of the time. The reactor is shut down if *either* system works.

(a) Explain how to simulate the response of System A to a dangerous temperature level.

(b) Explain how to simulate the response of System B to a dangerous temperature level.

(c) Both systems are in operation simultaneously. Combine your answers to (a) and (b) to simulate the response of both systems to a dangerous temperature level. Explain why you cannot use the same entry in Table B to simulate both responses.

(d) Now simulate 100 trials of the reactor's response to an emergency of this kind. Estimate the probability that it will shut down. This probability is higher than the probability that either system working alone will shut down the reactor.

5.73 SPREADING A RUMOR On a small island there are 25 inhabitants. One of these inhabitants, named Jack, starts a rumor which spreads around the isle. Any person who hears the rumor continues spreading it until he or she meets someone who has heard the story before. At that point, the person stops spreading it, since nobody likes to spread stale news.

(a) Do you think that all 25 inhabitants will eventually hear the rumor or will the rumor die out before that happens? Estimate the proportion of inhabitants who will hear the rumor.

(b) In the first time increment, Jack randomly selects one of the other inhabitants, named Jill, to tell the rumor to. In the second time increment, both Jack and Jill each randomly select one of the remaining 24 inhabitants to tell the rumor to. (*Note:* They could conceivably pick each other again.) In the next time increment, there are 4 rumor spreaders, and so on. If a randomly selected person has already heard the rumor, that rumor teller stops spreading the rumor. Design a record-keeping chart, and simulate this procedure. Use your TI-83/89 to help with the random selection. Continue until all 25 inhabitants hear the rumor or the rumor dies out. How many inhabitants out of 25 eventually heard the rumor?

(c) Combine your results with those of other students in the class. What is the mean number of inhabitants who hear the rumor?

CHAPTER REVIEW

Designs for producing data are essential parts of statistics in practice. Random sampling and randomized comparative experiments are perhaps the most important statistical inventions in this century. Both were slow to gain acceptance, and you will still see many voluntary response samples and uncontrolled experiments. This chapter has explained good techniques for producing data and has also explained why bad techniques often produce worthless data.The deliberate use of chance in producing data is a central idea in statistics. It allows use of the laws of probability to analyze data, as we will see in the following chapters. Here are the major skills you should have now that you have studied this chapter.

A. SAMPLING

1. Identify the population in a sampling situation.

2. Recognize bias due to voluntary response samples and other inferior sampling methods.

3. Use Table B of random digits to select a simple random sample (SRS) from a population.

4. Recognize the presence of undercoverage and nonresponse as sources of error in a sample survey. Recognize the effect of the wording of questions on the response.

5. Use random digits to select a stratified random sample from a population when the strata are identified.

B. EXPERIMENTS

1. Recognize whether a study is an observational study or an experiment.

2. Recognize bias due to confounding of explanatory variables with lurking variables in either an observational study or an experiment.

3. Identify the factors (explanatory variables), treatments, response variables, and experimental units or subjects in an experiment.

4. Outline the design of a completely randomized experiment using a diagram like those in Examples 5.12 and 5.13. The diagram in a specific case should show the sizes of the groups, the specific treatments, and the response variable.

5. Use Table B of random digits to carry out the random assignment of subjects to groups in a completely randomized experiment.

6. Recognize the placebo effect. Recognize when the double-blind technique should be used.

7. Recognize a block design when it would be appropriate. Know when a matched pairs design would be appropriate and how to design a matched pairs experiment.

8. Explain why a randomized comparative experiment can give good evidence for cause-and-effect relationships.

C. SIMULATIONS

1. Recognize that many random phenomena can be investigated by means of a carefully designed simulation.

2. Use the following steps to construct and run a simulation:

- **a.** State the problem or describe the experiment.
- **b.** State the assumptions.
- **c.** Assign digits to represent outcomes.
- **d.** Simulate many repetitions.
- **e.** Calculate relative frequencies and state your conclusions.

3. Use a random number table, the TI-83/89, or a computer utility such as Minitab, Data Desk, or a spreadsheet to conduct simulations.

CHAPTER 5 REVIEW EXERCISES

5.74 ONTARIO HEALTH SURVEY The Ministry of Health in the Province of Ontario, Canada, wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records, but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a random sample of 61,239 people who live in the Province of Ontario.[22]

(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?

5.75 TREATING BREAST CANCER What is the preferred treatment for breast cancer that is detected in its early stages? The most common treatment was once removal of the breast. It is now usual to remove only the tumor and nearby lymph nodes, followed by radiation. To study whether these treatments differ in their effectiveness, a medical team examines the records of 25 large hospitals and compares the survival times after surgery of all women who have had either treatment.

(a) What are the explanatory and response variables?

(b) Explain carefully why this study is not an experiment.

(c) Explain why confounding will prevent this study from discovering which treatment is more effective. (The current treatment was in fact recommended after a large randomized comparative experiment.)

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

(a) Are people generally satisfied with how things are going in the country right now?

(b) Do college students learn basic accounting better in the classroom or using an online course?

(c) How long do your teachers wait on the average after they ask their class a question?

5.77 COACH, I NEED OXYGEN! We often see players on the sidelines of a football game inhaling oxygen. Their coaches think it 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 25 football players as subjects. Discuss the design of such an experiment to investigate the effect of inhaling oxygen during the rest period.

5.78 POLLING THE FACULTY A labor organization wants to study the attitudes of college faculty members toward collective bargaining. These attitudes appear to be different depending on the type of college. The American Association of University Professors classifies colleges as follows:

Class I. Offer doctorate degrees and award at least 15 per year.

Class IIA. Award degrees above the bachelor's but are not in Class I.

Class IIB. Award no degrees beyond the bachelor's.

Class III. Two-year colleges.

Discuss the design of a sample of faculty from colleges in your state, with total sample size about 200.

5.79 FOOD FOR CHICKS New varieties of corn with altered amino acid content may have higher nutritional value than standard corn, which is low in the amino acid lysine. An experiment compares two new varieties, called opaque-2 and floury-2, with normal corn. The researchers mix corn-soybean meal diets using each type of corn at each of three protein levels, 12% protein, 16% protein, and 20% protein. They feed each diet to 10 one-day-old male chicks and record their weight gains after 21 days. The weight gain of the chicks is a measure of the nutritional value of their diet.

(a) What are the experimental units and the response variable in this experiment?

(b) How many factors are there? How many treatments? Use a diagram like Figure 5.2 to describe the treatments. How many experimental units does the experiment require?

(c) Use a diagram to describe a completely randomized design for this experiment. (You do not need to actually do the randomization.)

5.80 VITAMIN C FOR MARATHON 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 those 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.[23]

(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.

5.81 DELIVERING THE MAIL Is the number of days a letter takes to reach another city affected by the time of day it is mailed and whether or not the zip code is used? Describe briefly the design of a two-factor experiment to investigate this question. Be sure to specify the treatments exactly and to tell how you will handle lurking variables such as the day of the week on which the letter is mailed.

5.82 McDONALD'S VERSUS WENDY'S Do consumers prefer the taste of a cheeseburger from McDonald's or from Wendy's in a blind test in which neither burger is identified? Describe briefly the design of a matched pairs experiment to investigate this question.

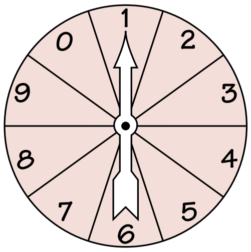
5.83 REPAIRING KNEES IN COMFORT Knee injurys are routinely repaired by arthroscopic surgery that does not require opening up the knee. Can we reduce patient discomfort by giving them a nonsteroidal anti-inflammatory drug (NSAID)? Eighty-three patients were placed in three groups. Group A received the NSAID both before and after the surgery. Group B was given a placebo before and the NSAID after. Group C received a placebo both before and after surgery. The patients recorded a pain score by answering questions one day after the surgery.[24]

(a) Outline the design of this experiment. You do not need to do the randomization that your design requires.

(b) You read that "the patients, physicians and physical therapists were blinded" during the study. What does this mean?

(c) You also read that "the pain scores for Group A were significanly lower than Group C but not significantly lower than Group B." What does this mean? What does this finding lead you to conclude about the use of NSAIDs?

5.84 A SPINNER GAME OF CHANCE A game of chance is based on spinning a 1–10 spinner like the one shown in the illustration two times in succession. The player wins if the larger of the two numbers is greater than 5.



(a) What constitutes a single run of this experiment? What are the possible outcomes resulting in win or lose?

(b) Describe a correspondence between random digits from a random number table and outcomes in the game.

(c) Describe a technique using the `randInt` command on the TI-83/89 to simulate the result of a single run of the experiment.

(d) Use either the random number table or your calculator to simulate 20 trials. Report the proportion of times you win the game. Then combine your results with those of other students to obtain results for a large number of trials.

5.85 GAUGING THE DEMAND FOR CHEESECAKE The owner of a bakery knows that the daily demand for a highly perishable cheesecake is as follows:

Number/day:	0	1	2	3	4	5
Relative frequency:	0.05	0.15	0.25	0.25	0.20	0.10

(a) Use simulation to find the demand for the cheesecake on 30 consecutive business days.

(b) Suppose that it cost the baker $5 to produce a cheesecake, and that the unused cheesecakes must be discarded at the end of the business day. Suppose also that the selling price of a cheesecake is $13. Use simulation to estimate the number of cheesecakes that he should produce each day in order to maximize his profit.

5.86 HOT STREAKS IN FOUL SHOOTING Joey is interested in investigating so-called hot streaks in foul shooting among basketball players. He's a fan of Carla, who has been making approximately 80% of her free throws. Specifically, Joey wants to use simulation methods to determine Carla's longest *run* of baskets on average, for 20 consecutive free throws.

(a) Describe a correspondence between random numbers and outcomes.

(b) What will constitute one repetition in this simulation? Carry out 20 repetitions and record the longest run for each repetition. Combine your results with those of 4 other students to obtain at least 100 replications.

(c) What is the mean run length? Are you surprised? Determine the five-number summary for the data.

(d) Construct a histogram of the results.

5.87 **SELF-PACED LEARNING, I** Elaine is enrolled in a self-paced course that allows three attempts to pass an examination on the material. She does not study and has 2 out of 10 chances of passing on any one attempt by luck. What is Elaine's likelihood of passing on at least one of the three attempts? (Assume the attempts are independent because she takes a different examination on each attempt.)

(a) Explain how you would use random digits to simulate one attempt at the exam. Elaine will of course stop taking the exam as soon as she passes.

(b) Simulate 50 repetitions. What is your estimate of Elaine's likelihood of passing the course?

(c) Do you think the assumption that Elaine's likelihood of passing the exam is the same on each trial is realistic? Why?

5.88 **SELF-PACED LEARNING, II** A more realistic model for Elaine's attempts to pass an exam in the previous exercise is as follows: On the first try she has probability 0.2 of passing. If she fails on the first try, her probability on the second try increases to 0.3 because she learned something from her first attempt. If she fails on two attempts, the probability of passing on a third attempt is 0.4. She will stop as soon as she passes. The course rules force her to stop after three attempts in any case.

(a) Explain how to simulate one repetition of Elaine's tries at the exam. Notice that she has different probabilities of passing on each successive try.

(b) Simulate 50 repetitions and estimate the probability that Elaine eventually passes the exam.

NOTES AND DATA SOURCES

1. Reported by D. Horvitz in his contribution to "Pseudo–opinion polls: SLOP or useful data?" *Chance*, 8, No. 2 (1995), pp. 16–25.
2. Based in part on Randall Rothenberger, "The trouble with mall interviewing," *New York Times*, August 16, 1989.
3. K. J. Mukamal et al., "Prior alcohol consumption and mortality following acute myocardial infarction," *Journal of the American Medical Association*, 285 (2001), pp. 1965–1970.
4. L. E. Moses and F. Mosteller, "Safety of anesthetics," in J. M. Tanur et al. (eds.), *Statistics: A Guide to the Unknown*, 3rd ed., Wadsworth, 1989, pp. 15–24.
5. The information in this example is taken from *The ASCAP Survey and Your Royalties*, ASCAP, New York, undated.
6. The most recent account of the design of the CPS is Bureau of Labor Statistics, *Design and Methodology*, Current Population Survey Technical Paper 63, March 2000 (available in print or online at www.bls.census.gov/cps/tp/tp63.htm). The account here omits many complications, such as the need to separately sample "group quarters" like college dormitories.

7. For more detail on the material of this section and complete references, see P. E. Converse and M. W. Traugott, “Assessing the accuracy of polls and surveys,” *Science*, 234 (1986), pp. 1094–1098.

8. The estimates of the census undercount come from Howard Hogan, “The 1990 post-enumeration survey: operations and results,” *Journal of the American Statistical Association*, 88 (1993), pp. 1047–1060. The information about nonresponse appears in Eugene P. Eriksen and Teresa K. DeFonso, “Beyond the net undercount: how to measure census error, *Chance*, 6, No. 4 (1993), pp. 38–43 and 14.

9. For more detail on the limits of memory in surveys, see N. M. Bradburn, L. J. Rips, and S. K. Shevell, “Answering autobiographical questions: the impact of memory and inference on surveys,” *Science*, 236 (1987), pp. 157–161.

10. Cynthia Crossen, “Margin of error: studies galore support products and positions, but are they reliable?” *Wall Street Journal*, November 14, 1991.

11. M. R. Kagay, “Poll on doubt of Holocaust is corrected,” *New York Times*, July 8, 1994.

12. Giuliana Coccia, “An overview of non-response in Italian telephone surveys,” *Proceedings of the 99th Session of the International Statistics Institute, 1993*, Book 3, pp. 271–272.

13. From the *New York Times* of August 21, 1989.

14. Steering Committee of the Physicians’ Health Study Research Group, “Final report on the aspirin component of the ongoing Physicians’ Health Study,” *New England Journal of Medicine*, 321 (1989), pp. 129–135.

15. L. L. Miao, “Gastric freezing: an example of the evaluation of medical therapy by randomized clinical trials,” in J. P. Bunker, B. A. Barnes, and F. Mosteller (eds.), *Costs, Risks, and Benefits of Surgery*, Oxford University Press, New York, 1977, pp. 198–211.

16. Based on Christopher Anderson, “Measuring what works in health care,” *Science*, 263 (1994), pp. 1080–1082.

17. K. Wang, Y. Li, and J. Erickson, “A new look at the Monday effect,” *Journal of Finance*, 52 (1997), pp. 2171–2186.

18. Based on Evan H. DeLucia et al., “Net primary production of a forest ecosystem with experimental CO_2 enhancement,” *Science*, 284 (1999), pp. 1177–1179. The investigators used the block design.

19. R. C. Shelton et al., “Effectiveness of St.-John’s-wort in major depression,” *Journal of the American Medical Association*, 285 (2001), pp. 1978–1986.

20. Based on the Electronic Encyclopedia of Statistical Examples and Exercises (EESEE) story “Surgery in a Blanket,” found on the TPS Web site www.whfreeman.com/tps.

21. The study is described in G. Kolata, “New study finds vitamins are not cancer preventers,” *New York Times*, July 21, 1994. Look in the *Journal of the American Medical Association* of the same date for the details.

22. Information from Warren McIsaac and Vivek Goel, “Is access to physician services in Ontario equitable?” Institute for Clinical Evaluative Sciences in Ontario, October 18, 1993.

23. E. M. Peters et al., “Vitamin C supplementation reduces the incidence of post-race symptoms of upper-respiratory tract infection in ultramarathon runners,” *American Journal of Clinical Nutrition*, 57 (1993), pp. 170–174.

24. This exercise is based on the EESEE story “Blinded Knee Doctors.” The study was reported in W. E. Nelson, R. C. Henderson, L. C. Almekinders, R. A. DeMasi, and T. N. Taft, “An evaluation of pre- and postoperative nonsteroidal antiinflammatory drugs in patients undergoing knee arthroscopy,” *Journal of Sports Medicine*, 21 (1994), pp. 510–516.