Chapter

1

INTRODUCTION

Objectives

In this chapter we will look at a series of examples of areas in the life sciences in which statistics is used, with the goal of understanding the scope of the field of statistics. We will also

- explain how experiments differ from observational studies.
- discuss the concepts of placebo effect, blinding, and confounding.
- discuss the role of random sampling in statistics.

1.1 Statistics and the Life Sciences

Researchers in the life sciences carry out investigations in various settings: in the clinic, in the laboratory, in the greenhouse, in the field. Generally, the resulting data exhibit some *variability*. For instance, patients given the same drug respond somewhat differently; cell cultures prepared identically develop somewhat differently; adjacent plots of genetically identical wheat plants yield somewhat different amounts of grain. Often the degree of variability is substantial even when experimental conditions are held as constant as possible.

The challenge to the life scientist is to discern the patterns that may be more or less obscured by the variability of responses in living systems. The scientist must try to distinguish the "signal" from the "noise."

Statistics is the science of understanding data and of making decisions in the face of variability and uncertainty. The discipline of statistics has evolved in response to the needs of scientists and others whose data exhibit variability. The concepts and methods of statistics enable the investigator to describe variability and to plan research so as to take variability into account (i.e., to make the "signal" strong in comparison to the background "noise" in data that are collected). Statistical methods are used to analyze data so as to extract the maximum information and also to quantify the reliability of that information.

We begin with some examples that illustrate the degree of variability found in biological data and the ways in which variability poses a challenge to the biological researcher. We will briefly consider examples that illustrate some of the statistical issues that arise in life sciences research and indicate where in this book the issues are addressed.

The first two examples provide a contrast between an experiment that showed no variability and another that showed considerable variability.

Example 1.1.1

Vaccine for Anthrax Anthrax is a serious disease of sheep and cattle. In 1881, Louis Pasteur conducted a famous experiment to demonstrate the effect of his vaccine against anthrax. A group of 24 sheep were vaccinated; another group of 24 unvaccinated sheep served as controls. Then, all 48 animals were inoculated with a virulent culture of anthrax bacillus. Table 1.1.1 shows the results.[1] The data of Table 1.1.1 show no variability; all the vaccinated animals survived and all the unvaccinated animals died. ■

Table 1.1.1 Response of sheep to anthrax

	Treatment	
Response	Vaccinated	Not vaccinated
Died of anthrax	0	24
Survived	24	0
Total	24	24
Percent survival	100%	0%

Example 1.1.2

Bacteria and Cancer To study the effect of bacteria on tumor development, researchers used a strain of mice with a naturally high incidence of liver tumors. One group of mice were maintained entirely germ free, while another group were exposed to the intestinal bacteria *Escherichia coli*. The incidence of liver tumors is shown in Table 1.1.2.[2]

Table 1.1.2 Incidence of liver tumors in mice

	Treatment	
Response	*E. coli*	Germ free
Liver tumors	8	19
No liver tumors	5	30
Total	13	49
Percent with liver tumors	62%	39%

In contrast to Table 1.1.1, the data of Table 1.1.2 show variability; mice given the same treatment did not all respond the same way. Because of this variability, the results in Table 1.1.2 are equivocal; the data suggest that exposure to *E. coli* increases the risk of liver tumors, but the possibility remains that the observed difference in percentages (62% versus 39%) might reflect only chance variation rather than an effect of *E. coli*. If the experiment were replicated with different animals, the percentages might change substantially.

One way to explore what might happen if the experiment were replicated is to simulate the experiment, which could be done as follows. Take 62 cards and write "liver tumors" on 27 (= 8 + 19) of them and "no liver tumors" on the other 35 (= 5 + 30). Shuffle the cards and randomly deal 13 cards into one stack (to correspond to the *E. coli* mice) and 49 cards into a second stack. Next, count the number of cards in the "*E. coli* stack" that have the words "liver tumors" on them—to correspond to mice exposed to *E. coli* who develop liver tumors—and record whether this number is greater than or equal to 8. This process represents distributing 27 cases of liver tumors to two groups of mice (*E. coli* and germ free) randomly, with *E. coli* mice no more likely, nor any less likely, than germ-free mice to end up with liver tumors.

If we repeat this process many times (say, 10,000 times, with the aid of a computer in place of a physical deck of cards), it turns out that roughly 12% of the time we get 8 or more *E. coli* mice with liver tumors. Since something that happens 12% of the time is not terribly surprising, Table 1.1.2 does not provide significant evidence that exposure to *E. coli* increases the incidence of liver tumors. ■

In Chapter 10 we will discuss statistical techniques for evaluating data such as those in Tables 1.1.1 and 1.1.2. Of course, in some experiments variability is minimal and the message in the data stands out clearly without any special statistical analysis. It is worth noting, however, that absence of variability is itself an experimental result that must be justified by sufficient data. For instance, because Pasteur's anthrax data (Table 1.1.1) show no variability at all, it is intuitively plausible to conclude that the data provide "solid" evidence for the efficacy of the vaccination. But note that this conclusion involves a judgment; consider how much *less* "solid" the evidence would be if Pasteur had included only 3 animals in each group, rather than 24. Statistical analyses can be used to make such a judgment, that is, to determine if the variability is indeed negligible. Thus, a statistical view can be helpful even in the absence of variability.

The next two examples illustrate additional questions that a statistical approach can help to answer.

Example 1.1.3

Flooding and ATP In an experiment on root metabolism, a plant physiologist grew birch tree seedlings in the greenhouse. He flooded four seedlings with water for one day and kept four others as controls. He then harvested the seedlings and analyzed the roots for adenosine triphosphate (ATP). The measured amounts of ATP (nmoles per mg tissue) are given in Table 1.1.3 and displayed in Figure 1.1.1.[3]

Table 1.1.3 ATP concentration in birch tree roots (nmol/mg)

Flooded	Control
1.45	1.70
1.19	2.04
1.05	1.49
1.07	1.91

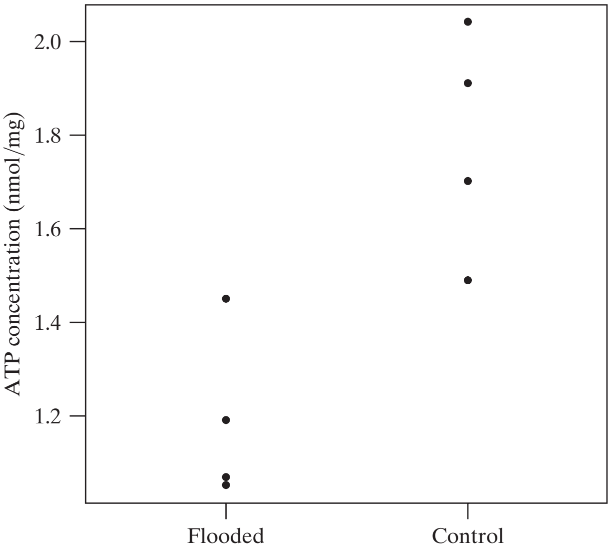


Figure 1.1.1 ATP concentration in birch tree roots

The data of Table 1.1.3 raise several questions: How should one summarize the ATP values in each experimental condition? How much information do the data provide about the effect of flooding? How confident can one be that the reduced ATP in the flooded group is really a response to flooding rather than just random variation? What size experiment would be required in order to firmly corroborate the apparent effect seen in these data? ■

Chapters 2, 6, and 7 address questions like those posed in Example 1.1.3. One question that we can address here is whether the data in Table 1.1.3 are consistent with the claim that flooding has no effect on ATP concentration, or instead provide significant evidence that flooding affects ATP concentrations. If the claim of no effect is true, then should we be surprised to see that all four of the flooded observations are smaller than each of the control observations? Might this happen by chance alone? If we wrote each of the numbers 1.05, 1.07, 1.19, 1.45, 1.49, 1.91, 1.70, and 2.04 on cards, shuffled the eight cards, and randomly dealt them into two piles, what is the chance that the four smallest numbers would end up in one pile and the four largest numbers in the other pile? It turns out that we could expect this to happen 1 time in 35 random shufflings, so "chance alone" would only create the kind of imbalance seen in Figure 1.1.1 about 2.9% of the time (since 1/35 = 0.029). Thus, we have some evidence that flooding has an effect on ATP concentration. We will develop this idea more fully in Chapter 7.

Example 1.1.4

MAO and Schizophrenia Monoamine oxidase (MAO) is an enzyme that is thought to play a role in the regulation of behavior. To see whether different categories of schizophrenic patients have different levels of MAO activity, researchers collected blood specimens from 42 patients and measured the MAO activity in the platelets. The results are given in Table 1.1.4 and displayed in Figure 1.1.2. (Values are expressed as nmol benzylaldehyde product per 10^8 platelets per hour.)[4] Note that it is much easier to get a feeling for the data by looking at the graph (Figure 1.1.2) than it is to read through the data in the table. The use of graphical displays of data is a very important part of data analysis. ■

Table 1.1.4 MAO activity in schizophrenic patients

Diagnosis	MAO activity				
I:	6.8	4.1	7.3	14.2	18.8
Chronic undifferentiated	9.9	7.4	11.9	5.2	7.8
schizophrenic	7.8	8.7	12.7	14.5	10.7
(18 patients)	8.4	9.7	10.6		
II:	7.8	4.4	11.4	3.1	4.3
Undifferentiated with	10.1	1.5	7.4	5.2	10.0
paranoid features	3.7	5.5	8.5	7.7	6.8
(16 patients)	3.1				
III:	6.4	10.8	1.1	2.9	4.5
Paranoid schizophrenic (8 patients)	5.8	9.4	6.8		

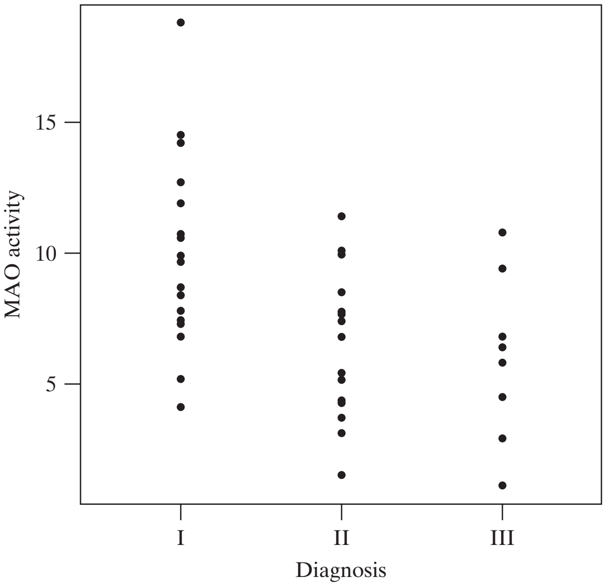


Figure 1.1.2 MAO activity in schizophrenic patients

To analyze the MAO data, one would naturally want to make comparisons among the three groups of patients, to describe the reliability of those comparisons, and to characterize the variability within the groups. To go beyond the data to a biological interpretation, one must also consider more subtle issues, such as the following: How were the patients selected? Were they chosen from a common hospital population, or were the three groups obtained at different times or places? Were precautions taken so that the person measuring the MAO was unaware of the patient's diagnosis? Did the investigators consider various ways of subdividing the patients before choosing the particular diagnostic categories used in Table 1.1.4? At first glance, these questions may seem irrelevant—can we not let the measurements speak for themselves? We will see, however, that the proper interpretation of data always requires careful consideration of how the data were obtained.

Chapters 2, 3, and 8 include discussions of selection of experimental subjects and of guarding against unconscious investigator bias. In Chapter 11 we will show how sifting through a data set in search of patterns can lead to serious misinterpretations and we will give guidelines for avoiding the pitfalls in such searches.

The next example shows how the effects of variability can distort the results of an experiment and how this distortion can be minimized by careful design of the experiment.

Example 1.1.5

Food Choice by Insect Larvae The clover root curculio, *Sitona hispidulus*, is a root-feeding pest of alfalfa. An entomologist conducted an experiment to study food choice by *Sitona* larvae. She wished to investigate whether larvae would preferentially choose alfalfa roots that were nodulated (their natural state) over roots whose nodulation had been suppressed. Larvae were released in a dish where both nodulated and nonnodulated roots were available. After 24 hours, the investigator counted the larvae that had clearly made a choice between root types. The results are shown in Table 1.1.5.[5]

The data in Table 1.1.5 appear to suggest rather strongly that *Sitona* larvae prefer nodulated roots. But our description of the experiment has obscured an important point—we have not stated how the roots were arranged. To see the relevance of the arrangement, suppose the experimenter had used only one dish, placing all the nodulated roots on one side of the dish and all the nonnodulated roots on the other side, as shown in Figure 1.1.3(a), and had then released 120 larvae in the center of the dish. This experimental arrangement would be seriously deficient, because the data of Table 1.1.5 would then permit several competing interpretations—for instance, (a) perhaps the larvae really do prefer nodulated roots; or (b) perhaps the two sides of the dish were at slightly different temperatures and the larvae were responding to temperature rather than nodulation; or (c) perhaps one larva chose the nodulated roots just by chance and the other larvae followed its trail. Because of these possibilities the experimental arrangement shown in Figure 1.1.3(a) can yield only weak information about larval food preference.

Table 1.1.5 Food choice by *Sitona* larvae

Choice	Number of larvae
Chose nodulated roots	46
Chose nonnodulated roots	12
Other (no choice, died, lost)	62
Total	120

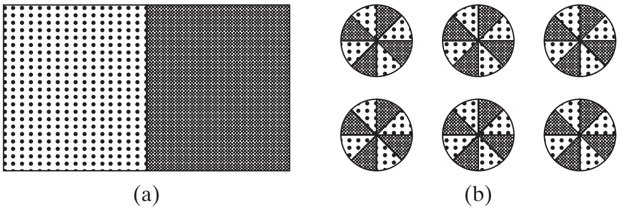


Figure 1.1.3 Possible arrangements of food choice experiment. The dark-shaded areas contain nodulated roots and the light-shaded areas contain nonnodulated roots.
(a) A poor arrangement.
(b) A good arrangement.

The experiment was actually arranged as in Figure 1.1.3(b), using six dishes with nodulated and nonnodulated roots arranged in a symmetric pattern. Twenty larvae were released into the center of each dish. This arrangement avoids the pitfalls of the arrangement in Figure 1.1.3(a). Because of the alternating regions of nodulated and nonnodulated roots, any fluctuation in environmental conditions (such as temperature) would tend to affect the two root types equally. By using several dishes, the experimenter has generated data that can be interpreted even if the larvae do tend to follow each other. To analyze the experiment properly, we would need to know the results in each dish; the condensed summary in Table 1.1.5 is not adequate. ■

In Chapter 11 we will describe various ways of arranging experimental material in space and time so as to yield the most informative experiment, as well as how to analyze the data to extract as much information as possible and yet resist the temptation to overinterpret patterns that may represent only random variation.

The following example is a study of the relationship between two measured quantities.

Example 1.1.6

Body Size and Energy Expenditure How much food does a person need? To investigate the dependence of nutritional requirements on body size, researchers used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure during conditions of quiet sedentary activity; this was repeated twice for each subject. The results are shown in Table 1.1.6 and plotted in Figure 1.1.4.[6]

Table 1.1.6 Fat-free mass and energy expenditure

Subject	Fat-free mass (kg)	24-hour energy expenditure (kcal)	
1	49.3	1,851	1,936
2	59.3	2,209	1,891
3	68.3	2,283	2,423
4	48.1	1,885	1,791
5	57.6	1,929	1,967
6	78.1	2,490	2,567
7	76.1	2,484	2,653

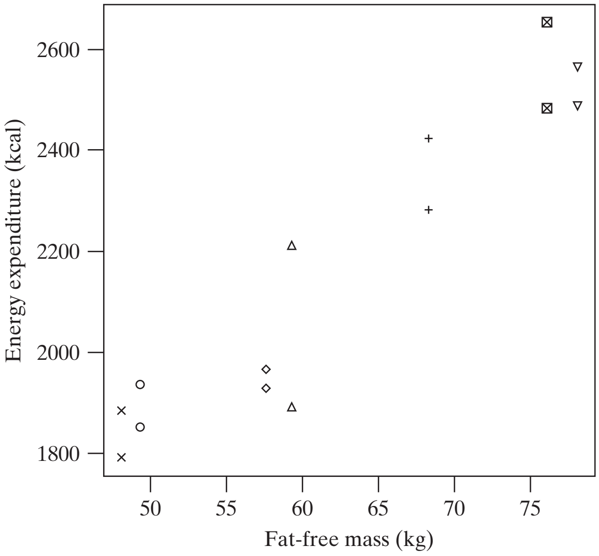


Figure 1.1.4 Fat-free mass and energy expenditure in seven men. Each man is represented by a different symbol.

A primary goal in the analysis of these data would be to describe the relationship between fat-free mass and energy expenditure—to characterize not only the overall trend of the relationship, but also the degree of scatter or variability in the relationship. (Note also that, to analyze the data, one needs to decide how to handle the duplicate observations on each subject.) ■

The focus of Example 1.1.6 is on the relationship between two variables: fat-free mass and energy expenditure. Chapter 12 deals with methods for describing such relationships, and also for quantifying the reliability of the descriptions.

A Look Ahead

Where appropriate, statisticians make use of the computer as a tool in data analysis; computer-generated output and statistical graphics appear throughout this book. The computer is a powerful tool, but it must be used with caution. Using the computer to perform calculations allows us to concentrate on concepts. The danger when using a computer in statistics is that we will jump straight to the calculations without looking closely at the data and asking the right questions about the data. Our goal is to analyze, understand, and interpret data—which are numbers *in a specific context*—not just to perform calculations.

In order to understand a data set it is necessary to know how and why the data were collected. In addition to considering the most widely used methods in statistical inference, we will consider issues in data collection and experimental design. Together, these topics should provide the reader with the background needed to read the scientific literature and to design and analyze simple research projects.

The preceding examples illustrate the kind of data to be considered in this book. In fact, each of the examples will reappear as an exercise or example in an appropriate chapter. As the examples show, research in the life sciences is usually concerned with the comparison of two or more groups of observations, or with the relationship between two or more variables. We will begin our study of statistics by focusing on a simpler situation—observations of a *single* variable for a *single* group. Many of the basic ideas of statistics will be introduced in this oversimplified context. Two-group comparisons and more complicated analyses will then be discussed in Chapter 7 and later chapters.

1.2 Types of Evidence

Researchers gather information and make inferences about the state of nature in a variety of settings. Much of statistics deals with the *analysis* of data, but statistical considerations often play a key role in the planning and *design* of a scientific investigation. We begin with examples of the three major kinds of evidence that one encounters.

Example 1.2.1

Lightning and Deafness On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing, which led to a headline in the *New York Times*, "Lightning Cures Deafness."[7] Is this compelling evidence that lightning is a cure for deafness? Could this event have been a coincidence? Are there other explanations for her cure? ■

The evidence discussed in Example 1.2.1 is **anecdotal evidence**. An anecdote is a short story or an example of an interesting event, in this case, of lightning curing deafness. The accumulation of anecdotes often leads to conjecture and to scientific investigation, but it is predictable pattern, not anecdote, that establishes a scientific theory.

Example 1.2.2 Sexual Orientation Some research has suggested that there is a genetic basis for sexual orientation. One such study involved measuring the midsagittal area of the anterior commissure (AC) of the brain for 30 homosexual men, 30 heterosexual men, and 30 heterosexual women. The researchers found that the AC tends to be larger in heterosexual women than in heterosexual men and that it is even larger in homosexual men. These data are summarized in Table 1.2.1 and are shown graphically in Figure 1.2.1.

Table 1.2.1 Midsagittal area of the anterior commissure (mm^2)

Group	Average midsagittal area (mm^2) of the anterior commissure
Homosexual men	14.20
Heterosexual men	10.61
Heterosexual women	12.03

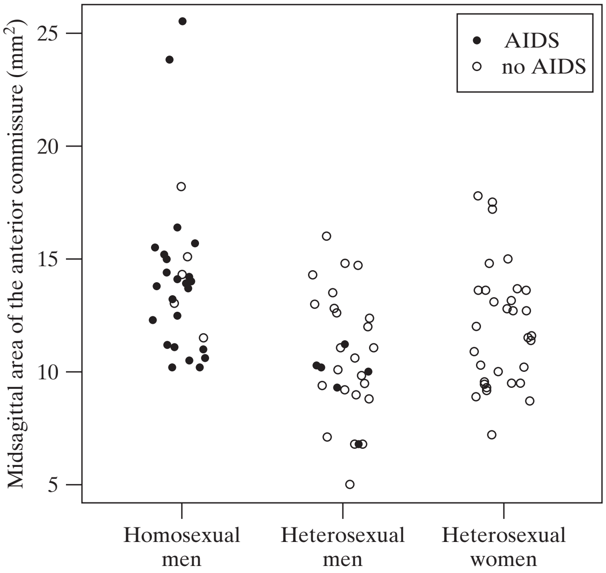


Figure 1.2.1 Midsagittal area of the anterior commissure (mm^2)

The data suggest that the size of the AC in homosexual men is more like that of heterosexual women than that of heterosexual men. When analyzing these data, we should take into account two things. (1) The measurements for two of the homosexual men were much larger than any of the other measurements; sometimes one or two such outliers can have a big impact on the conclusions of a study. (2) Twenty-four of the 30 homosexual men had AIDS, as opposed to 6 of the 30 heterosexual men; if AIDS affects the size of the anterior commissure, then this factor could account for some of the difference between the two groups of men.[8] ■

Example 1.2.2 presents an **observational study**. In an observational study the researcher systematically collects data from subjects, but only as an observer and not as someone who is manipulating conditions. By systematically examining all the data that arise in observational studies, one can guard against selectively viewing and reporting only evidence that supports a previous view. However, observational studies can be misleading due to *confounding variables*. In Example 1.2.2 we noted that having AIDS may affect the size of the anterior commissure. We would say that the effect of AIDS is confounded with the effect of sexual orientation in this study.

Note that the *context* in which the data arose is of central importance in statistics. This is quite clear in Example 1.2.2. The numbers themselves can be used to compute averages or to make graphs, like Figure 1.2.1, but if we are to understand what the data have to say, we must have an understanding of the context in which they arose. This context tells us to be on the alert for the effects that other factors, such as the impact of AIDS, may have on the size of the anterior commissure. Data analysis without reference to context is meaningless.

Example 1.2.3

Health and Marriage A study conducted in Finland found that people who were married at midlife were less likely to develop cognitive impairment (particularly Alzheimer's disease) later in life.[9] However, from an observational study such as this we don't know whether marriage *prevents* later problems or whether persons who are likely to develop cognitive problems are less likely to get married.

Example 1.2.4

Toxicity in Dogs Before new drugs are given to human subjects, it is common practice to first test them in dogs or other animals. In part of one study, a new investigational drug was given to eight male and eight female dogs at doses of 8 mg/kg and 25 mg/kg. Within each sex, the two doses were assigned at random to the eight dogs. Many "endpoints" were measured, such as cholesterol, sodium, glucose, and so on, from blood samples, in order to screen for toxicity problems in the dogs before starting studies on humans. One endpoint was alkaline phosphatase level (or APL, measured in U/l). The data are shown in Table 1.2.2 and plotted in Figure 1.2.2.[10]

Table 1.2.2 Alkaline phosphatase level (U/l)

Dose (mg/kg)	Male	Female
8	171	150
	154	127
	104	152
	143	105
Average	**143**	**133.5**
25	80	101
	149	113
	138	161
	131	197
Average	**124.5**	**143**

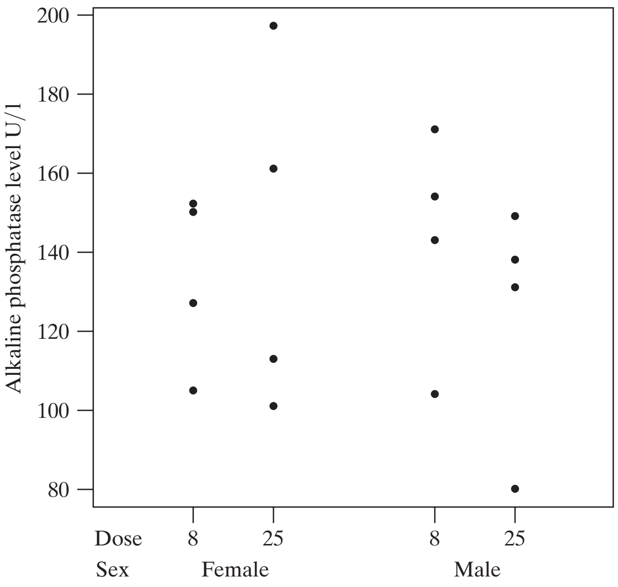


Figure 1.2.2 Alkaline phosphatase level in dogs

The design of this experiment allows for the investigation of the interaction between two factors: sex of the dog and dose. These factors interacted in the following sense: For females, the effect of increasing the dose from 8 to 25 mg/kg was positive, although small (the average APL increased from 133.5 to 143 U/l), but for males the effect of increasing the dose from 8 to 25 mg/kg was negative (the average APL dropped from 143 to 124.5 U/l). Techniques for studying such interactions will be considered in Chapter 11.

Example 1.2.4 presents an **experiment**, in that the researchers imposed the conditions—in this case, doses of a drug—on the subjects (the dogs). By randomly assigning treatments (drug doses) to subjects (dogs), we can get around the problem of confounding that complicates observational studies and limits the conclusions that we can reach from them. Randomized experiments are considered the "gold standard" in scientific investigation, but they can also be plagued by difficulties.

Often human subjects in experiments are given a **placebo**—an inert substance, such as a sugar pill. It is well known that people often exhibit a *placebo response*; that is, they tend to respond favorably to *any* treatment, even if it is only inert. This psychological effect can be quite powerful. Research has shown that placebos are effective for roughly one-third of people who are in pain; that is, one-third of pain sufferers report their pain ending after being giving a "painkiller" that is, in fact, an inert pill. For diseases such as bronchial asthma, angina pectoris (recurrent chest pain caused by decreased blood flow to the heart), and ulcers, the use of placebos has been shown to produce clinically beneficial results in over 60% of patients.[11] Of course, if a placebo control is used, then the subjects must not be told which group they are in—the group getting the active treatment or the group getting the placebo.

Example 1.2.5

Autism Autism is a serious condition in which children withdraw from normal social interactions and sometimes engage in aggressive or repetitive behavior. In 1997, an autistic child responded remarkably well to the digestive enzyme secretin. This led to an experiment (a "clinical trial") in which secretin was compared to a placebo. In this experiment, children who were given secretin improved considerably. However, the children given the placebo also improved considerably. There was no statistically significant difference between the two groups. Thus, the favorable response in the secretin group was considered to be only a "placebo response," meaning, unfortunately, that secretin was not found to be beneficial (beyond inducing a positive response associated simply with taking a substance as part of an experiment).[12] ■

The word *placebo* means "I shall please." The word *nocebo* ("I shall harm") is sometimes used to describe adverse reactions to perceived, but nonexistent, risks. The following example illustrates the strength that psychological effects can have.

Example 1.2.6

Bronchial Asthma A group of patients suffering from bronchial asthma were given a substance that they were told was a chest-constricting chemical. After being given this substance, several of the patients experienced bronchial spasms. However, during part of the experiment, the patients were given a substance that they were told would alleviate their symptoms. In this case, bronchial spasms were prevented. In reality, the second substance was identical to the first substance: Both were distilled water. It appears that it was the power of suggestion that brought on the bronchial spasms; the same power of suggestion prevented spasms.[13] ■

Similar to placebo treatment is *sham* treatment, which can be used on animals as well as humans. An example of sham treatment is injecting control animals with an inert substance such as saline. In some studies of surgical treatments, control animals (even, occasionally, humans) are given a "mock" surgery.

Example 1.2.7

Mammary Artery Ligation In the 1950s, the surgical technique of internal mammary artery ligation became a popular treatment for patients suffering from angina pectoris. In this operation the surgeon would ligate (tie) the mammary artery, with the goal of increasing collateral blood flow to the heart. Doctors and patients alike enthusiastically endorsed this surgery as an effective treatment. In 1958, studies of internal mammary artery ligation in animals found that it was not effective and this raised doubts about its usefulness on humans. A study was conducted in which patients were randomly assigned to one of two groups. Patients in the treatment group received the standard surgery. Patients in the control group received a sham operation in which an incision was made, the mammary artery was exposed as in the real operation, but the incision was closed *without* the artery being ligated. These patients had no way of knowing that their operation was a sham. The rates of improvement in the two groups of patients were nearly identical. (Patients who had the sham operation did slightly better than patients who had the real operation, but the difference was small.) A second randomized, controlled study also found that patients who received the sham surgery did as well as those who had the real operation. As a result of these studies, physicians stopped using internal mammary artery ligation.[14] ■

Blinding

In experiments on humans, particularly those that involve the use of placebos, **blinding** is often used. This means that the treatment assignment is kept secret from the experimental subject. The purpose of blinding the subject is to minimize the extent to which his or her expectations influence the results of the experiment. If subjects exhibit a psychological reaction to getting a medication, that placebo response will tend to balance out between the two groups, so that any difference between the groups can be attributed to the effect of the active treatment.

In many experiments the persons who evaluate the responses of the subjects are also kept blind; that is, during the experiment they are kept ignorant of the treatment assignment. Consider, for instance, the following:

> In a study to compare two treatments for lung cancer, a radiologist reads X-rays to evaluate each patient's progress. The X-ray films are coded so that the radiologist cannot tell which treatment each patient received.
>
> Mice are fed one of three diets; the effects on their liver are assayed by a research assistant who does not know which diet each mouse received.

Of course, *someone* needs to keep track of which subject is in which group, but that person should not be the one who measures the response variable. The most obvious reason for blinding the person making the evaluations is to reduce the possibility of subjective bias influencing the observation process itself: Someone who *expects* or *wants* certain results may unconsciously influence those results. Such bias can enter even apparently "objective" measurements through subtle variation in dissection techniques, titration procedures, and so on.

In medical studies of human beings, blinding often serves additional purposes. For one thing, a patient must be asked whether he or she consents to participate in a medical study. If the physician who asks the question already knows which treatment the patient would receive, then by discouraging certain patients and encouraging others, the physician can (consciously or unconsciously) create noncomparable treatment groups. The effect of such biased assignment can be surprisingly large, and it has been noted that it generally favors the "new" or "experimental" treatment.[15] Another reason for blinding in medical studies is that a physician may (consciously or unconsciously) provide more psychological encouragement, or even better care, to the patients who are receiving the treatment that the physician regards as superior.

An experiment in which both the subjects and the persons making the evaluations of the response are blinded is called a **double-blind** experiment. The first mammary artery ligation experiment described in Example 1.2.7 was conducted as a double-blind experiment.

The Need for Control Groups

Example 1.2.8

Clofibrate An experiment was conducted in which subjects were given the drug clofibrate, which was intended to lower cholesterol and reduce the chance of death from coronary disease. The researchers noted that many of the subjects did not take all the medication that the experimental protocol called for them to take. They calculated the percentage of the prescribed capsules that each subject took and divided the subjects into two groups according to whether or not the subjects took at least 80% of the capsules they were given. Table 1.2.3 shows that the five-year mortality rate for those who took at least 80% of their capsules was much lower than the corresponding rate for subjects who did not adhere to the protocol. On the surface, this suggests that taking the medication lowers the chance of death. However, there was a placebo control group in the experiment and many of the placebo subjects took fewer than 80% of their capsules. The mortality rates for the two placebo groups—those who adhered to the protocol and those who did not—are quite similar to the rates for the clofibrate groups.

Table 1.2.3 Mortality rates for the clofibrate experiment

	Clofibrate		Placebo	
Adherence	*n*	5-year mortality	*n*	5-year mortality
≥80%	708	15.0%	1813	15.1%
<80%	357	24.6%	882	28.2%

The clofibrate experiment seems to indicate that there are two kinds of subjects: those who adhere to the protocol and those who do not. The first group had a much lower mortality rate than the second group. This might be due simply to better health habits among people who are willing to follow a scientific protocol for five years than among people who don't adhere to the protocol. A further conclusion from the experiment is that clofibrate does not appear to be any more effective than placebo in reducing the death rate. Were it not for the presence of the placebo control group, the researchers might well have drawn the wrong conclusion from the study and attributed the lower death rate among adherers to clofibrate itself, rather than to other confounded effects that make the adherers different from the nonadherers.[16] ■

Example 1.2.9

The Common Cold Many years ago, investigators invited university students who believed themselves to be particularly susceptible to the common cold to be part of an experiment. Volunteers were randomly assigned to either the treatment group, in which case they took capsules of an experimental vaccine, or to the control group, in which case they were told that they were taking a vaccine, but in fact were given a placebo—capsules that looked like the vaccine capsules but that contained lactose in place of the vaccine.[17] As shown in Table 1.2.4, both groups reported having dramatically fewer colds during the study than they had had in the previous year.

Table 1.2.4 Number of colds in cold-vaccine experiment

	Vaccine	Placebo
n	201	203
Average number of colds		
Previous year (from memory)	5.6	5.2
Current year	1.7	1.6
% reduction	70%	69%

The average number of colds per person dropped 70% in the treatment group. This would have been startling evidence that the vaccine had an effect, except that the corresponding drop in the control group was 69%. ■

We can attribute much of the large drop in colds in Example 1.2.9 to the placebo effect. However, another statistical concern is **panel bias**, which is bias attributable to the study having influenced the behavior of the subjects—that is, people who know they are being studied often change their behavior. The students in this study reported from memory the number of colds they had suffered in the previous year. The fact that they were part of a study might have influenced their behavior, so that they were less likely to catch a cold during the study. Being in a study might also have affected the way in which they defined having a cold—during the study, they were "instructed to report to the health service whenever a cold developed"—so that some illness may have gone unreported during the study. (How sick do you have to be before you classify yourself as having a cold?)

Historical Controls

Researchers may be particularly reluctant to use randomized allocation in medical experiments on human beings. Suppose, for instance, that researchers want to evaluate a promising new treatment for a certain illness. It can be argued that it would be unethical to withhold the treatment from any patients, and that therefore all current patients should receive the new treatment. But then who would serve as a control group? One possibility is to use historical controls—that is, previous patients with the same illness who were treated with another therapy. One difficulty with historical controls is that there is often a tendency for later patients to show a better response—even to the same therapy—than earlier patients with the same diagnosis. This tendency has been confirmed, for instance, by comparing experiments conducted at the same medical centers in different years.[18] One major reason for the tendency is that the overall characteristics of the patient population may change with time. For instance, because diagnostic techniques tend to improve, patients with a given diagnosis (say, breast cancer) in 2001 may have a better chance of recovery (even with the same treatment) than those with the same diagnosis in 1991, because they were diagnosed earlier in the course of the disease.

Medical researchers do not agree on the validity and value of historical controls. The following example illustrates the importance of this controversial issue.

Example 1.2.10

Coronary Artery Disease Disease of the coronary arteries is often treated by surgery (such as bypass surgery), but it can also be treated with drugs only. Many studies have attempted to evaluate the effectiveness of surgical treatment for this common disease. In a review of 29 of these studies, each study was classified as to whether it used randomized controls or historical controls; the conclusions of the 29 studies are summarized in Table 1.2.5.[19]

Table 1.2.5 Coronary artery disease studies

	Conclusion about effectiveness of surgery		
Type of controls	Effective	Not effective	Total number of studies
Randomized	1	7	8
Historical	16	5	21

It would appear from Table 1.2.5 that enthusiasm for surgery is much more common among researchers who use historical controls than among those who use randomized controls. ■

Proponents of the use of historical controls argue that statistical adjustment can provide meaningful comparison between a current group of patients and a group of historical controls; for instance, if the current patients are younger than the historical controls, then the data can be analyzed in a way that adjusts, or corrects, for the effect of age. Critics reply that such adjustment may be grossly inadequate.

The concept of historical controls is not limited to medical studies. The issue arises whenever a researcher compares current data with past data. Whether the data are from the lab, the field, or the clinic, the researcher must confront the question: Can the past and current results be meaningfully compared? One should always at least ask whether the experimental material, and/or the environmental conditions, may have changed enough over time to distort the comparison.

Exercises 1.2.1–1.2.8

1.2.1 Fluoridation of drinking water has long been a controversial issue in the United States. One of the first communities to add fluoride to their water was Newburgh, New York. In March 1944, a plan was announced to begin to add fluoride to the Newburgh water supply on April 1 of that year. During the month of April, citizens of Newburgh complained of digestive problems, which were attributed to the fluoridation of the water. However, there had been a delay in the installation of the fluoridation equipment, so that fluoridation did not begin until May 2.[20] Explain how the placebo effect/nocebo effect is related to this example.

1.2.2 Olestra is a no-calorie, no-fat additive that is used in the production of some potato chips. After the Food and Drug Administration approved the use of olestra, some consumers complained that olestra caused stomach cramps and diarrhea. A randomized, double-blind experiment was conducted in which some subjects were given bags of potato chips made with olestra and other subjects were given ordinary potato chips. In the olestra group, 38% of the subjects reported having gastrointestinal symptoms. However, in the group given regular potato chips the corresponding percentage was 37%. (The two percentages are not statistically significantly different.)[21] Explain how the placebo effect/nocebo effect is related to this example. Also explain why it was important for this experiment to be double-blind.

1.2.3 (Hypothetical) In a study of acupuncture, patients with headaches are randomly divided into two groups. One group is given acupuncture and the other group is given aspirin. The acupuncturist evaluates the effectiveness of the acupuncture and compares it to the results from the aspirin group. Explain how lack of blinding biases the experiment in favor of acupuncture.

1.2.4 Randomized, controlled experiments have found that vitamin C is not effective in treating terminal cancer patients.[22] However, a 1976 research paper reported that terminal cancer patients given vitamin C survived much longer than did historical controls. The patients treated with vitamin C were selected by surgeons from a group of cancer patients in a hospital.[23] Explain how this experiment was biased in favor of vitamin C.

1.2.5 On 3 November 2009, the blog lifehacker.com contained a posting by an individual with chronic toenail fungus. He remarked that after many years of suffering and trying all sorts of cures, he resorted to sanding his toenail as thin as he could tolerate, followed by daily application of vinegar and hydrogen-peroxide-soaked bandaids on his toenail. He repeated the vinegar peroxide bandaging for 100 days. After this time his nail grew out and the fungus was gone. Using the language of statistics, what kind of evidence is this? Is this convincing evidence that this procedure is an effective cure of toenail fungus?

1.2.6 For each of the following cases [(a), (b), and (c)],

(I) state whether the study should be observational or experimental.

(II) state whether the study should be run blind, double-blind, or neither. If the study should be run blind or double-blind, who should be blinded?

(a) An investigation of whether taking aspirin reduces one's chance of having a heart attack.

(b) An investigation of whether babies born into poor families (family income below $25,000) are more likely to weigh less than 5.5 pounds at birth than babies born into wealthy families (family income above $65,000).

(c) An investigation of whether the size of the midsagittal plane of the anterior commisssure (a part of the brain) of a man is related to the sexual orientation of the man.

1.2.7 (Hypothetical) In order to assess the effectiveness of a new fertilizer, researchers applied the fertilizer to the tomato plants on the west side of a garden but did not fertilize the plants on the east side of the garden. They later measured the weights of the tomatoes produced by each plant and found that the fertilized plants grew larger tomatoes than did the nonfertilized plants. They concluded that the fertilizer works.

(a) Was this an experiment or an observational study? Why?

(b) This study is seriously flawed. Use the language of statistics to explain the flaw and how this affects the validity of the conclusion reached by the researchers.

(c) Could this study have used the concept of blinding (i.e., does the word "blind" apply to this study)? If so, how? Could it have been double-blind? If so, how?

1.2.8 Reseachers studied 1,718 persons over age 65 living in North Carolina. They found that those who attended religious services regularly were more likely to have strong immune systems (as determined by the blood levels of the protein interleukin-6) than those who didn't.[24] Does this mean that attending religious services improves one's health? Why or why not?

1.3 Random Sampling

In order to address research questions with data, we first must consider how those data are to be gathered. How we gather our data has tremendous implications on our choice of analysis methods and even on the validity of our studies. In this section we will examine some common types of data-gathering methods with special emphasis on the **simple random sample**.

Samples and Populations

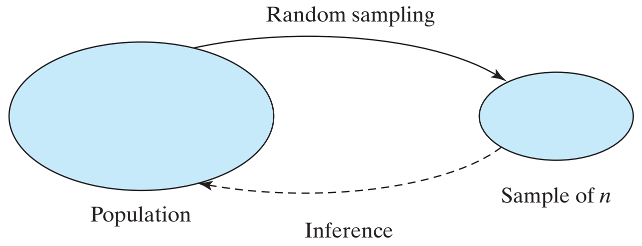
Before gathering data, we first consider the scope of our study by identifying the **population**. The population consists of all subjects/animals/specimens/plants, and so on, of interest. The following are all examples of populations:

- All birch tree seedlings in Florida
- All raccoons in Montaña de Oro State Park
- All people with schizophrenia in the United States
- All 100-ml water specimens in Chorro Creek

Typically we are unable to observe the entire population and therefore we must be content with gathering data from a subset of the population, a **sample** of size n. From this sample we make inferences about the population as a whole (see Figure 1.3.1). The following are all examples of samples:

- A selection of eight ($n = 8$) Florida birch seedlings grown in a greenhouse.
- Thirteen ($n = 13$) raccoons captured in traps at the Montaña de Oro campground.
- Forty-two ($n = 42$) schizophrenic patients who respond to an advertisement in a U.S. newspaper.
- Ten ($n = 10$) 100-ml vials of water collected one day at 10 locations along Chorro Creek.

Figure 1.3.1 Sampling from a population



Remark There is some potential for confusion between the statistical meaning of the term *sample* and the sense in which this word is sometimes used in biology. If a biologist draws blood from 20 people and measures the glucose concentration in each, she might say she has 20 samples of blood. However, the statistician says she has *one* sample of 20 glucose measurements; the sample size is $n = 20$. In the interest of clarity, throughout this book we will use the term *specimen* where a biologist might prefer *sample.* So we would speak of glucose measurements on a sample of 20 specimens of blood.

Ideally our sample will be a representative subset of the population; however, unless we are careful, we may end up obtaining a **biased** sample. A biased sample systematically overestimates or systematically underestimates a characteristic of the population. For example, consider the raccoons from the sample described previously that are captured in traps at a campground. These raccoons may systematically differ from the population; they may be larger (from having ample access to food from dumpsters and campers), less timid (from being around people who feed them), and may be even longer lived than the general population of raccoons in the entire park.

One method to ensure that samples will be (in the long run) representative of the population is to use random sampling.

Definition of a Simple Random Sample

Informally, the process of obtaining a simple random sample can be visualized in terms of labeled tickets, such as those used in a lottery or raffle. Suppose that each member of the population (e.g., raccoon, patient, plant) is represented by one ticket, and that the tickets are placed in a large box and thoroughly mixed. Then n tickets are drawn from the box by a blindfolded assistant, with new mixing after each ticket is removed. These n tickets constitute the sample. (Equivalently, we may visualize that n assistants reach in the box simultaneously, each assistant drawing one ticket.)

More abstractly, we may define random sampling as follows.

A Simple Random Sample

A **simple random sample** of n items is a sample in which (a) every member of the population has the same chance of being included in the sample, and (b) the members of the sample are chosen independently of each other. [Requirement (b) means that the chance of a given member of the population being chosen does not depend on which other members are chosen.]*

Simple random sampling can be thought of in other, equivalent, ways. We may envision the sample members being chosen one at a time from the population; under simple random sampling, at each stage of the drawing, every remaining member of the population is equally likely to be the next one chosen. Another view is to consider the totality of possible samples of size n. If all possible samples are equally likely to be obtained, then the process gives a simple random sample.

*Technically, requirement (b) is that every pair of members of the population has the same chance of being selected for the sample, every group of 3 members of the population has the same chance of being selected for the sample, and so on. In contrast to this, suppose we had a population with 30 persons in it and we wrote the names of 3 persons on each of 10 tickets. We could then choose one ticket in order to get a sample of size $n = 3$, but this would not be a simple random sample, since the pair (1,2) could end up in the sample but the pair (1,4) could not. Here the selections of members of the sample are not independent of each other. [This kind of sampling is known as "cluster sampling," with 10 clusters of size 3.] If the population is infinite, then the technical definition that all subsets of a given size are equally likely to be selected as part of the sample is equivalent to the requirement that the members of the sample are chosen independently.

Employing Randomness

When conducting statistical investigations, we will need to make use of randomness. As previously discussed, we obtain simple random samples randomly—every member of the population has the same chance of being selected. In Chapter 7 we shall discuss experiments in which we wish to compare the effects of different treatments on members of a sample. To conduct these experiments we will have to assign the treatments to subjects randomly—so that every subject has the same chance of receiving treatment A as they do treatment B.

Unfortunately, as a practical matter, humans are not very capable of mentally employing randomness. We are unable to eliminate unconscious bias that often leads us to systematically excluding or including certain individuals in our sample (or at least decreasing or increasing the chance of choosing certain individuals). For this reason, we must use external resources for selecting individuals when we want a random sample: mechanical devices such as dice, coins, and lottery tickets; electronic devices that produce random digits such as computers and calculators; or tables of random digits such as Table 1 in the back of this book. Although straightforward, using mechanical devices such as tickets in a box is impractical, so we will focus on the use of random digits for sample selection.

How to Choose a Random Sample

The following is a simple procedure for choosing a random sample of n items from a finite population of items.

(a) Create the **sampling frame**: a list of all members of the population with unique identification numbers for each member. All identification numbers must have the same number of digits; for instance, if the population contains 75 items, the identification numbers could be 01, 02, . . . , 75.

(b) Read numbers from Table 1, a calculator, or computer. Reject any numbers that do not correspond to any population member. (For example, if the population has 75 items that have been assigned identification numbers 01, 02, . . . , 75, then skip over the numbers 76, 77, . . . , 99 and 00.) Continue until n numbers have been acquired. (Ignore any repeated occurrence of the same number.)

(c) The population members with the chosen identification numbers constitute the sample.

The following example illustrates this procedure.

Example 1.3.1

Suppose we are to choose a random sample of size 6 from a population of 75 members. Label the population members 01, 02, . . . , 75. Use Table 1, a calculator, or a computer to generate a string of random digits.* For example, our calculator might produce the following string:

8 3 8 7 1 7 9 4 0 1 6 2 5 3 4 5 9 7 5 3 9 8 2 2

As we examine two-digit pairs of numbers, we ignore numbers greater than 75 as well as any pairs that identify a previously chosen individual.

~~8 3~~ ~~8 7~~ 1 7 ~~9 4~~ 0 1 6 2 5 3 4 5 ~~9 7~~ ~~5 3~~ ~~9 8~~ 2 2

Thus, the population members with the following identification numbers will constitute the sample: 17, 01, 62, 53, 45, 22. ■

*Most calculators generate random numbers expressed as decimal numbers between 0 and 1; to convert these to random digits, simply ignore the leading zero and decimal and read the digits that follow the decimal. To generate a long string of random digits, simply call the random number function on the calculator repeatedly.

Remark In calling the digits in Table 1 or your calculator or computer *random* digits, we are using the term *random* loosely. Strictly speaking, random digits are digits produced by a random *process*—for example, tossing a 10-sided die. The digits in Table 1 or in your calculator or computer are actually *pseudorandom* digits; they are generated by a deterministic (although possibly very complex) process that is designed to produce sequences of digits that mimic randomly generated sequences.

Remark If the population is large, then computer software can be quite helpful in generating a sample. If you need a random sample of size 15 from a population with 2,500 members, have the computer (or calculator) generate 15 random numbers between 1 and 2,500. (If there are duplicates in the set of 15, then go back and get more random numbers.)

Practical Concerns When Random Sampling

In many cases, obtaining a proper simple random sample is difficult or impossible. For example, to obtain a random sample of raccoons from Montaña de Oro State Park, one would first have to create the sampling frame, which provides a unique number for each raccoon in the park. Then, after generating the list of random numbers to identify our sample, one would have to capture those particular raccoons. This is likely an impossible task.

In practice, when it is possible to obtain a proper random sample, one should. When a proper random sample is impractical, it is important to take all precautions to ensure that the subjects in the study may be viewed *as if* they were obtained by random sampling from some population. That is, the sample should be comprised of individuals that all have the same chance of being selected from the population, and the individuals should be chosen independently. To do this, the first step is to define the population. The next step is to scrutinize the procedure by which the observational units are selected and to ask: Could the *observations* have been chosen at random? With the raccoon example, this might mean that we first define the population of raccoons by creating a sharp geographic boundary based on raccoon habitat and place traps at randomly chosen locations within the population habitat using a variety of baits and trap sizes. (We could use random numbers to generate latitude and longitude coordinates within the population habitat). While still less than ideal (some raccoons might be trap shy and baby raccoons may not enter the traps at all), this is certainly better than simply capturing raccoons at one nonrandomly chosen atypical location (e.g., the campground) within the park. Presumably, the vast majority of raccoons now have the same chance of being trapped (i.e., equally likely to be selected) and capturing one raccoon has little or no bearing on the capture of any other (i.e., they can be considered to be independently chosen). Thus, it seems reasonable to treat the observations as if they were chosen at random.

Nonsimple Random Sampling Methods

There are other kinds of sampling that are random in a sense, but that are not simple. Two common nonsimple random sampling techniques are the **random cluster sample** and **stratified random sample**. To illustrate the concept of a cluster sample, consider a modification to the lottery method of generating a simple random sample. With cluster sampling, rather than assigning a unique ticket (or ID number) for each

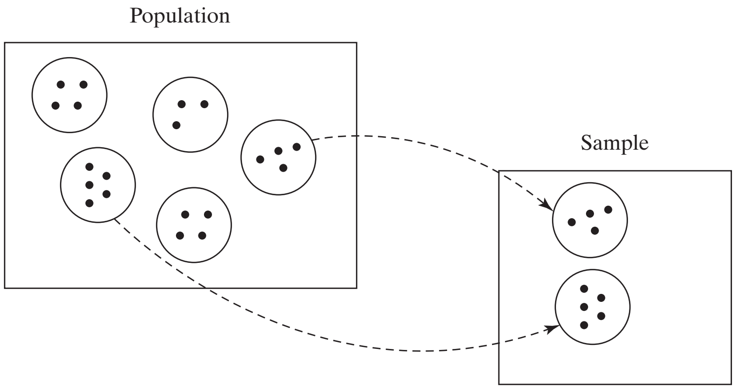


Figure 1.3.2 Random cluster sampling. The dots represent individuals within the population that are grouped into clusters (circles). Individuals in entire clusters are sampled from the population to form the sample.

member of the population, IDs are assigned to entire groups of individuals. As tickets are drawn from the box, entire groups of individuals are selected for the sample as in the following example and Figure 1.3.2.

Example 1.3.2

La Graciosa Thistle The La Graciosa thistle (*Cirsium loncholepis*) is an endangered plant native to the Guadalupe Dunes on the central coast of California. In a seed germination study, 30 plants were randomly chosen from the population of plants in the Guadalupe dunes and all seeds from the 30 plants were harvested. The seeds form a cluster sample from the population of all La Graciosa thistle seeds in Guadalupe while the individual plants were used to identify the clusters.[25] ■

A stratified random sample is chosen by first dividing the population into **strata**—homogeneous collections of individuals. Then, many simple random samples are taken—one within each stratum—and combined to comprise the sample (see Figure 1.3.3). The following is an example of a stratified random sample.

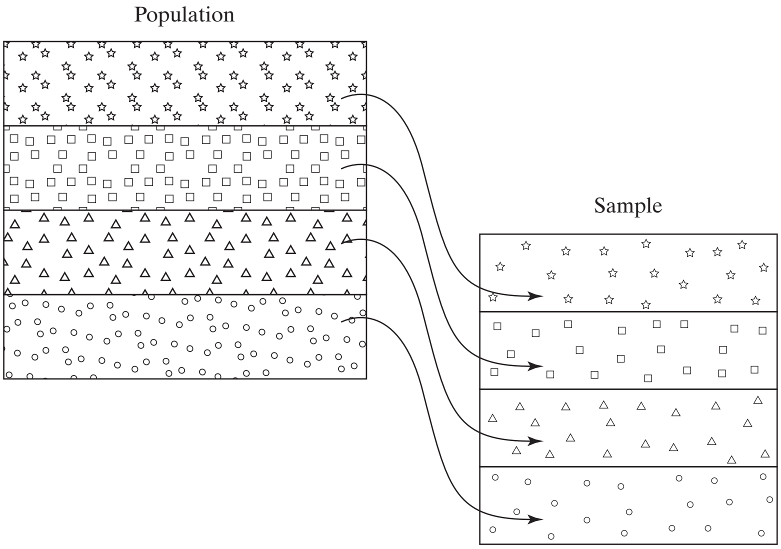


Figure 1.3.3 Stratified random sampling. The dots represent individuals within the population that are grouped into strata. Individuals from each stratum are randomly sampled and combined to form the sample.

Example 1.3.3

Sand Crabs In a study of parasitism of sand crabs (*Emerita analoga*), researchers obtained a stratified random sample of crabs by dividing a beach into 5-meter strips parallel to the water's edge. These strips were chosen as the strata because crab parasite loads may differ systematically based on the distance to the water's edge, thus making the parasite load for crabs within each stratum more similar than loads across strata. The first stratum was the 5-meter strip of beach just under the water's edge parallel to the shoreline. The second stratum was the 5-meter strip of beach just above the shoreline, followed by the third and fourth strata—the next two 5-meter strips above the shoreline. Within each strata, 25 crabs were randomly sampled, yielding a total sample size of 100 crabs.[26]

The majority of statistical methods discussed in this textbook will assume we are working with data gathered from a simple random sample. A sample chosen by simple random sampling is often called a *random sample.* But note that it is actually the *process* of sampling rather than the sample itself that is defined as random; randomness is not a property of the particular sample that happens to be chosen.

Sampling Error

How can we provide a rationale for inference from a limited sample to a much larger population? The approach of statistical theory is to refer to an idealized model of the sample–population relationship. In this model, which is called the **random sampling model**, the sample is chosen from the population by random sampling. The model is represented schematically in Figure 1.3.1.

The random sampling model is useful because it provides a basis for answering the question, How representative (of the population) is a sample likely to be? The model can be used to determine how much an inference might be influenced by chance, or "luck of the draw." More explicitly, a randomly chosen sample will usually not exactly resemble the population from which it was drawn. The discrepancy between the sample and the population is called **chance error due to sampling** or **sampling error**. We will see in later chapters how statistical theory derived from the random sampling model enables us to set limits on the likely amount of error due to sampling in an experiment. The quantification of such error is a major contribution that statistical theory has made to scientific thinking.

Because our samples are chosen randomly, there will always be sampling error present. If we sample nonrandomly, however, we may exacerbate the sampling error in unpredictable ways such as by introducing **sampling bias**, which is a systematic tendency for some individuals of the population to be selected more readily than others. The following two examples illustrate sampling bias.

Example 1.3.4

Lengths of Fish A biologist plans to study the distribution of body length in a certain population of fish in the Chesapeake Bay. The sample will be collected using a fishing net. Smaller fish can more easily slip through the holes in the net. Thus, smaller fish are less likely to be caught than larger ones, so that the sampling procedure is biased.

Example 1.3.5

Sizes of Nerve Cells A neuroanatomist plans to measure the sizes of individual nerve cells in cat brain tissue. In examining a tissue specimen, the investigator must decide which of the hundreds of cells in the specimen should be selected for measurement. Some of the nerve cells are incomplete because the microtome cut through them when the tissue was sectioned. If the size measurement can be made only on complete cells, a bias arises because the smaller cells had a greater chance of being missed by the microtome blade.

When the sampling procedure is biased, the sample may not accurately represent the population, because it is systematically distorted. For instance, in Example 1.3.4

smaller fish will tend to be underrepresented in the sample, so that the length of the fish in the sample will tend to be larger than those in the population.

The following example illustrates a kind of nonrandomness that is different from bias.

Example 1.3.6

Sucrose in Beet Roots An agronomist plans to sample beet roots from a field in order to measure their sucrose content. Suppose she were to take all her specimens from a randomly selected small area of the field. This sampling procedure would not be biased but would tend to produce *too homogeneous* a sample, because environmental variation across the field would not be reflected in the sample. ■

Example 1.3.6 illustrates an important principle that is sometimes overlooked in the analysis of data: In order to check applicability of the random sampling model, one needs to ask not only whether the sampling procedure might be biased, but also whether the sampling procedure will adequately reflect the variability inherent in the population. Faulty information about variability can distort scientific conclusions just as seriously as bias can.

We now consider some examples where the random sampling model might reasonably be applied.

Example 1.3.7

Fungus Resistance in Corn A certain variety of corn is resistant to fungus disease. To study the inheritance of this resistance, an agronomist crossed the resistant variety with a nonresistant variety and measured the degree of resistance in the progeny plants. The actual progeny in the experiment can be regarded as a random sample from a conceptual population of all *potential* progeny of that particular cross. ■

When the purpose of a study is to *compare* two or more experimental conditions, a very narrow definition of the population may be satisfactory, as illustrated in the next example.

Example 1.3.8

Nitrite Metabolism To study the conversion of nitrite to nitrate in the blood, researchers injected four New Zealand White rabbits with a solution of radioactively labeled nitrite molecules. Ten minutes after injection, they measured for each rabbit the percentage of the nitrite that had been converted to nitrate.[27] Although the four animals were not literally chosen at random from a specified population, nevertheless it might be reasonable to view the measurements of nitrite metabolism as a random sample from similar measurements made on all New Zealand White rabbits. (This formulation assumes that age and sex are irrelevant to nitrite metabolism.) ■

Example 1.3.9

Treatment of Ulcerative Colitis A medical team conducted a study of two therapies, A and B, for treatment of ulcerative colitis. All the patients in the study were referral patients in a clinic in a large city. Each patient was observed for satisfactory "response" to therapy. In applying the random sampling model, the researchers might want to make an inference to the population of all ulcerative colitis patients in urban referral clinics. First, consider inference about the actual probabilities of response; such an inference would be valid if the probability of response to each therapy is the same at all urban referral clinics. However, this assumption might be somewhat questionable, and the investigators might believe that the population should be defined very narrowly—for instance, as "the type of ulcerative colitis patients who are referred to this clinic." Even such a narrow population can be of interest in a comparative study. For instance, if treatment A is better than treatment B for the narrow population, it might be reasonable to infer that A would be better than B for a broader population (even if the actual response probabilities might be different in the broader population). In fact, it might even be argued that the broad population should include all ulcerative colitis patients, not merely those in urban referral clinics. ■

It often happens in research that, for practical reasons, the population actually studied is narrower than the population that is of real interest. In order to apply the kind of rationale illustrated in Example 1.3.9, one must argue that the results in the narrowly defined population (or, at least, some aspects of those results) can meaningfully be extrapolated to the population of interest. This extrapolation is not a *statistical* inference; it must be defended on biological, not statistical, grounds.

In Section 2.8 we will say more about the connection between samples and populations as we further develop the concept of statistical inference.

Nonsampling Errors

In addition to sampling errors, other concerns can arise in statistical studies. A **nonsampling error** is an error that is not caused by the sampling method; that is, a nonsampling error is one that would have arisen even if the researcher had a census of the entire population. For example, the way in which questions are worded can greatly influence how people answer them, as Example 1.3.10 shows.

Example 1.3.10

Abortion Funding In 1991, the U.S. Supreme Court made a controversial ruling upholding a ban on abortion counseling in federally financed family-planning clinics. Shortly after the ruling, a sample of 1,000 people were asked, "As you may know, the U.S. Supreme Court recently ruled that the federal government is not required to use taxpayer funds for family planning programs to perform, counsel, or refer for abortion as a method of family planning. In general, do you favor or oppose this ruling?" In the sample, 48% favored the ruling, 48% were opposed, and 4% had no opinion.

A separate opinion poll conducted at nearly the same time, but by a different polling organization, asked over 1,200 people, "Do you favor or oppose that Supreme Court decision preventing clinic doctors and medical personnel from discussing abortion in family-planning clinics that receive federal funds?" In this sample, 33% favored the decision and 65% opposed it.[28] The difference in the percentages favoring the opinion is too large to be attributed to chance error in the sampling. It seems that the way in which the question was worded had a strong impact on the respondents. ■

Another type of nonsampling error is **nonresponse bias**, which is bias caused by persons not responding to some of the questions in a survey or not returning a written survey. It is common to have only one-third of those receiving a survey in the mail complete the survey and return it to the researchers. (We consider the people receiving the survey to be part of the sample, even if some of them don't complete the entire survey, or even return the survey at all.) If the people who respond are unlike those who choose not to respond—and this is often the case, since people with strong feelings about an issue tend to complete a questionnaire, while others will ignore it—then the data collected will not accurately represent the population.

Example 1.3.11

HIV Testing A sample of 949 men were asked if they would submit to an HIV test of their blood. Of the 782 who agreed to be tested, 8 (1.02%) were found to be HIV positive. However, some of the men refused to be tested. The health researchers conducting the study had access to serum specimens that had been taken earlier from these 167 men and found that 9 of them (5.4%) were HIV positive.[29] Thus, those who refused to be tested were much more likely to have HIV than those who agreed to be tested. An estimate of the HIV rate based only on persons who agree to be tested is likely to substantially underestimate the true prevalence. ■

There are other cases in which an experimenter is faced with the vexing problem of **missing data**—that is, observations that were planned but could not be made. In addition to nonresponse, this can arise because experimental animals or plants die, because equipment malfunctions, or because human subjects fail to return for a follow-up observation.

A common approach to the problem of missing data is to simply use the remaining data and ignore the fact that some observations are missing. This approach is temptingly simple but must be used with extreme caution, because comparisons based on the remaining data may be seriously biased. For instance, if observations on some experimental mice are missing because the mice died of causes related to the treatment they received, it is obviously not valid to simply compare the mice that survived. As another example, if patients drop out of a medical study because they think their treatment is not working, then analysis of the remaining patients could produce a greatly distorted picture.

Naturally, it is best to make every effort to avoid missing data. But if data are missing, it is crucial that the possible reasons for the omissions be considered in interpreting and reporting the results.

Data can also be misleading if there is bias in how the data are collected. People have difficulty remembering the dates on which events happen and they tend to give unreliable answers if asked a question such as "How many times per week do you exercise?" They may also be biased as they make observations, as the following example shows.

Example 1.3.12

Sugar and Hyperactivity Mothers who thought that their young sons were "sugar sensitive" were randomly divided into two groups. Those in the first group were told that their sons had been given a large dose of sugar, whereas those in the second group were told that their sons had been given a placebo. In fact, all the boys had been given the placebo. Nonetheless, the mothers in the first group rated their sons to be much more hyperactive during a 25-minute study period than did the mothers in the second group.[30] Neutral measurements found that boys in the first group were actually a bit *less* active than those in the second group. Numerous other studies have failed to find a link between sugar consumption and activity in children, despite the widespread belief that sugar causes hyperactive behavior. It seems that the expectations that these mothers had colored their observations.[31] ■

Exercises 1.3.1–1.3.6

1.3.1 In each of the following studies, identify which sampling technique best describes the way the data were collected (or could be treated as if they were collected): simple random sampling, random cluster sampling, or stratified random sampling. For cluster samples identify the clusters and for stratified samples identify the strata.

(a) All 257 leukemia patients from three randomly chosen pediatric clinics in the United States were enrolled in a clinical trial for a new drug.

(b) A total of twelve 10-g soil specimens were collected from random locations on a farm to study physical and chemical soil profiles.

(c) In a pollution study three 100-ml air specimens were collected at each of four specific altitudes (100 m, 500 m, 1000 m, 2000 m) for a total of twelve 100-ml specimens.

(d) A total of 20 individual grapes were picked from random vines in a vineyard to evaluate readiness for harvest.

(e) Twenty-four dogs (eight randomly chosen small breed, eight randomly chosen medium breed, and eight randomly chosen large breed) were enrolled in an experiment to evaluate a new training program.

1.3.2 For each of the following studies, identify the source(s) of sampling bias and describe (i) how it might affect the study conclusions and (ii) how you might alter the sampling method to avoid the bias.

(a) Eight hundred volunteers were recruited from nightclubs to enroll in an experiment to evaluate a new treatment for social anxiety.

(b) In a water pollution study, water specimens were collected from a stream on 15 rainy days.

(c) To study the size (radius) distribution of scrub oaks (shrubby oak trees), 20 oak trees were selected by using random latitude/longitude coordinates. If the random coordinate fell within the canopy of a tree, the tree was selected; if not, another random location was generated.

(d) To study the size distribution of rock cod (*Epinephelus puscus*) off the coast of southeastern Australia, the lengths and weights were recorded for all cod captured by a commercial fishing vessel on one day (using standard hook-and-line fishing methods).

1.3.3 (A fun activity) Write the digits 1, 2, 3, 4 in order on an index card. Bring this card to a busy place (e.g., dining hall, library, university union) and ask at least 30 people to look at the card and select one of the digits at random in their head. Record their responses.

(a) If people can think "randomly," about what fraction of the people should respond with the digit 1? 2? 3? 4?

(b) What fraction of those surveyed responded with the digit 1? 2? 3? 4?

(c) Do the results suggest anything about people's ability to choose randomly?

1.3.4 Consider a population consisting of 600 individuals with unique IDs: 001, 002, . . . , 600. Use the following string of random digits to select a simple random sample of 5 individuals. List the IDs of the individuals selected for your sample.

7 2 8 1 2 1 8 7 6 4 4 2 1 2 1 5 9 3 7 8 7 8 0 3 5 4 7 2 1 6 5 9 6 8 5 1

1.3.5 (Sampling exercise) Refer to the collection of 100 ellipses shown in the accompanying figure, which can be thought of as representing a natural population of the mythical organism *C. ellipticus*. The ellipses have been given identification numbers 00, 01, . . ., 99 for convenience in sampling. Certain individuals of *C. ellipticus* are mutants and have two tail bristles.

(a) Use your *judgment* to choose a sample of size 10 from the population that you think is representative of the entire population. Note the number of mutants in the sample.

(b) Use *random digits* (from Table 1 or your calculator or computer) to choose a random sample of size 10 from the population and note the number of mutants in the sample.

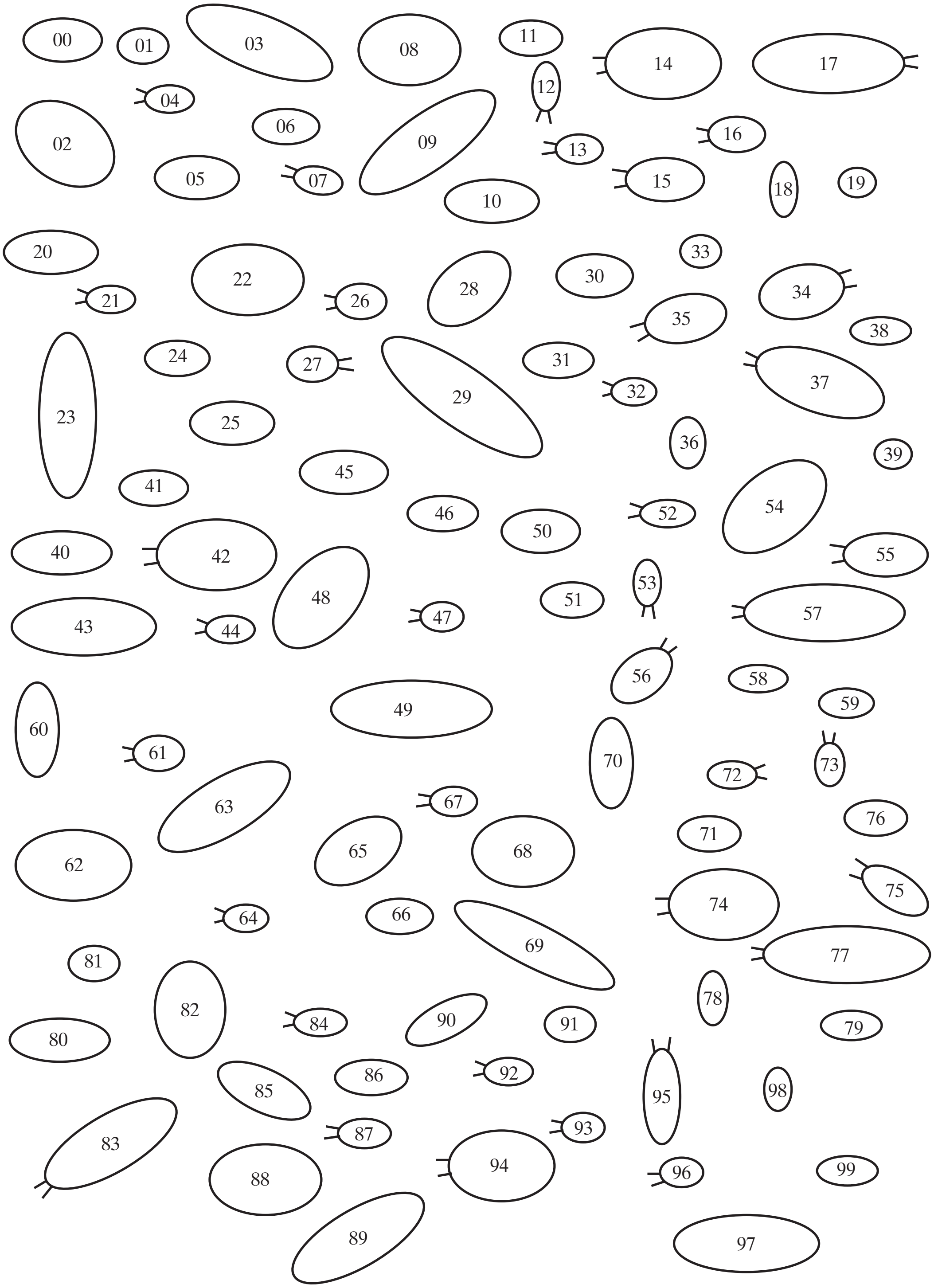
1.3.6 (Sampling exercise) Refer to the collection of 100 ellipses.

(a) Use random digits (for Table 1 or your calculator or computer) to choose a random sample of size 5 from the population and note the number of mutants in the sample.

(b) Repeat part (a) nine more times, for a total of 10 samples. (Some of the 10 samples may overlap.)

To facilitate pooling of results from the entire class, report your results in the following format:

NUMBER OF MUTANTS	NONMUTANTS	FREQUENCY (NO. OF SAMPLES)
0	5	
1	4	
2	3	
3	2	
4	1	
5	0	
		Total: 10


00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99