Pearson wishes to thank and acknowledge the following people for their work on the Global Edition:

Contributor

C. V. Vinay, *JSS Academy of Technical Education*
Dilip Nath, *Gauhati University*

Reviewers

D. V. Chandrashekhar, *Vivekananda Institute of Technology*
Sunil Jacob John, *National Institute of Technology Calicut*
D. V. Jayalakshmamma, *Vemana Institute of Technology*

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

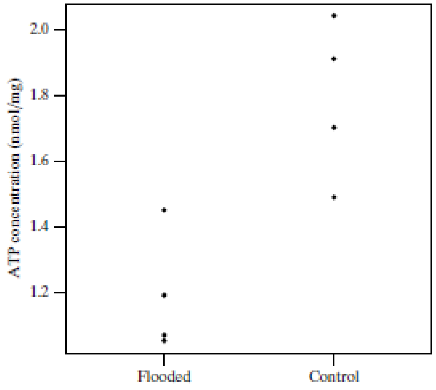


Figure 1.1.1 ATP concentration in birch tree roots

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 patients with schizophrenia 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 patients with schizophrenia

Diagnosis	MAO activity				
I:	6.8	4.1	7.3	14.2	18.8
Chronic	9.9	7.4	11.9	5.2	7.8
undifferentiated schizophrenia	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	10.1	1.5	7.4	5.2	10.0
with paranoid	3.7	5.5	8.5	7.7	6.8
features (16 patients)	3.1				
III:	6.4	10.8	1.1	2.9	4.5
Paranoid schizophrenia (8 patients)	5.8	9.4	6.8		

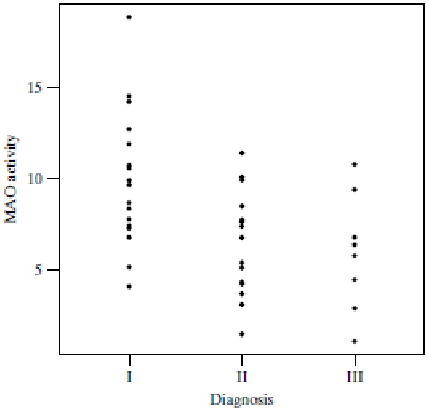


Figure 1.1.2 MAO activity in patients with schizophrenia

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.

Sections 1.2 and 1.3. as well as Chapters 2 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

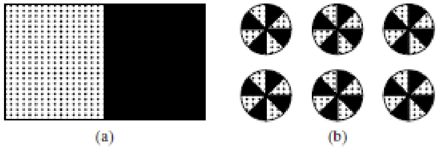


Figure 1.1.3 Possible arrangements of food choice experiment. The dark-shaded areas contain nodulated roots and the light-shaded areas contain nonnodulated roots.
(a) A poor arrangement.
(b) A good arrangement.

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

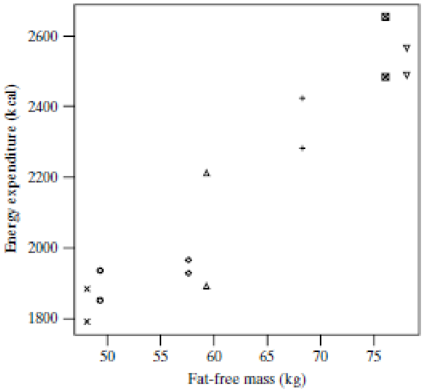


Figure 1.1.4 Fat-free mass and energy expenditure in seven men. Each man is represented by a different symbol.

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

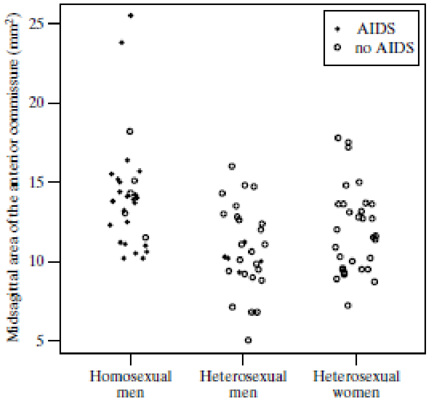


Figure 1.2.1 Midsagittal area of the anterior commissure (mm^2)

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**

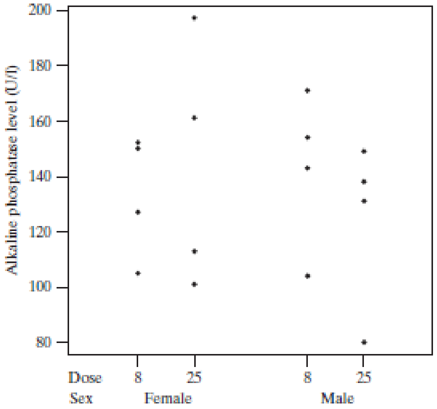


Figure 1.2.2 Alkaline phosphatase level in dogs

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

Renal Denervation A surgical procedure called "renal denervation" was developed to help people with hypertension who do not respond to medication. An early study suggested that renal denervation (which uses radiotherapy to destroy some nerves in arteries feeding the kidney) reduces blood pressure. In that experiment, patients who received surgery had an average improvement in systolic blood pressure of 33 mmHg more than did control patients who received no surgery. Later an experiment was conducted in which patients were randomly assigned to one of two groups. Patients in the treatment group received the renal denervation surgery. Patients in the control group received a sham operation in which a catheter was inserted, as in the real operation, but 20 minutes later the catheter was removed *without* radiotherapy being used. 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.[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. Suppose the physician who asks the question already knows which treatment the patient will receive. 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 5-year mortality rate for those who took at least 80% of their capsules was much lower than the corresponding rate for subjects who took fewer than 80% of the capsules. 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 show stronger adherence to a scientific protocol for 5 years than among people who only adhere weakly, if at all. 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 strong adherers to clofibrate itself, rather than to other confounded effects that make the strong 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. 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%.

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%

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?)

Example 1.2.10

Diet and Cancer Prevention A diet that is high in fruits and vegetables may yield many health benefits, but how can we be sure? During the 1990s, the medical community believed that such a diet would reduce the risk of cancer. This belief was based on comparisons from **case-control studies.** In such studies patients with cancer were matched with "control subjects"—persons of the same age, race, sex, and so on—who did not have cancer; then the diets of the two groups were compared, and it was found that the control patients ate more fruits and vegetables than did the cancer patients. This would seem to indicate that cancer rates go down as consumption of fruits and vegetables goes up. The use of case-control studies is quite sensible because it allows researchers to make comparisons (e.g., of diets, etc.) while taking into consideration important characteristics such as age.

Nonetheless, a case-control study is not perfect. Not all people agree to be interviewed and to complete health information surveys, and these individuals thus might be excluded from a case-control study. People who agree to be interviewed about their health are generally more healthy than those who decline to participate. In addition to eating more fruits and vegetables than the average person, they are also less likely to smoke and more likely to exercise.[18] Thus, even though case-control studies took into consideration age, race, and other characteristics, they overstated the benefits of fruits and vegetables. The observed benefits are likely also the result of other healthy lifestyle factors.* Drawing a cause–effect conclusion that fruit and vegetable consumption protects against cancer is dangerous.

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.[19] One major reason for the tendency is that the overall characteristics of the patient population may change with time. For

*A more informative kind of study is a prospective study or cohort study in which people with varying diets are followed over time to see how many of them develop cancer; however, such a study can be difficult to carry out.

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. This is one reason that patients diagnosed with kidney cancer in 1995 had a 61% chance of surviving for at least 5 years but those with the same diagnosis in 2005 had a 75% 5-year survival rate.[20]

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

Example 1.2.11 **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.[21]

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.

Example 1.2.12 **Healthcare Trials** A medical intervention, such as a new surgical procedure or drug, will often be used at one time in a nonrandomized clinical trial and at another time in a clinical trial of patients with the same condition who are assigned to groups randomly. Nonrandomized trials, which include the use of historical controls, tend to overstate the effectiveness of interventions. One analysis of many pairs of studies found that the nonrandomized trial showed a larger intervention effect than the corresponding randomized trial 22 times out of 26 comparisons; see Table 1.2.6.[22] Researchers concluded that overestimates of effectiveness are "due to poorer prognosis in non-randomly selected control groups compared with randomly selected control groups."[23] That is, if you give a new drug to relatively healthy patients and compare them to very sick patients taking the standard drug, the new drug is going to look better than it really is.

Even when randomization is used, trials may or may not be run double-blind. A review of 250 controlled trials found that trials that were not run double-blind produced significantly larger estimates of treatment effects than did trials that were double-blind.[24]

Table 1.2.6 Randomized versus nonrandomized trials

	Larger estimate of effect of the (common) intervention		
	Not randomized	Randomized	Total
Number of studies	22	4	26

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 1.2.1–1.2.10

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.[25] 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.)[26] 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.[27] 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.[28] 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)],

(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).

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

(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 the size of the midsagittal plane of the anterior commissure (a part of the brain) of a man is related to the sexual orientation of the man.

(b) An investigation of whether drinking more than 1 liter of water per day helps with weight loss for people who are trying to lose weight.

1.2.8 **(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?

1.2.9 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.[29] Does this mean that attending religious services improves one's health? Why or why not?

1.2.10 Researchers studied 300,818 golfers in Sweden and found that the "standardized mortality ratios" for golfers, adjusting for age, sex, and socioeconomic status, were lower than for nongolfers, meaning that golfers tend to live longer.[30] Does this mean that playing golf 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; 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.

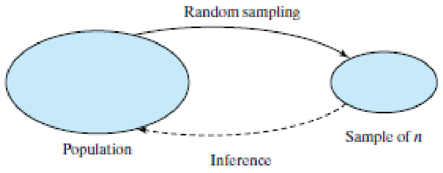


Figure 1.3.1 Sampling from a population

- Thirteen ($n = 13$) raccoons captured in traps at the Montaña de Oro campground.
- Forty-two ($n = 42$) patients with schizophrenia 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.

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.]*

*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.

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.

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 exclude or include certain individuals in our sample (or at least decrease or increase 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:

838717940162534597539822

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.

~~83~~ ~~87~~ 17 ~~94~~ 01 62 53 45 ~~97~~ ~~53~~ ~~98~~ 22

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

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.) Although 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**

*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.

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 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.

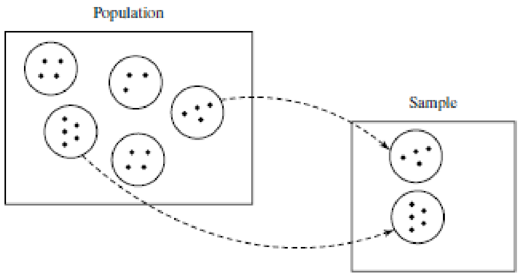


Figure 1.3.2 Random cluster sampling. The dots represent individuals within the population that are grouped into clusters (circles). Individuals in entire clusters are sampled from the population to form the sample.

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.[31]

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.

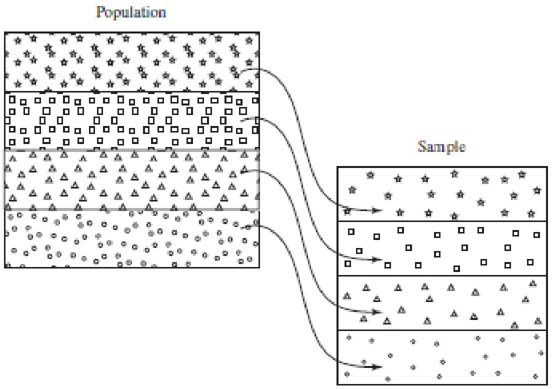


Figure 1.3.3 Stratified random sampling. The dots represent individuals within the population that are grouped into strata. Individuals from each stratum are randomly sampled and combined to form the sample.

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.[32]

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 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 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.[33] Although the four animals were not literally chosen at random from a specified population, it might be reasonable, nevertheless, 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.[34] 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.[35] 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.[36] 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.[37] ■

Exercises 1.3.1–1.3.7

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, one from each of 20 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.

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

(b) A nutritionist is interested in the eating habits of college students and observes what each student who enters a dining hall between 8:00 A.M. and 8:30 A.M. chooses for breakfast on a Monday morning.

(c) To study how fast an experimental painkiller relieves headache pain residents of a nursing home who complain of headaches are given the painkiller and are later asked how quickly their headaches subsided.

1.3.4 **(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?

1.3.5 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.

728121876442121593787803547216596851

1.3.6 **(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.

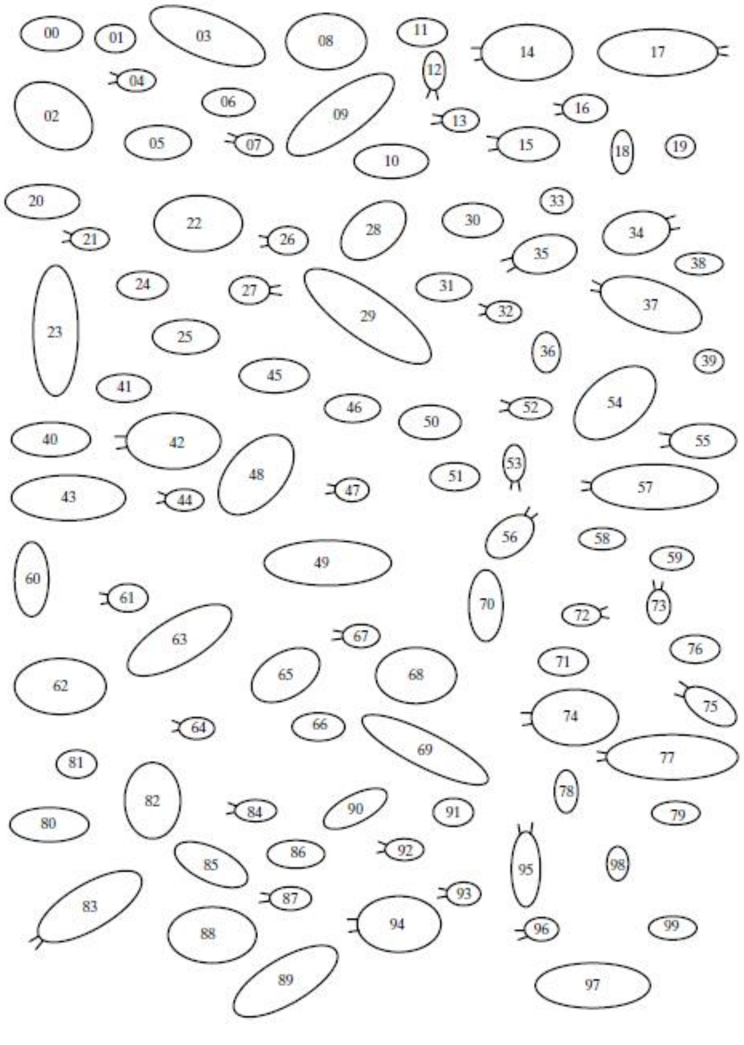
1.3.7 **(Sampling exercise)** Refer to the collection of 100 ellipses.

(a) Use random digits (from 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



Chapter 2

Description of Samples and Populations

OBJECTIVES

In this chapter we will study how to describe data. In particular, we will

- show how frequency distributions are used to make bar charts and histograms.
- compare the mean and median as measures of center.
- demonstrate how to construct and read a variety of graphics including dotplots, boxplots, and scatterplots.
- compare several measures of variability with emphasis on the standard deviation.
- examine how transformations of variables affect distributions.
- consider the relationship between populations and samples.

2.1 Introduction

Statistics is the science of analyzing and learning from data. In this section we introduce some terminology and notation for dealing with data.

VARIABLES

We begin with the concept of a **variable**. A variable is a characteristic of a person or a thing that can be assigned a number or a category. For example, blood type (A, B, AB, O) and age are two variables we might measure on a person.

Blood type is an example of a **categorical variable***: A categorical variable is a variable that records which of several categories a person or thing is in. Examples of categorical variables are

Blood type of a person: A, B, AB, O
Sex of a fish: male, female
Color of a flower: red, pink, white
Shape of a seed: wrinkled, smooth

Age is an example of a **numeric variable**, that is, a variable that records the amount of something. A **continuous variable** is a numeric variable that is measured on a continuous scale. Examples of continuous variables are

Weight of a baby
Cholesterol concentration in a blood specimen
Optical density of a solution

A variable such as weight is continuous because, in principle, two weights can be arbitrarily close together. Some types of numeric variables are not continuous but fall on a discrete scale, with spaces between the possible values. A **discrete variable** is a numeric variable for which we can list the possible values. For example, the number of eggs in a bird's nest is a discrete variable because only the values 0, 1, 2, 3, . . . , are possible. Other examples of discrete variables are

Number of bacteria colonies in a petri dish
Number of cancerous lymph nodes detected in a patient
Length of a DNA segment in basepairs

*For some categorical variables, the categories can be arrayed in a meaningful rank order. Such a variable is said to be **ordinal**. For example, the response of a patient to therapy might be none, partial, or complete.

The distinction between continuous and discrete variables is not a rigid one. After all, physical measurements are always rounded off. We may measure the weight of a steer to the nearest kilogram, of a rat to the nearest gram, or of an insect to the nearest milligram. The scale of the actual measurements is always discrete, strictly speaking. The continuous scale can be thought of as an approximation to the actual scale of measurement.

OBSERVATIONAL UNITS

When we collect a sample of *n* persons or things and measure one or more variables on them, we call these persons or things **observational units** or cases. The following are some examples of samples.

Sample	Variable	Observational unit
150 babies born in a certain hospital	Birthweight (kg)	A baby
73 *Cecropia* moths caught in a trap	Sex	A moth
81 plants that are a progeny of a single parental cross	Flower color	A plant
Bacterial colonies in each of six petri dishes	Number of colonies	A petri dish

NOTATION FOR VARIABLES AND OBSERVATIONS

We will adopt a notational convention to distinguish between a variable and an observed value of that variable. We will denote variables by uppercase letters such as Y. We will denote the observations themselves (that is, the data) by lowercase letters such as y. Thus, we distinguish, for example, between $Y =$ birthweight (the variable) and $y = 7.9$ lb (the observation). This distinction will be helpful in explaining some fundamental ideas concerning variability.

Exercises 2.1.1–2.1.5

For each of the following settings in Exercises 2.1.1–2.1.5, (i) identify the variable(s) in the study, (ii) for each variable tell the type of variable (e.g., categorical and ordinal, discrete, etc.), (iii) identify the observational unit (the thing sampled), and (iv) determine the sample size.

2.1.1

(a) A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*.

(b) The birthweight, date of birth, and the mother's race were recorded for each of 65 babies.

2.1.2

(a) A physician measured the height and weight of each of 37 children.

(b) During a blood drive, a blood bank offered to check the cholesterol of anyone who donated blood. A total of 129 persons donated blood. For each of them, the blood type and cholesterol levels were recorded.

2.1.3

(a) A biologist measured the number of leaves on each of 25 plants.

(b) A physician recorded the number of seizures that each of 20 patients with severe epilepsy had during an eight-week period.

2.1.4

(a) A conservationist recorded the weather (clear, partly cloudy, cloudy, rainy) and number of cars parked at noon at a trailhead on each of 18 days.

(b) An enologist measured the pH and residual sugar content (g/l) of seven barrels of wine.

2.1.5

(a) A biologist measured the body mass (g) and sex of each of 123 blue jays.

(b) A biologist measured the lifespan (in days), the thorax length (in mm), and the percent of time spent sleeping for each of 125 fruit flies.

2.2 Frequency Distributions

A first step toward understanding a set of data on a given variable is to explore the data and describe the data in summary form. In this chapter we discuss three mutually complementary aspects of data description: frequency distributions, measures of center, and measures of dispersion. These tell us about the shape, center, and spread of the data.

A **frequency distribution** is simply a display of the **frequency**, or number of occurrences, of each value in the data set. The information can be presented in tabular form or, more vividly, with a graph. A **bar chart** is a graph of categorical data showing the number of observations in each category. Here are two examples of frequency distributions for categorical data.

Example 2.2.1

Color of Poinsettias Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.[1] The bar graph in Figure 2.2.1 is a visual display of the results given in Table 2.2.1.

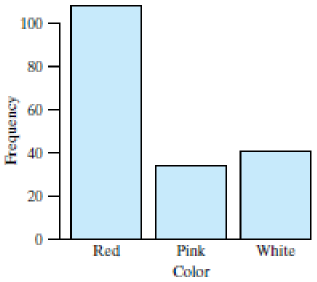


Figure 2.2.1 Bar chart of color of 182 poinsettias

Table 2.2.1 Color of 182 poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

Example 2.2.2

School Bags and Neck Pain Physiologists in Australia were concerned that carrying a school bag loaded with heavy books was a cause of neck pain in adolescents, so they asked a sample of 585 teenage girls how often they get neck pain when carrying their school bag (never, almost never, sometimes, often, always). A summary of the results reported to them is given in Table 2.2.2 and displayed as a bar graph in Figure 2.2.2(a).[2] As the variable incidence is an ordinal categorical variable, our tables and graphs should respect the natural ordering. Figure 2.2.2(b) shows the same data but with the categories in alphabetical order (a default setting for much software), which obscures the information in the data.

Table 2.2.2 Neck pain associated with carrying a school bag

Incidence	Frequency (number of girls)
Never	179
Almost never	159
Sometimes	173
Often	64
Always	10
Total	585

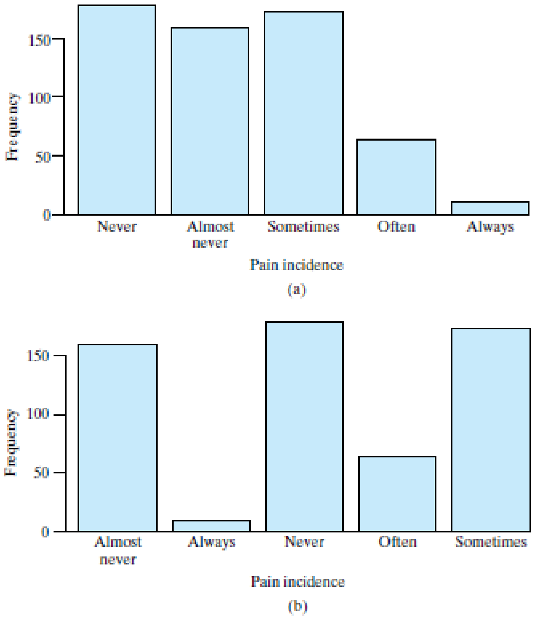


Figure 2.2.2 (a) Bar chart of incidence of neck pain reported by 585 adolescents; (b) the same data but with the categories in alphabetical order

A **dotplot** is a simple graph that can be used to show the distribution of a numeric variable when the sample size is small. To make a dotplot, we draw a number line covering the range of the data and then put a dot above the number line for each observation, as the following example shows.

Example 2.2.3

Infant Mortality Table 2.2.3 shows the infant mortality rate (infant deaths per 1,000 live births) in each of seven countries in South Asia, as of 2013.[3] The distribution is shown in Figure 2.2.3. ■

Table 2.2.3 Infant mortality in seven South Asian countries

Country	Infant mortality rate (deaths per 1,000 live births)
Bangladesh	47.3
Bhutan	40.0
India	44.6
Maldives	25.5
Nepal	41.8
Pakistan	59.4
Sri Lanka	9.2

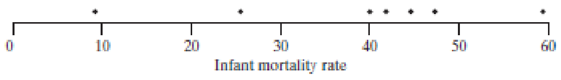


Figure 2.2.3 Dotplot of infant mortality in seven South Asian countries

When two or more observations take on the same value, we stack the dots in a dotplot on top of each other. This gives an effect similar to the effect of the bars in a bar chart. If we create bars in place of the stacks of dots, we then have a **histogram**. A histogram is like a bar chart, except that a histogram displays a numeric variable, which means that there is a natural order and scale for the variable. In a bar chart the amount of space between the bars (if any) is arbitrary, since the data being displayed are categorical. In a histogram the scale of the variable determines the placement of the bars. The following example shows a dotplot and a histogram for a frequency distribution.

Example 2.2.4

Litter Size of Sows A group of thirty-six 2-year-old sows of the same breed ($\frac{3}{4}$ Duroc, $\frac{1}{4}$ Yorkshire) were bred to Yorkshire boars. The number of piglets surviving to 21 days of age was recorded for each sow.[4] The results are given in Table 2.2.4 and displayed as a dotplot in Figure 2.2.4 and as a histogram in Figure 2.2.5. ■

Table 2.2.4 Number of surviving piglets of 36 sows

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

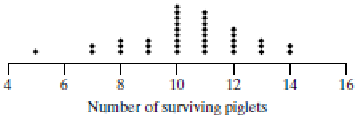


Figure 2.2.4 Dotplot of number of surviving piglets of 36 sows

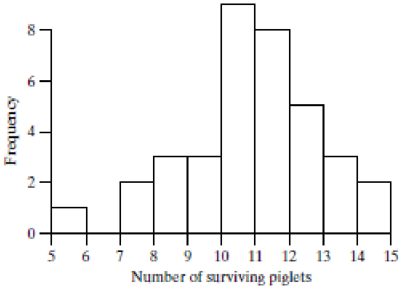


Figure 2.2.5 Histogram of number of surviving piglets of 36 sows

RELATIVE FREQUENCY

The frequency scale is often replaced by a **relative frequency** scale:

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

The relative frequency scale is useful if several data sets of different sizes (n's) are to be displayed together for comparison. As another option, a relative frequency can be expressed as a percentage frequency. The shape of the display is not affected by the choice of frequency scale, as the following example shows.

Example 2.2.5

Color of Poinsettias The poinsettia color distribution of Example 2.2.1 is expressed as frequency, relative frequency, and percent frequency in Table 2.2.5 and Figure 2.2.6. ■

Table 2.2.5 Color of 182 poinsettias

Color	Frequency	Relative frequency	Percent frequency
Red	108	.59	59
Pink	34	.19	19
White	40	.22	22
Total	182	1.00	100

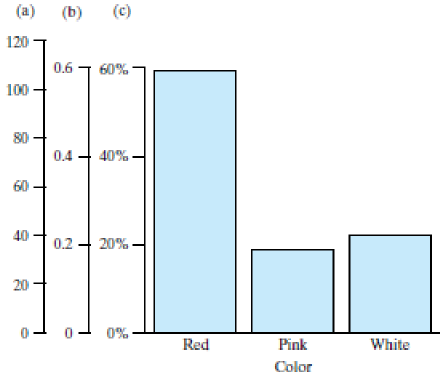


Figure 2.2.6 Bar chart of poinsettia colors on three scales:
(a) Frequency
(b) Relative frequency
(c) Percent frequency

GROUPED FREQUENCY DISTRIBUTIONS

In the preceding examples, simple ungrouped frequency distributions provided concise summaries of the data. For many data sets, it is necessary to group the data in order to condense the information adequately. (This is usually the case with continuous variables.) The following example shows a grouped frequency distribution.

Example 2.2.6

Serum CK Creatine phosphokinase (CK) is an enzyme related to muscle and brain function. As part of a study to determine the natural variation in CK concentration, blood was drawn from 36 male volunteers. Their serum concentrations of CK (measured in U/l) are given in Table 2.2.6.[5] Table 2.2.7 shows these data grouped into **classes**. For instance, the frequency of the class [20,40) (all values in the interval $20 \le y < 40$) is 1, which means that one CK value fell in this range. The grouped frequency distribution is displayed as a histogram in Figure 2.2.7. ■

Table 2.2.6 Serum CK values for 36 men

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Table 2.2.7 Frequency distribution of serum CK values for 36 men

Serum CK (U/l)	Frequency (number of men)
[20,40)	1
[40,60)	4
[60,80)	7
[80,100)	8
[100,120)	8
[120,140)	3
[140,160)	2
[160,180)	1
[180,200)	0
[200,220)	2
Total	36

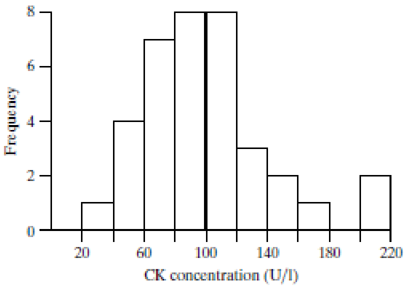


Figure 2.2.7 Histogram of serum CK concentrations for 36 men

A grouped frequency distribution should display the essential features of the data. For instance, the histogram of Figure 2.2.7 shows that the average CK value is about 100 U/l, with the majority of the values falling between 60 and 140 U/l. In addition, the histogram shows the *shape* of the distribution. Note that the CK values are piled up around a central peak, or **mode**. On either side of this mode, the frequencies decline and ultimately form the **tails** of the distribution. These shape features are labeled in Figure 2.2.8. The CK distribution is not symmetric but is a bit **skewed to the right**, which means that the right tail is more stretched out than the left.*

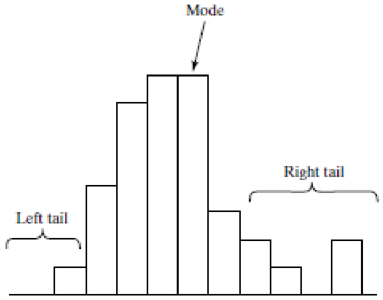


Figure 2.2.8 Shape features of the CK distribution

*To help remember which tail of a skewed distribution is the longer tail, think of skew as stretch. Which side of the distribution is more stretched away from the center? A distribution that is skewed to the right is one in which the right tail stretches out more than the left.

When making a histogram, we need to decide how many classes to have and how wide the classes should be. If we use computer software to generate a histogram, the program will choose the number of classes and the class width for us, but most software allows the user to change the number of classes and to specify the class width. If a data set is large and is quite spread out, it is a good idea to look at more than one histogram of the data, as is done in Example 2.2.7.

Example 2.2.7

Heights of Students A sample of 510 college students were asked how tall they were. Note that they were not measured; rather, they just reported their heights.[6] Figure 2.2.9 shows the distribution of the self-reported values, using 7 classes and a class width of 3 (inches). By using only 7 classes, the distribution appears to be reasonably symmetric, with a single peak around 66 inches.

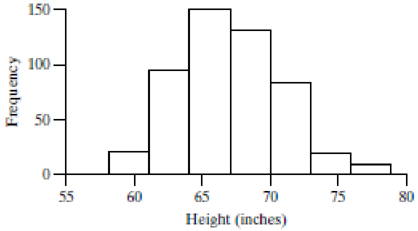


Figure 2.2.9 Heights of students, using 7 classes (class width = 3)

Figure 2.2.10 shows the height data, but in a histogram that uses 18 classes and a class width of 1.1. This view of the data shows two modes—one for women and one for men.

Figure 2.2.11 shows the height data again, this time using 37 classes, each of width 0.5. Using such a large number of classes makes the distribution look jagged. In this case, we see an alternating pattern between classes with lots of observations and classes with few observations. In the middle of the distribution we see that there were many students who reported a height of 63 inches, few who reported a height of 63.5 inches, many who reported a height of 64 inches, and so on. It seems that most students round off to the nearest inch! ■

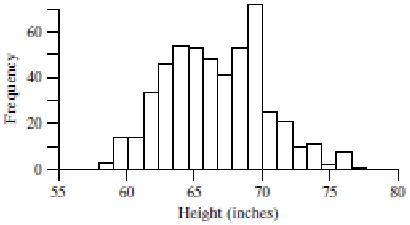


Figure 2.2.10 Heights of students, using 18 classes (class width = 1.1)

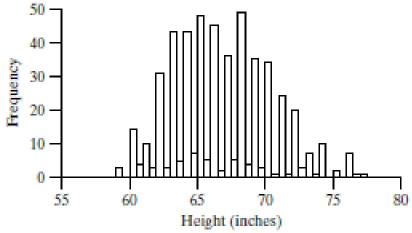


Figure 2.2.11 Heights of students, using 37 classes (class width = 0.5)

INTERPRETING AREAS IN A HISTOGRAM

A histogram can be looked at in two ways. The tops of the bars sketch out the shape of the distribution. But the *areas* within the bars also have a meaning. The area of each bar is proportional to the corresponding frequency. Consequently, the area of one or several bars can be interpreted as expressing the number of observations in the classes represented by the bars. For example, Figure 2.2.12 shows a histogram of the CK distribution of Example 2.2.6. The shaded area is 42% of the total area in all the bars. Accordingly, 42% of the CK values are in the corresponding classes; that is, 15 of 36 or 42% of the values are between 60 U/l and 100 U/l.*

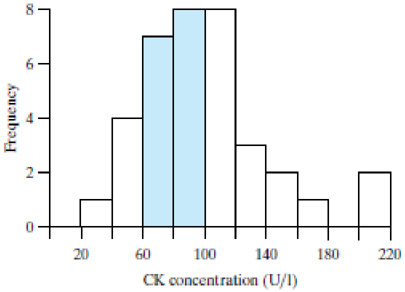


Figure 2.2.12 Histogram of CK distribution. The shaded area is 42% of the total area and represents 42% of the observations.

The area interpretation of histograms is a simple but important idea. In our later work with distributions we will find the idea to be indispensable.

SHAPES OF DISTRIBUTIONS

When discussing a set of data, we want to describe the shape, center, and spread of the distribution. In this section we concentrate on the shapes of frequency distributions and illustrate some of the diversity of distributions encountered in the life sciences. The shape of a distribution can be indicated by a smooth curve that approximates the histogram, as shown in Figure 2.2.13.

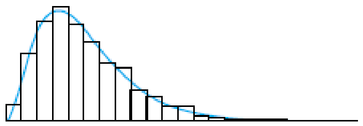


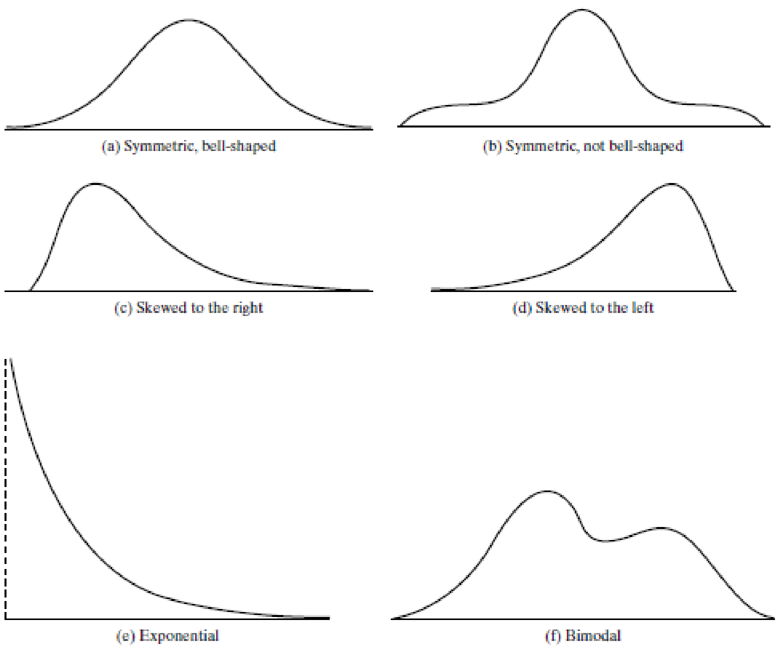
Figure 2.2.13 Approximation of a histogram by a smooth curve

Some distributional shapes are shown in Figure 2.2.14. A common shape for biological data is **unimodal** (has one mode) and is somewhat skewed to the right, as in (c). Approximately bell-shaped distributions, as in (a), also occur. Sometimes a distribution is symmetric but differs from a bell in having long tails; an exaggerated version is shown in (b). Left-skewed (d) and exponential (e) shapes are less common. **Bimodality** (two modes), as in (f), can indicate the existence of two distinct subgroups of observational units.

Notice that the shape characteristics we are emphasizing, such as number of modes and degree of symmetry, are *scale free*; that is, they are not affected by the arbitrary choices of vertical and horizontal scale in plotting the distribution. By contrast, a characteristic such as whether the distribution appears short and fat, or tall and skinny, is affected by how the distribution is plotted and so is not an inherent feature of the biological variable.

*Strictly speaking, between 60 U/l and 99 U/l, inclusive.

The following three examples illustrate biological frequency distributions with various shapes. In the first example, the shape provides evidence that the distribution is in fact biological rather than nonbiological.


(a) Symmetric, bell-shaped

(b) Symmetric, not bell-shaped

(c) Skewed to the right

(d) Skewed to the left

(e) Exponential

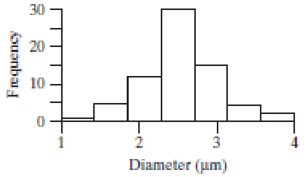
(f) Bimodal

Figure 2.2.14 Shapes of distributions

Example 2.2.8

Microfossils In 1977, paleontologists discovered microscopic fossil structures, resembling algae, in rocks 3.5 billion years old. A central question was whether these structures were biological in origin. One line of argument focused on their size distribution, which is shown in Figure 2.2.15. This distribution, with its unimodal and rather symmetric shape, resembles that of known microbial populations, but not that of known nonbiological structures.[7]

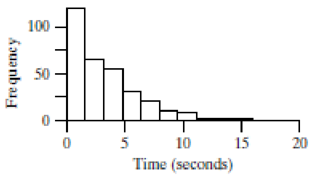
Figure 2.2.15 Sizes of microfossils



Example 2.2.9

Cell Firing Times A neurobiologist observed discharges from rat muscle cells grown in culture together with nerve cells. The time intervals between 308 successive discharges were distributed as shown in Figure 2.2.16. Note the exponential shape of the distribution.[8]

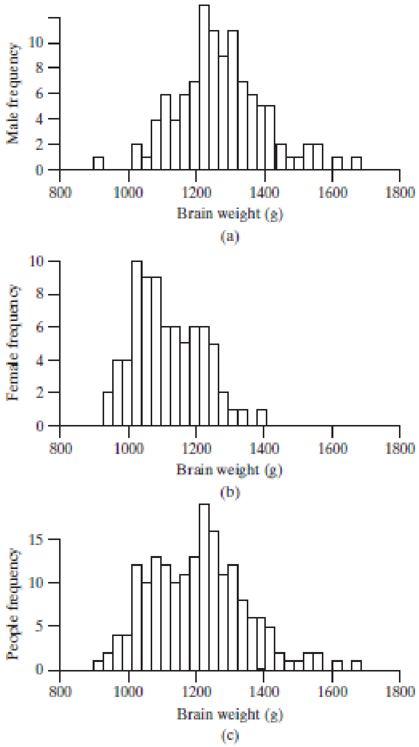
Figure 2.2.16 Time intervals between electrical discharges in rat muscle cells



Example 2.2.10

Brain Weight In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The data for males and females are shown in Figure 2.2.17(a) and (b). The male distribution is fairly symmetric and bell shaped; the female distribution is somewhat skewed to the right. Part (c) of the figure shows the brain weight distribution for males and females combined. This combined distribution is slightly bimodal.[9]

Figure 2.2.17 Brain weights



SOURCES OF VARIATION

In interpreting biological data, it is helpful to be aware of sources of variability. The variation among observations in a data set often reflects the combined effects of several underlying factors. The following two examples illustrate such situations.

Example 2.2.11

Weights of Seeds In a classic experiment to distinguish environmental from genetic influence, a geneticist weighed seeds of the princess bean *Phaseolus vulgaris*. Figure 2.2.18 shows the weight distributions of (a) 5,494 seeds from a commercial seed lot, and (b) 712 seeds from a highly inbred line that was derived from a single seed from the original lot. The variability in (a) is due to both environmental and genetic factors; in (b), because the plants are nearly genetically identical, the variation in weights is due largely to environmental influence.[10] Thus, there is less variability in the inbred line. ■

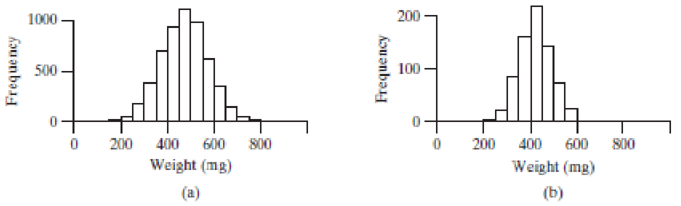


Figure 2.2.18 Weights of princess bean seeds: (a) from an open-bred population; (b) from an inbred line

Example 2.2.12

Serum ALT Alanine aminotransferase (ALT) is an enzyme found in most human tissues. Part (a) of Figure 2.2.19 shows the serum ALT concentrations for 129 adult volunteers. The following are potential sources of variability among the measurements:

1. Interindividual
   (a) Genetic
   (b) Environmental
2. Intraindividual
   (a) Biological: changes over time
   (b) Analytical: imprecision in assay

The effect of the last source—analytical variation—can be seen in part (b) of Figure 2.2.19, which shows the frequency distribution of 109 assays of the *same* specimen of serum; the figure shows that the ALT assay is fairly imprecise.[11] ■

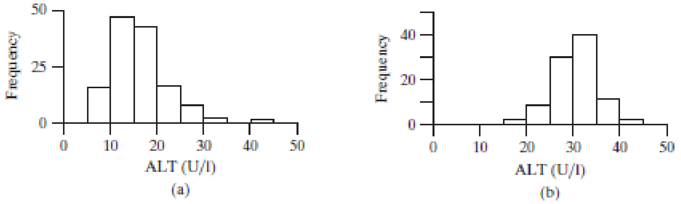


Figure 2.2.19 Distribution of serum ALT measurements (a) for 129 volunteers; (b) for 109 assays of the same specimen

Exercises 2.2.1–2.2.9

2.2.1 A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*. The results were as follows:[12]

6.1	5.7	6.0	6.5	6.0	5.7
6.1	5.8	5.9	6.1	6.2	6.0
6.3	6.2	6.1	6.2	6.0	5.7
6.2	5.8	5.7	6.3	6.2	5.7
6.2	6.1	5.9	6.5	5.4	6.7
5.9	6.1	5.9	5.9	6.1	6.1

(a) Construct a frequency distribution and display it as a table and as a histogram.

(b) Describe the shape of the distribution.

2.2.2 In a study of schizophrenia, researchers measured the activity of the enzyme monoamine oxidase (MAO) in the blood platelets of 18 patients. The results (expressed as nmoles benzylaldehyde product per 108 platelets) were as follows:[13]

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

Construct a dotplot of the data.

2.2.3 Consider the data presented in Exercise 2.2.2. Construct a frequency distribution and display it as a table and as a histogram.

2.2.4 A dendritic tree is a branched structure that emanates from the body of a nerve cell. As part of a study of brain development, 36 nerve cells were taken from the brains of newborn guinea pigs. The investigators counted the number of dendritic branch segments emanating from each nerve cell. The numbers were as follows:[14]

23	30	54	28	31	29	34	35	30
27	21	43	51	35	51	49	35	24
26	29	21	29	37	27	28	33	33
23	37	27	40	48	41	20	30	57

Construct a dotplot of the data.

2.2.5 Consider the data presented in Exercise 2.2.4. Construct a frequency distribution and display it as a table and as a histogram.

2.2.6 The total amount of protein produced by a dairy cow can be estimated from periodic testing of her milk. The following are the total annual protein production values (lb) for twenty-eight 2-year-old Holstein cows. Diet, milking procedures, and other conditions were the same for all the animals.[15]

425	481	477	434	410	397	438
545	528	496	502	529	500	465
539	408	513	496	477	445	546
471	495	445	565	499	508	426

Construct a frequency distribution and display it as a table and as a histogram.

2.2.7 For each of 31 healthy dogs, a veterinarian measured the glucose concentration in the anterior chamber of the right eye and also in the blood serum. The following data are the anterior chamber glucose measurements, expressed as a percentage of the blood glucose.[16]

81	85	93	93	99	76	75	84
78	84	81	82	89	81	96	82
74	70	84	86	80	70	131	75
88	102	115	89	82	79	106	

Construct a frequency distribution and display it as a table and as a histogram.

2.2.8 Agronomists measured the yield of a variety of hybrid corn in 16 locations in Illinois. The data, in bushels per acre, were[17]

241	230	207	219	266	167
204	144	178	158	153	
187	181	196	149	183	

(a) Construct a dotplot of the data.

(b) Describe the shape of the distribution.

2.2.9 (Computer problem) Trypanosomes are parasites that cause disease in humans and animals. In an early study of trypanosome morphology, researchers measured the lengths of 500 individual trypanosomes taken from the blood of a rat. The results are summarized in the accompanying frequency distribution.[18]

Length (μm)	Frequency (number of individuals)	Length (μm)	Frequency (number of individuals)
15	1	27	36
16	3	28	41
17	21	29	48
18	27	30	28
19	23	31	43
20	15	32	27
21	10	33	23
22	15	34	10
23	19	35	4
24	21	36	5
25	34	37	1
26	44	38	1

(a) Construct a histogram of the data using 24 classes (i.e., one class for each integer length, from 15 to 38).

(b) What feature of the histogram suggests the interpretation that the 500 individuals are a mixture of two distinct types?

(c) Construct a histogram of the data using only 6 classes. Discuss how this histogram gives a qualitatively different impression than the histogram from part (a).

2.3 Descriptive Statistics: Measures of Center

For categorical data, the frequency distribution provides a concise and complete summary of a sample. For numeric variables, the frequency distribution can usefully be supplemented by a few numerical measures. A numerical measure calculated from sample data is called a **statistic.*** **Descriptive statistics** are statistics that describe a set of data. Usually the descriptive statistics for a sample are calculated in order to provide information about a population of interest (see Section 2.8). In this section we discuss measures of the center of the data. There are several different ways to define the "center" or "typical value" of the observations in a sample. We will consider the two most widely used measures of center: the median and the mean.

THE MEDIAN

Perhaps the simplest measure of the center of a data set is the sample **median**. The sample median is the value that most nearly lies in the middle of the sample—it is the data value that splits the ordered data into two equal halves. To find the median, first arrange the observations in increasing order. In the array of ordered observations, the median is the middle value (if n is odd) or midway between the two middle values (if n is even). We denote the median of the sample by the symbol $\tilde{y}$ (read "y-tilde"). Example 2.3.1 illustrates these definitions.

Example 2.3.1

Weight Gain of Lambs The following are the 2-week weight gains (lb) of six young lambs of the same breed that had been raised on the same diet:[19]

11 13 19 2 10 1

The ordered observations are

1 2 10 11 13 19

The median weight gain is

$$\tilde{y} = \frac{10 + 11}{2} = 10.5 \text{ lb}$$

The median divides the sorted data into two equal pieces (the same number of observations fall above and below the median). Figure 2.3.1 shows a dotplot of the lamb weight-gain data, along with the location of $\tilde{y}$. ■

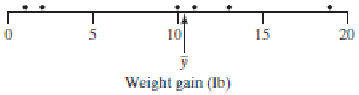


Figure 2.3.1 Plot of the lamb weight-gain data

*Numerical measures based on the entire population are called **parameters**, which are discussed in greater detail in Section 2.8.

Example 2.3.2

Weight Gain of Lambs Suppose the sample contained one more lamb, with the seven ranked observations as follows:

1 2 10 10 11 13 19

For this sample, the median weight gain is

$$\tilde{y} = 10 \text{ lb}$$

(Notice that in this example there are two lambs whose weight gain is equal to the median. The fourth observation—the second 10—is the median.) ■

A more formal way to define the median is in terms of rank position in the ordered array (counting the smallest observation as rank 1, the next as 2, and so on). The rank position of the median is equal to

$$(0.5)(n + 1)$$

Thus, if $n = 7$, we calculate $(0.5)(n + 1) = 4$, so that the median is the fourth largest observation; if $n = 6$, we have $(0.5)(n + 1) = 3.5$, so that the median is midway between the third and fourth largest observations. Note that the formula $(0.5)(n + 1)$ does not give the median, it gives the location of the median within the ordered list of the data.

THE MEAN

The most familiar measure of center is the ordinary average or **mean** (sometimes called the arithmetic mean). The mean of a sample (or "the sample mean") is the sum of the observations divided by the number of observations. If we denote a variable by Y, then we denote the observations in a sample by $y_1, y_2, \ldots, y_n$ and we denote the mean of the sample by the symbol $\bar{y}$ (read "y-bar"). Example 2.3.3 illustrates this notation.

Example 2.3.3

Weight Gain of Lambs The following are the data from Example 2.3.1:

11 13 19 2 10 1

Here $y_1 = 11, y_2 = 13$, and so on, and $y_6 = 1$. The sum of the observations is $11 + 13 + \cdots + 1 = 56$. We can write this using "summation notation" as $\sum_{i=1}^{n} y_i = 56$. The symbol $\sum_{i=1}^{n} y_i$ means to "add up the y_i's." Thus, when $n = 6$, $\sum_{i=1}^{n} y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6$. In this case we get $\sum_{i=1}^{n} y_i = 11 + 13 + 19 + 2 + 10 + 1 = 56$.

The mean weight gain of the six lambs in this sample is

$$\begin{aligned}\bar{y} &= \frac{11 + 13 + 19 + 2 + 10 + 1}{6} \\ &= \frac{56}{6} \\ &= 9.33 \text{ lb}\end{aligned}$$

THE SAMPLE MEAN The general definition of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where the y_i's are the observations in the sample and n is the sample size (that is, the number of y_i's).

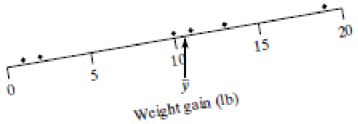


Figure 2.3.2 Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance

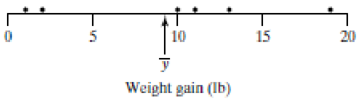


Figure 2.3.3 Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance

While the median divides the data into two equal pieces (i.e., the same number of observations above and below), the mean is the "point of balance" of the data. Figure 2.3.2 shows a dotplot of the lamb weight-gain data, along with the location of $\tilde{y}$. If the data points were children on a weightless seesaw, then the seesaw would tip if the fulcrum were placed at $\tilde{y}$ despite there being the same number of children on either side. The children on the left side (below $\tilde{y}$) tend to sit further from $\tilde{y}$ than the children on the right (above $\tilde{y}$) causing the seesaw to tip. However, if the fulcrum were placed at $\bar{y}$, the seesaw would exactly balance as in Figure 2.3.3. ■

The difference between a data point and the mean is called a **deviation**: deviation$_i = y_i - \bar{y}$. The mean has the property that the sum of the deviations from the mean is zero—that is, $\Sigma_{i=1}^{n}(y_i - \bar{y}) = 0$. In this sense, the mean is a center of the distribution—the positive deviations balance the negative deviations.

Example 2.3.4

Weight Gain of Lambs For the lamb weight-gain data, the deviations are as follows:

$$\text{deviation}_1 = y_1 - \bar{y} = 11 - 9.33 = 1.67$$
$$\text{deviation}_2 = y_2 - \bar{y} = 13 - 9.33 = 3.67$$
$$\text{deviation}_3 = y_3 - \bar{y} = 19 - 9.33 = 9.67$$
$$\text{deviation}_4 = y_4 - \bar{y} = 2 - 9.33 = -7.33$$
$$\text{deviation}_5 = y_5 - \bar{y} = 10 - 9.33 = 0.67$$
$$\text{deviation}_6 = y_6 - \bar{y} = 1 - 9.33 = -8.33$$

The sum of the deviations is $\Sigma_{i=1}^{n}(y_i - \bar{y}) = 1.67 + 3.67 + 9.67 - 7.33 + 0.67 - 8.33 = 0$. ■

Robustness A statistic is said to be **robust** if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes are dramatic ones. The median is a robust statistic, but the mean is not robust because it can be greatly shifted by changes in even one observation. Example 2.3.5 illustrates this behavior.

Example 2.3.5

Weight Gain of Lambs Recall that for the lamb weight-gain data

$$1 \quad 2 \quad 10 \quad 11 \quad 13 \quad 19$$

we found

$$\bar{y} = 9.33 \text{ and } \tilde{y} = 10.5$$

Suppose now that the observation 19 is changed. How would the mean and median be affected? You can visualize the effect by imagining moving the right-hand dot in Figure 2.3.3. Clearly the mean could change a great deal; the median would not be affected. For instance,

If the 19 is changed to 14, the mean becomes 8.5 and the median does not change.

If the 19 is changed to 29, the mean becomes 11 and the median does not change.

These changes are not wild ones; that is, the changed samples might well have arisen from the same feeding experiment. Of course, a huge change, such as changing the 19 to 100, would shift the mean very drastically. Note that it would not shift the median at all. ■

VISUALIZING THE MEAN AND MEDIAN

We can visualize the mean and the median in relation to the histogram of a distribution. The median divides the area under the histogram roughly in half because it divides the observations roughly in half ["roughly" because some observations may be tied at the median, as in Example 2.3.3(b), and because the observations within each class are not uniformly distributed across the class]. The mean can be visualized as the point of balance of the histogram: If the histogram were made out of plywood, it would balance if supported at the mean.

If the frequency distribution is symmetric, the mean and the median are equal and fall in the center of the distribution. If the frequency distribution is skewed, both measures are pulled toward the longer tail, but the mean is usually pulled farther than the median. The effect of skewness is illustrated by the following example.

Example 2.3.6

Cricket Singing Times Male Mormon crickets *(Anabrus simplex)* sing to attract mates. A field researcher measured the duration of 51 unsuccessful songs—that is, the time until the singing male gave up and left his perch.[20] Figure 2.3.4 shows the histogram of the 51 singing times. Table 2.3.1 gives the raw data. The median is 3.7 min and the mean is 4.3 min. The discrepancy between these measures is due largely to the long straggly tail of the distribution; the few unusually long singing times influence the mean, but not the median. ■

Table 2.3.1 Fifty-one cricket singing times (min)

4.3	3.9	17.4	2.3	0.8	1.5	0.7	3.7
24.1	9.4	5.6	3.7	5.2	3.9	4.2	3.5
6.6	6.2	2.0	0.8	2.0	3.7	4.7	
7.3	1.6	3.8	0.5	0.7	4.5	2.2	
4.0	6.5	1.2	4.5	1.7	1.8	1.4	
2.6	0.2	0.7	11.5	5.0	1.2	14.1	
4.0	2.7	1.6	3.5	2.8	0.7	8.6	

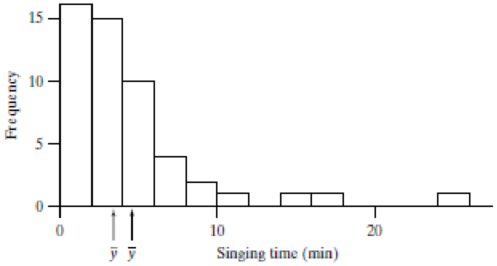


Figure 2.3.4 Histogram of cricket singing times

MEAN VERSUS MEDIAN

Both the mean and the median are usually reasonable measures of the center of a data set. The mean is related to the sum; for example, if the mean weight gain of 100 lambs is 9 lb, then the total weight gain is 900 lb, and this total may be of primary interest since it translates more or less directly into profit for the farmer. In some situations the mean makes very little sense. Suppose, for example, that the observations are survival times of cancer patients on a certain treatment protocol, and that most patients survive less than 1 year, while a few respond well and survive for 5 or even 10 years. In this case, the mean survival time might be greater than the survival time of most patients; the median would more nearly represent the experience of a "typical" patient. Note also that the mean survival time cannot be computed until the last patient has died; the median does not share this disadvantage. Situations in which the median can readily be computed, but the mean cannot, are not uncommon in bioassay, survival, and toxicity studies.

We have noted that the median is more robust than the mean. If a data set contains a few observations rather distant from the main body of the data—that is, a long, straggly tail—then the mean may be unduly influenced by these few unusual observations. Thus, the "tail" may "wag the dog"—an undesirable situation. In such cases, the robustness of the median may be advantageous.

An advantage of the mean is that in some circumstances it is more efficient than the median. Efficiency is a technical notion in statistical theory; roughly speaking, a method is efficient if it takes full advantage of all the information in the data. Partly because of its efficiency, the mean has played a major role in classical methods in statistics.

Exercises 2.3.1–2.3.14

2.3.1 Invent a sample of size 5 for which the sample mean is 20 and not all the observations are equal.

2.3.2 Invent a sample of size 5 for which the sample mean is 20 and the sample median is 15.

2.3.3 A researcher applied the carcinogenic (cancer-causing) compound benzo(a)pyrene to the skin of five mice, and measured the concentration in the liver tissue after 48 hours. The results (nmol/gm) were as follows:[21]

6.3 5.9 7.0 6.9 5.9

Determine the mean and the median.

2.3.4 Consider the data from Exercise 2.3.3. Do the calculated mean and median support the claim that, in general, liver tissue concentration after 48 hours differs from 6.3 nmol/gm?

2.3.5 Six men with high serum cholesterol participated in a study to evaluate the effects of diet on cholesterol level. At the beginning of the study their serum cholesterol levels (mg/dl) were as follows:[22]

366 327 274 292 274 230

Determine the mean and the median.

2.3.6 Consider the data from Exercise 2.3.5. Suppose an additional observation equal to 400 were added to the sample. What would be the mean and the median of the seven observations?

2.3.7 The weight gains of beef steers were measured over a 140-day test period. The average daily gains (lb/day) of 9 steers on the same diet were as follows:[23]

3.89 3.51 3.97 3.31 3.21
3.36 3.67 3.24 3.27

Determine the mean and median.

2.3.8 Consider the data from Exercise 2.3.7. Are the calculated mean and median consistent with the claim that, in general, steers gain 3.5 lb/day? Are they consistent with a claim of 4.0 lb/day?

2.3.9 Consider the data from Exercise 2.3.7. Suppose an additional observation equal to 2.46 were added to the sample. What would be the mean and the median of the 10 observations?

2.3.10 As part of a classic experiment on mutations, 10 aliquots of identical size were taken from the same culture of the bacterium *E. coli*. For each aliquot, the number of bacteria resistant to a certain virus was determined. The results were as follows:[24]

14 15 13 21 15
14 26 16 20 13

(a) Construct a frequency distribution of these data and display it as a histogram.

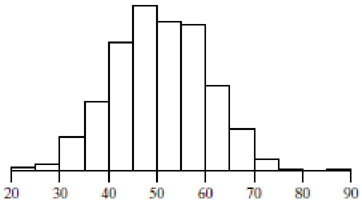
(b) Determine the mean and the median of the data and mark their locations on the histogram.

2.3.11 The accompanying table gives the litter size (number of piglets surviving to 21 days) for each of 36 sows (as in Example 2.2.4). Determine the median litter size. (*Hint*: Note that there is one 5, but there are two 7's, three 8's, etc.)

Number of piglets	Frequency (Number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

2.3.12 Consider the data from Exercise 2.3.11. Determine the mean of the 36 observations. (*Hint*: Note that there is one 5 but there are two 7's, three 8's, etc. Thus, $\Sigma y_i = 5 + 7 + 7 + 8 + 8 + 8 + \cdots = 5 + 2(7) + 3(8) + \cdots$)

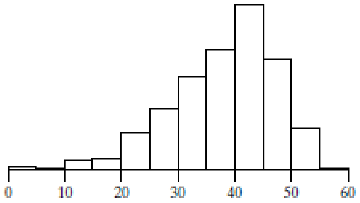
2.3.13 Here is a histogram.



(a) Estimate the median of the distribution.

(b) Estimate the mean of the distribution.

2.3.14 Here is a histogram.



(a) Estimate the median of the distribution.

(b) Estimate the mean of the distribution.

2.4 Boxplots

One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions, is known as a boxplot, which is the topic of this section. Before discussing boxplots, however, we need to discuss quartiles.

QUARTILES AND THE INTERQUARTILE RANGE

The median of a distribution splits the distribution into two parts, a lower part and an upper part. The **quartiles** of a distribution divide each of these parts in half, thereby dividing the distribution into four quarters. The **first quartile**, denoted by Q_1, is the median of the data values in the lower half of the data set. The **third quartile**, denoted by Q_3, is the median of the data values in the upper half of the data set.* The following example illustrates these definitions.

*Some authors use other definitions of quartiles, as does some computer software. A common alternative definition is to say that the first quartile has rank position $(0.25)(n + 1)$ and that the third quartile has rank position $(0.75)(n + 1)$. Thus, if $n - 10$, the first quartile would have rank position $(0.25)(11) - 2.75$—that is, to find the first quartile we would have to interpolate between the second and third largest observations. If n is large, then there is little practical difference between the definitions that various authors use.

Example 2.4.1

Blood Pressure The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:[25]

151 124 132 170 146 124 113

Putting these values in rank order, the sample is

113 124 124 132 146 151 170

The median is the fourth largest observation, which is 132. There are three data points in the lower part of the distribution: 113, 124, and 124. The median of these three values is 124. Thus, the first quartile, Q_1, is 124.

Likewise, there are three data points in the upper part of the distribution: 146, 151 and 170. The median of these three values is 151. Thus, the third quartile, Q_3, is 151.

$$\begin{array}{ccccccc} 113 & 124 & 124 & 132 & 146 & 151 & 170 \\ & \uparrow & & \vdots & & \uparrow & \\ & \text{first quartile} & & \text{median} & & \text{third quartile} & \\ & Q_1 & & & & Q_3 & \end{array}$$

Note that the median is not included in either the lower part or the upper part of the distribution. If the sample size, n, is even, then exactly one-half of the observations are in the lower part of the distribution and one-half are in the upper part.

The **interquartile range** is the difference between the first and third quartiles and is abbreviated as **IQR**: IQR $= Q_3 - Q_1$. For the blood pressure data in Example 2.4.1, the IQR is $151 - 124 = 27$. Note that the IQR is a *number*, not an interval; the IQR measures the spread of the middle 50% of the distribution.

Example 2.4.2

Pulse The pulses of 12 college students were measured.[26] Here are the data, arranged in order, with the position of the median indicated by a dashed line:

62 64 68 70 70 74 ⋮ 74 76 76 78 78 80

The median is $\frac{74 + 74}{2} = 74$. There are six observations in the lower part of the distribution: 62, 64, 68, 70, 70, 74. Thus, the first quartile is the average of the third and fourth largest data values:

$$Q_1 = \frac{68 + 70}{2} = 69$$

There are six observations in the upper part of the distribution: 74, 76, 76, 78, 78, 80. Thus, the third quartile is the average of the ninth and tenth largest data values (the third and fourth values in the upper part of the distribution):

$$Q_3 = \frac{76 + 78}{2} = 77$$

Thus, the interquartile range is

$$\text{IQR} = 77 - 69 = 8$$

We have

$$\begin{array}{ccccccccccccccc} 62 & 64 & 68 & & 70 & 70 & 74 & \vdots & 74 & 76 & 76 & & 78 & 78 & 80 \\ & & & \uparrow & & & & \text{median} & & & & \uparrow & & & \\ & & & \text{first quartile} & & & & & & & & \text{third quartile} & & & \\ & & & Q_1 & & & & & & & & Q_3 & & & \end{array}$$

The minimum pulse value is 62 and the maximum is 80.

The minimum, the maximum, the median, and the quartiles, taken together, are referred to as the **five-number summary** of the data.

OUTLIERS

Sometimes a data point differs so much from the rest of the data that it doesn't seem to belong with the other data. Such a point is called an **outlier**. An outlier might occur because of a recording error or typographical error when the data are recorded, because of an equipment failure during an experiment, or for many other reasons. Outliers are the most interesting points in a data set. Sometimes outliers tell us about a problem with the experimental protocol (e.g., an equipment failure, a failure of a patient to take his or her medication consistently during a medical trial). At other times an outlier might alert us to the fact that a special circumstance has happened (e.g., an abnormally high or low value on a medical test could indicate the presence of a disease in a patient).

People often use the term "outlier" informally. There is, however, a common definition of "outlier" in statistical practice. To give a definition of outlier, we first discuss what are known as fences. The **lower fence** of a distribution is

$$\text{lower fence} = Q_1 - 1.5 \times \text{IQR}$$

The **upper fence** of a distribution is

$$\text{upper fence} = Q_3 + 1.5 \times \text{IQR}$$

Note that the fences need not be data values; indeed, there might be no data near the fences. The fences just locate limits within the sample distribution. These limits give us a way to define outliers. *An outlier is a data point that falls outside of the fences.* That is, if

$$\text{data point} < Q_1 - 1.5 \times \text{IQR}$$

or

$$\text{data point} > Q_3 + 1.5 \times \text{IQR}$$

then we call the point an outlier.

Example 2.4.3

Pulse In Example 2.4.2 we saw that $Q_1 = 69$, $Q_3 = 77$, and IQR $= 8$. Thus, the lower fence is $69 - 1.5 \times 8 = 69 - 12 = 57$. Any point less than 57 would be an outlier. The upper fence is $77 + 1.5 \times 8 = 77 + 12 = 89$. Any point greater than

89 would be an outlier. Since there are no points less than 57 or greater than 89, there are no outliers in this data set.

Example 2.4.4 **Radish Growth in Light** A common biology experiment involves growing radish seedlings under various conditions. In one experiment students grew 14 radish seedlings in constant light. The observations, in order, are

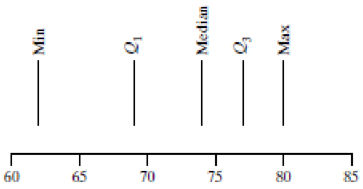
3 5 5 7 7 8 9 ⋮ 10 10 10 10 14 20 21

↑ first quartile Q_1 — median — ↑ third quartile Q_3

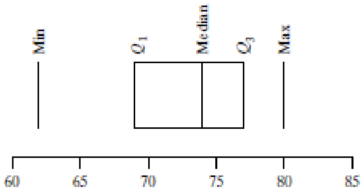
Thus, the median is $\frac{9 + 10}{2} = 9.5$, Q_1 is 7, and Q_3 is 10. The interquartile range is $\text{IQR} = 10 - 7 = 3$. The lower fence is $7 - 1.5 \times 3 = 7 - 4.5 = 2.5$, so any point less than 2.5 would be an outlier. The upper fence is $10 + 1.5 \times 3 = 10 + 4.5 = 14.5$, so any point greater than 14.5 is an outlier. Thus, the two largest observations in this data set are outliers: 20 and 21.

BOXPLOTS FOR DATA WITH NO OUTLIERS

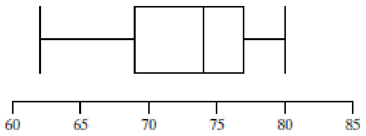
A **boxplot** is a visual representation of the five-number summary. To make a boxplot for a data set with no outliers, we first make a number line; then we mark the positions minimum, Q_1, the median, Q_3, and the maximum:



Next, we make a box connecting the quartiles:



Note that the interquartile range is equal to the length of the box. Finally, provided there are no outliers* we extend "whiskers" from Q_1 down to the minimum and from Q_3 up to the maximum:



A boxplot gives a quick visual summary of the distribution. We can immediately see where the center of the data is from the line within the box that locates the median. We see the spread of the total distribution, from the minimum up to the maximum, as well as the spread of the middle half of the distribution—the interquartile range—from the length of the box. The boxplot also gives an indication of the shape of the distribution; the preceding boxplot has a long lower whisker, indicating that the distribution is skewed to the left. Example 2.4.5 shows a boxplot for data from a radish growth experiment that had no outliers.[†]

Example 2.4.5 **Radish Growth** In another version of the experiment in Example 2.4.4, a moist paper towel is put into a plastic bag. About one third of the way from the bottom of the bag a seam of staples was created; the radish seeds were placed along the seam. One group of students kept their radish seed bags in total darkness for 3 days and then measured the length, in mm, of each radish shoot at the end of the 3 days. They collected 14 observations; the data are shown in Table 2.4.1.[27]

Table 2.4.1 Radish growth, in mm, after three days in total darkness

15	20	11	30	33
20	29	35	8	10
22	37	15	25	

Here are the data in order from smallest to largest:

8 10 11 **15** 15 20 **20** ⋮ **22** 25 29 **30** 33 35 37

↑ first quartile Q_1 — median — ↑ third quartile Q_3

The quartiles are $Q_1 = 15$ and $Q_3 = 30$. The median, $\tilde{y} = 21$, is the average of the two middle values of 20 and 22. Figure 2.4.1 shows a boxplot of the same data.

*We will consider situations with outliers after the next example.

[†] This and subsequent boxplots in our text are slightly stylized. Different computer packages present the plot somewhat differently, but all boxplots have the same basic five-number summary.

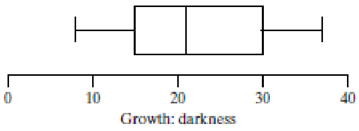


Figure 2.4.1 Boxplot of data on radish growth in darkness

BOXPLOTS FOR DATA WITH OUTLIERS

If there are outliers in the upper part of the distribution, then we can identify them with dots (or other plotting symbols) on the boxplot. We then extend a whisker from Q_3 up to the largest data point that is *not* an outlier. Likewise, if there are outliers in the lower part of the distribution, we identify them with dots and extend a whisker from Q_1 down to the smallest observation that is not an outlier. Figure 2.4.2 shows the distribution of radish seedlings grown under constant light. The area between the lower and upper fences is white, while the outlying region is blue.

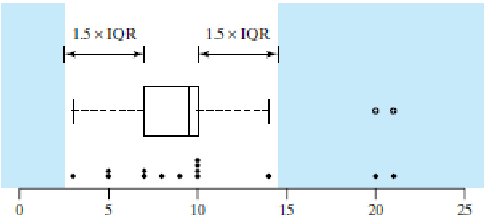


Figure 2.4.2 Dotplot and boxplot of data on radish growth in constant light. The points in the blue region are outliers.

Figure 2.4.3 shows a boxplot of the data on radish seedlings grown in constant light.*

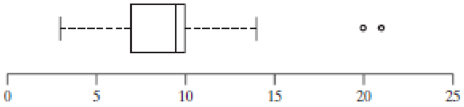


Figure 2.4.3 Boxplot of data on radish growth in constant light

The method we have defined for identifying outliers allows the bulk of the data to determine how extreme an observation must be before we consider it to be an outlier, since the quartiles and the IQR are determined from the data themselves. Thus, a point that is an outlier in one data set might not be an outlier in another data set. We label a point as an outlier if it is unusual relative to the inherent variability in the entire data set.

*Most computer software has options that can alter how outliers are determined and displayed.

After an outlier has been identified, people are often tempted to remove the outlier from the data set. In general this is not a good idea. If we can identify that an outlier occurred due to an equipment error, for example, then we have good reason to remove the outlier before analyzing the rest of the data. However, quite often outliers appear in data sets without any identifiable, external reason for them. In such cases, we simply proceed with our analysis, aware that there is an outlier present. In some cases, we might want to calculate the mean, for example, with and without the outlier and then report both calculations to show the effect of the outlier in the overall analysis. This is preferable to removing the outlier, which obscures the fact that there was an unusual data point present.

Exercises 2.4.1–2.4.8

2.4.1 Here are the data from Exercise 2.3.10 on the number of virus-resistant bacteria in each of 10 aliquots:

14	15	13	21	15
14	26	16	20	13

(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

(c) How large would an observation in this data set have to be in order to be an outlier?

2.4.2 Here are the 18 measurements of MAO activity reported in Exercise 2.2.2:

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

(c) How large would an observation in this data set have to be in order to be an outlier?

(d) Construct a boxplot of the data.

2.4.3 In a study of milk production in sheep (for use in making cheese), a researcher measured the 3-month milk yield for each of 11 ewes. The yields (liters) were as follows:[28]

56.5	89.8	110.1	65.6	63.7	82.6
75.1	91.5	102.9	44.4	108.1	

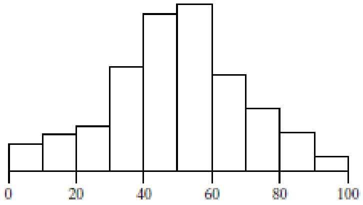
(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

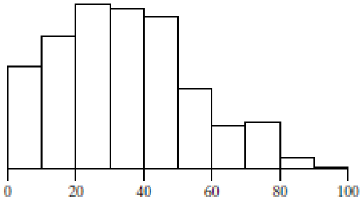
(c) Construct a boxplot of the data.

2.4.4 For each of the following histograms, use the histogram to estimate the median and the quartiles; then construct a boxplot for the distribution.

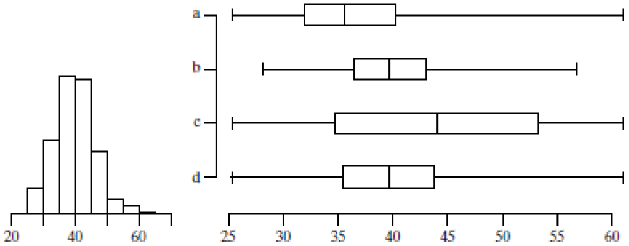
(a)



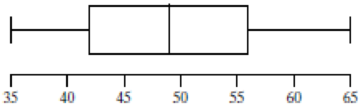
(b)



2.4.5 The following histogram shows the same data that are shown in one of the four boxplots. Which boxplot goes with the histogram? Explain your answer.



2.4.6 The following boxplot shows the five-number summary for a data set. For these data the minimum is 35, Q_1 is 42, the median is 49, Q_3 is 56, and the maximum is 65. Is it possible that no observation in the data set equals 42? Explain your answer.



2.4.7 Statistics software can be used to find the five-number summary of a data set. Here is an example of MINITAB's descriptive statistics summary for a variable stored in column 1 (C1) of MINITAB's worksheet.

```
Variable  N   Mean   Median  TrMean  StDev  SEMean
   C1    75  119.94  118.40  119.98   9.98    1.15

Variable  Min     Max      Q1      Q3
   C1    95.16  145.11  113.59  127.42
```

(a) Use the MINITAB output to calculate the interquartile range.

(b) Are there any outliers in this set of data?

2.4.8 Consider the data from Exercise 2.4.7. Use the five-number summary that is given to create a boxplot of the data.

2.5 Relationships between Variables

In the previous sections we have studied **univariate** summaries of both numeric and categorical variables. A univariate summary is a graphical or numeric summary of a single variable.

The histogram, boxplot, sample mean, and median are all examples of univariate summaries for numeric data. The bar chart, frequency, and relative frequency tables are examples of univariate summaries for categorical data. In this section we present some common **bivariate** graphical summaries used to examine the *relationship* between pairs of variables.

CATEGORICAL–CATEGORICAL RELATIONSHIPS

To understand the relationship between two categorical variables, we first summarize the data in a **bivariate frequency table**. Unlike the frequency table presented in Section 2.2 (a univariate table), the bivariate frequency table has both rows and columns—one dimension for each variable. The choice of which variable to list with the rows and which to list with the columns is arbitrary. The following example considers the relationship between two categorical variables: *E. Coli* Source and Sampling Location.

Example 2.5.1

***E. Coli* Watershed Contamination** In an effort to determine if there are differences in the primary sources of fecal contamination at different locations in the Morro Bay watershed, $n = 623$ water specimens were collected at three primary locations that feed into Morro Bay: Chorro Creek ($n_1 = 241$), Los Osos Creek ($n_2 = 256$), and Baywood Seeps ($n_3 = 126$).[29] DNA fingerprinting techniques were used to determine the intestinal origin of the dominant *E. coli* strain in each water specimen. *E. coli* origins were classified into the following five categories: bird, domestic pet (e.g., cat or dog), farm animal (e.g., horse, cow, pig), human, or other terrestrial mammal (e.g., fox, mouse, coyote . . .). Thus, each water specimen had *two* categorical variables measured: location (Chorro, Los Osos, or Baywood) and *E. coli* source (bird, . . . , terrestrial mammal). Table 2.5.1 presents a frequency table of the data. ■

Table 2.5.1 Frequency table of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	46	29	106	38	22	**241**
Los Osos Creek	79	56	32	63	26	**256**
Baywood Seeps	35	23	0	60	8	**126**
Total	**160**	**108**	**138**	**161**	**56**	**623**

While Table 2.5.1 provides a concise summary of the data, it is difficult to discover any patterns in the data. Examining relative frequencies (row or column proportions) often helps us make meaningful comparisons as seen in the following example.

Example 2.5.2

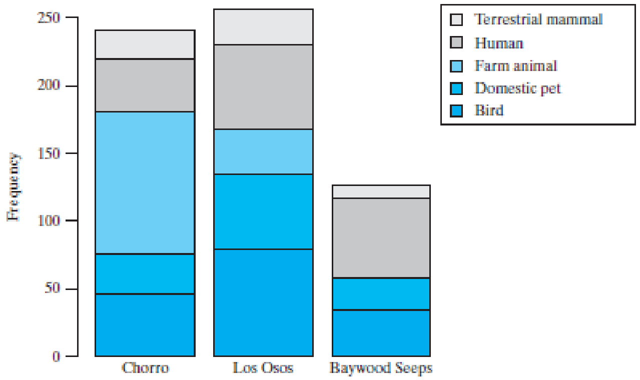
***E. Coli* Watershed Contamination** Are domestic pets more of an *E. coli* problem (i.e., source) at Chorro Creek or Baywood Seeps? Table 2.5.1 shows that the domestic pet *E. coli* source count at Chorro (29) is higher than Baywood (23), so at first glance it seems that pets are more problematic at Chorro. However, as more water specimens were collected at Chorro ($n_1 = 241$) than Baywood ($n_2 = 126$), the relative frequency of domestic pet source *E. coli* is actually lower at Chorro ($29/241 = 0.120$) than Baywood ($23/126 = 0.183$). Table 2.5.2 displays row percentages and thus facilitates comparisons of *E. coli* sources among the locations. (Note that column percentages would not be meaningful in this context since the water was sampled by location and not by *E. coli* source.) ■

Table 2.5.2 Bivariate relative frequency table (row percentages) of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	19.1	12.0	44.0	15.8	9.1	**100**
Los Osos Creek	30.9	21.9	12.5	24.6	10.2	**100**
Baywood Seeps	27.8	18.3	0.0	47.6	6.3	**100**
All locations	**25.7**	**17.3**	**22.2**	**25.8**	**9.0**	**100**

To visualize the data in Tables 2.5.1 and 2.5.2, we can examine **stacked bar charts**. With a stacked frequency bar chart, the overall height of each bar reflects the sample size for a level of the *X* categorical variable (e.g., location), while the height or thickness of a slice that makes up a bar represents the count of the *Y* categorical variable (e.g., *E. coli* source) for that level of *X*. Figure 2.5.1 displays a stacked bar chart for the *E. coli* watershed count data in Table 2.5.1.

Figure 2.5.1 Stacked frequency chart of *E. coli* source by location



Like the frequency table, the stacked frequency bar chart is not conducive to making comparisons across the three locations as the sample sizes differ for these locations. (This graph does help highlight the difference in sample sizes; for example, it is very clear that many fewer water specimens were collected at Baywood Seeps.) A chart that better displays the distribution of one categorical variable across levels of another is a **stacked relative frequency** (or percentage) bar chart, which graphs the summaries from a bivariate relative frequency table such as Table 2.5.2. Figure 2.5.2 provides an example using the *E. coli* watershed contamination data. This plot normalizes the bars of Figure 2.5.1 to have the same height (100%) to facilitate comparisons across the three locations.

Figure 2.5.2 Stacked relative frequency (percentage) chart of *E. coli* source by location

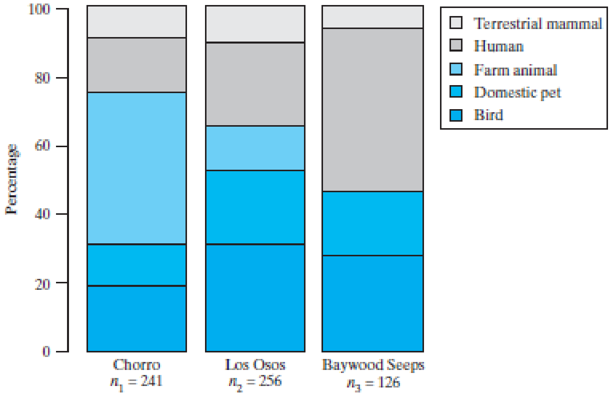


Figure 2.5.2 makes it very easy to see that farm animals are the largest contributors of *E. coli* to Chorro Creek while humans are primarily responsible for the pollution at Baywood Seeps. The distribution of the slices in the three bars appears quite different, suggesting that the distribution of *E. coli* sources is not the same at the three locations. In Chapter 10 we will learn how to determine if these apparent differences are large enough to be compelling evidence for real differences in the distribution of *E. coli* source by location, or whether they are likely due to chance variation.

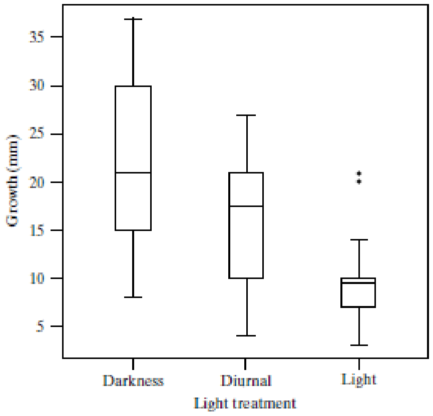
NUMERIC–CATEGORICAL RELATIONSHIPS

In Section 2.4 we learned that boxplots are graphs based on only five numbers: the minimum, first quartile, median, third quartile, and maximum. They are appealing plots because they are very simple and uncluttered, yet contain easy to read information about center, spread, skewness, and even outliers of a data set. By displaying **side-by-side boxplots** on the same graph, we are able to compare numeric data among several groups. We now consider an extension of the radish shoot growth problem in Example 2.4.3.

Example 2.5.3

Radish Growth Does light exposure alter initial radish shoot growth? The complete radish growth experiment of Examples 2.4.4 and 2.4.5 actually involved a total of 42 radish seeds randomly divided to receive one of three lighting conditions for germination (14 seeds in each lighting condition): 24-hour light, diurnal light (12 hours of light and 12 hours of darkness each day), and 24 hours of darkness. At the end of 3 days, shoot length was measured (mm). Thus, each shoot has two variables that are measured in this study: the categorical variable lighting condition (light, diurnal, dark) and the numeric variable sprout length (mm). Figure 2.5.3 displays side-by-side boxplots of the data. The boxplots make it very easy to compare the growth under the three conditions: It appears that light inhibits shoot growth. Are the observed differences in growth among the lighting conditions just due to chance variation, or is light really altering growth? We will learn how to numerically measure the strength of this evidence and answer this question in Chapters 7 and 11.

Figure 2.5.3 Side-by-side boxplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light



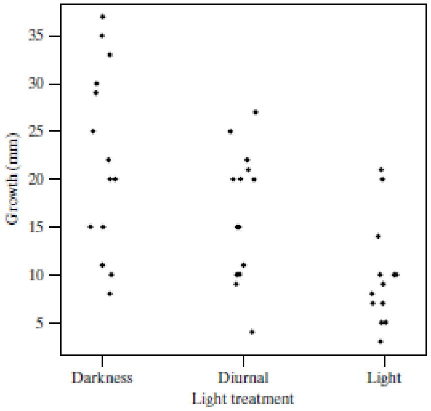


Figure 2.5.4 Side-by-side jittered dotplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light

For smaller data sets, we also may consider side-by-side dotplots of the data. Figure 2.5.4 displays a jittered side-by-side dotplot of the radish growth data of Example 2.5.3. The "jitter" is a common software option that adds horizontal scatter to the plot, helping to reduce the overlap of the dots. Choosing between side-by-side boxplots and dotplots is matter of personal preference. A good rule of thumb is to choose the plot that accurately reflects patterns in the data in the cleanest (least ink on the paper) way possible. For the radish growth example, the boxplot enables a very clean comparison of the growth under the three light treatments without hiding any information revealed by the dotplot.

NUMERIC–NUMERIC RELATIONSHIPS

Each of the previous examples considered comparing the distribution of one variable (either categorical or numeric) among several groups (i.e., across levels of a categorical variable). In the next example we illustrate the **scatterplot** as a tool to examine the relationship between two numeric variables, X and Y. A scatterplot plots each observed (x,y) pair as a dot on the x–y plane.

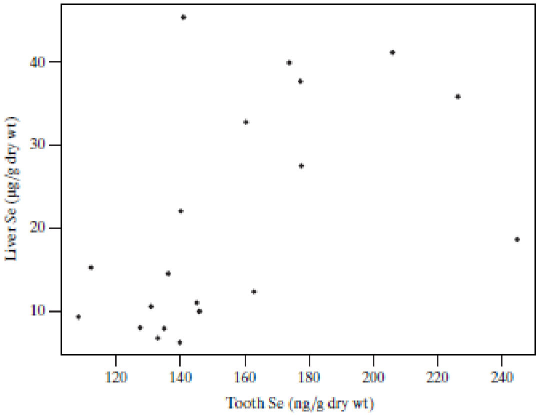
Example 2.5.4

Whale Selenium Can metal concentration in marine mammal teeth be used as a bioindicator for body burden? Selenium (Se) is an essential element that has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. Twenty beluga whales (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, as part of an annual traditional Inuit hunt.[30] Each whale yielded two numeric measurements: Tooth Se (μg/g) and Liver Se (ng/g). Selenium concentrations for the whales are listed in Table 2.5.3. Liver Se concentration (Y) is graphed against Tooth Se concentration (X) in the scatterplot of Figure 2.5.5. ■

Table 2.5.3 Liver and tooth selenium concentrations of 20 belugas

Whale	Liver Se (μg/g)	Tooth Se (ng/g)	Whale	Liver Se (μg/g)	Tooth Se (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

Figure 2.5.5 Scatterplot of liver selenium concentration against tooth selenium concentration for 20 belugas



Scatterplots are helpful in revealing relationships between numeric variables. In Figure 2.5.6 two lines have been added to the whale selenium scatterplot of Figure 2.5.5 to highlight the increasing trend in the data: Tooth Se concentration tends to increase with liver Se concentration. The dashed line is called a **lowess smooth**, whereas the straight solid line is called a **regression line**. Many software packages allow one to easily add these lines to a scatterplot. The lowess smooth is particularly helpful in visualizing curved or nonlinear relationships in data, while the regression line is used to highlight a linear trend. Generally speaking, we would choose only one of these to display on our graph. In this case, since the pattern is fairly linear (the lowess smooth is fairly straight), we would choose the solid regression line. In Chapter 12 we will learn how to identify the equation of the regression line that best summarizes the data and determine if the apparent trend in the data is likely to be just due to chance or if there is evidence for a real relationship between X and Y.

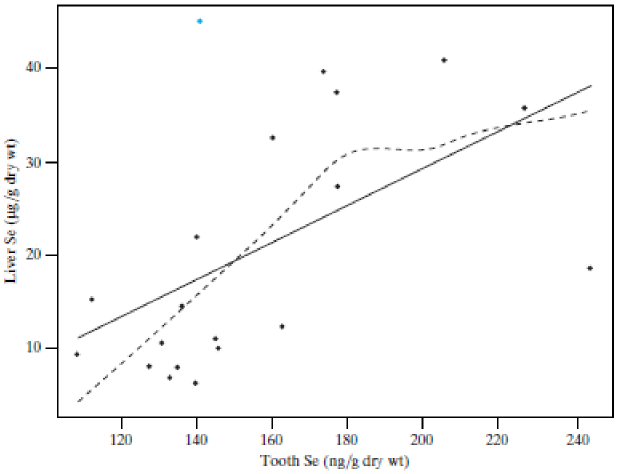


Figure 2.5.6 Scatterplot of liver selenium concentration against tooth selenium concentration for 20 belugas with regression (solid) and lowess (dashed) summary lines and outlier marked in blue

In addition to revealing relationships between two numeric variables, scatterplots also help reveal outliers that might otherwise be unnoticed in univariate plots (e.g., histograms, single boxplots). The colored point on Figure 2.5.6 falls far from the scatter of the other points. The X value of this point is not unusual in any way, and even the Y value, although large, doesn't appear extreme. The scatterplot, however, shows that the particular (x,y) pair for this whale is unusual.

Exercises 2.5.1–2.5.4

2.5.1 The two claws of the lobster (*Homarus americanus)* are identical in the juvenile stages. By adulthood, however, the two claws normally have differentiated into a stout claw called a "crusher" and a slender claw called a "cutter." In a study of the differentiation process, 26 juvenile animals were reared in smooth plastic trays and 18 were reared in trays containing oyster chips (which they could use to exercise their claws). Another 23 animals were reared in trays containing only one oyster chip. The claw configurations of all the animals as adults are summarized in the table.[31]

	Claw Configuration		
Treatment	Right crusher, left cutter	Right cutter, left crusher	Right and left cutter (no crusher)
Oyster chips	8	9	1
Smooth plastic	2	4	20
One oyster chip	7	9	7

(a) Create a stacked frequency bar chart to display these data.

(b) Create a stacked relative frequency bar chart to display these data.

(c) Of the two charts you created in parts (a) and (b), which is more useful for comparing the claw configurations across the three treatments? Why?

2.5.2 Does the length (mm) of the golden mantled ground squirrel (*Spermophilus lateralis*) differ by latitude in California? A graduate student captured squirrels at four locations across California. Listed from south to north the locations are Hemet, Big Bear, Susanville, and Loop Hill.[32]

Hemet	Big Bear	Susanville	Loop Hill
263	274	245	273
256	256	272	291
251	249	263	278
242	264	260	281
248		271	
281			

(a) Create side-by-side dotplots of the data. Consider the geography of these four locations when making your plot. Is alphabetic order of the locations the most appropriate, or is there a better way to order the location categories?

(b) Create side-by-side boxplots of the data. Again, consider the geography of these four locations when making your plot.

(c) Of the two plots created in parts (a) and (b), which do you prefer and why?

2.5.3 The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts to its varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made of the dark respiration rate. The accompanying table shows the altitude of origin (in meters) of each batch of buds and the dark respiration rate (expressed as μl of oxygen per hour per mg dry weight of tissue).[33]

Tree	Altitude of origin X (m)	Respiration rate Y (μl/hr · mg)
1	90	0.11
2	230	0.20
3	240	0.13
4	260	0.15
5	330	0.18
6	400	0.16
7	410	0.23
8	550	0.18
9	590	0.23
10	610	0.26
11	700	0.32
12	790	0.37

(a) Create a scatterplot of the data.

(b) If your software allows, add a regression line to summarize the trend.

(c) If your software allows, create a scatterplot with a lowess smooth to summarize the trend.

2.5.4 A group of college students were asked how many hours per week they exercise.[34]

The answers given by 12 men were as follows:

6 0 2 1 2 4.5 8 3 17 4.5 4 5

The answers given by 13 women were as follows:

5 13 3 2 6 14 3 1 1.5 1.5 3 8 4

(a) Construct parallel boxplots of the male and female distributions.

(b) Describe the two boxplots, including how they compare to each other.

2.6 Measures of Dispersion

We have considered the shapes and centers of distributions, but a good description of a distribution should also characterize how spread out the distribution is—are the observations in the sample all nearly equal, or do they differ substantially? In Section 2.4 we defined the interquartile range, which is one measure of dispersion. We will now consider other measures of dispersion: the range and the standard deviation.

THE RANGE

The sample **range** is the difference between the largest and smallest observations in a sample. Here is an example.

Example 2.6.1 **Blood Pressure** The systolic blood pressures (mm Hg) of seven middle-aged men were given in Example 2.4.1 as follows:

113 124 124 132 146 151 170

For these data, the sample range is

$$170 - 113 = 57 \text{ mm Hg}$$ ■

The range is easy to calculate, but it is very sensitive to extreme values; that is, it is not robust. If the maximum in the blood pressure sample had been 190 rather than 170, the range would have been changed from 57 to 77.

We defined the interquartile range (IQR) in Section 2.4 as the difference between the quartiles. Unlike the range, the IQR is robust. The IQR of the blood pressure data is $151 - 124 = 17$. If the maximum in the blood pressure sample had been 190 rather than 170, the IQR would not have changed; it would still be 17.

THE STANDARD DEVIATION

The standard deviation is the classical and most widely used measure of dispersion. Recall that a *deviation* is the difference between an observation and the sample mean:

$$\text{deviation} = \text{observation} - \bar{y}$$

The standard deviation of the sample, or sample **standard deviation**, is determined by combining the deviations in a special way, as described in the following box.

THE SAMPLE STANDARD DEVIATION The sample standard deviation is denoted by s and is defined by the following formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

In this formula, the expression $\Sigma_{i=1}^{n}(y_i - \bar{y})^2$ denotes the sum of the squared deviations.

So, to find the standard deviation of a sample, first find the deviations. Then

1. square
2. add
3. divide by $n - 1$
4. take the square root

To illustrate the use of the formula, we have chosen a data set that is especially simple to handle because the mean happens to be an integer.

Example 2.6.2

Chrysanthemum Growth In an experiment on chrysanthemums, a botanist measured the stem elongation (mm in 7 days) of five plants grown on the same greenhouse bench. The results were as follows:[35]

76 72 65 70 82

The data are tabulated in the first column of Table 2.6.1. The sample mean is

$$\bar{y} = \frac{365}{5} = 73 \text{ mm}$$

The deviations $(y_i - \bar{y})$ are tabulated in the second column of Table 2.6.1; the first observation is 3 mm above the mean, the second is 1 mm below the mean, and so on.

The third column of Table 2.6.1 shows that the sum of the squared deviations is

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 = 164$$

Table 2.6.1 Illustration of the formula for the sample standard deviation

Observation (y_i)	Deviation $(y_i - \bar{y})$	Squared deviation $(y_i - \bar{y})^2$
76	3	9
72	−1	1
65	−8	64
70	−3	9
82	9	81
Sum $365 = \sum_{i=1}^{n} y_i$	0	$164 = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Since $n = 5$, the standard deviation is

$$s = \sqrt{\frac{164}{4}}$$
$$= \sqrt{41}$$
$$= 6.4 \text{ mm}$$

Note that the units of s (mm) are the same as the units of Y. This is because we have squared the deviations and then later taken the square root. ■

The sample **variance**, denoted by s^2, is simply the standard deviation squared: variance $= s^2$. Thus, $s = \sqrt{\text{variance}}$.

Example 2.6.3

Chrysanthemum Growth The variance of the chrysanthemum growth data is

$$s^2 = 41 \text{ mm}^2$$

Note that the units of the variance (mm^2) are not the same as the units of Y. ■

An abbreviation We will frequently abbreviate "standard deviation" as "SD"; the symbol "s" will be used in formulas.

INTERPRETATION OF THE DEFINITION OF s

The magnitude (disregarding sign) of each deviation $(y_i - \bar{y})$ can be interpreted as the *distance* of the corresponding observation from the sample mean $\bar{y}$. Figure 2.6.1 shows a plot of the chrysanthemum growth data (Example 2.6.2) with each distance marked.

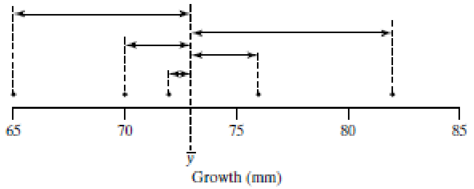


Figure 2.6.1 Plot of chrysanthemum growth data with deviations indicated as distances

From the formula for s, you can see that each deviation contributes to the SD. Thus, a sample of the same size but with less dispersion will have a smaller SD, as illustrated in the following example.

Example 2.6.4

Chrysanthemum Growth If the chrysanthemum growth data of Example 2.6.2 are changed to

75 72 73 75 70

then the mean is the same ($\bar{y} = 73$ mm), but the SD is smaller ($s = 2.1$ mm), because the observations lie closer to the mean. The relative dispersion of the two samples can easily be seen from Figure 2.6.2.

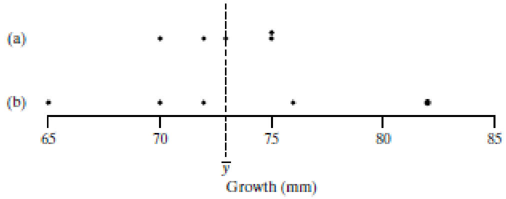


Figure 2.6.2 Two samples of chrysanthemum growth data with the same mean but different standard deviations: (a) $s = 2.1$ mm; (b) $s = 6.3$ mm

Let us look more closely at the way in which the deviations are combined to form the SD. The formula calls for dividing by $(n - 1)$. If the divisor were n instead of $(n - 1)$, then the quantity inside the square root sign would be the average (the mean) of the squared deviations. Unless n is very small, the inflation due to dividing by $(n - 1)$ instead of n is not very great, so that the SD can be interpreted approximately as

$$s \approx \sqrt{\text{sample average value of } (y_i - \bar{y})^2}$$

Thus, it is roughly appropriate to think of the SD as a "typical" distance of the observations from their mean.

Why $n - 1$? Since dividing by n seems more natural, you may wonder why the formula for the SD specifies dividing by $(n - 1)$. Note that the sum of the deviations $y_i - \bar{y}$ is always zero. Thus, once the first $n - 1$ deviations have been calculated, the last deviation is constrained. This means that in a sample with n observations, there are only $n - 1$ units of information concerning deviation from the average. The quantity $n - 1$ is called the **degrees of freedom** of the standard deviation or variance. We can also give an intuitive justification of why $n - 1$ is used by considering the extreme case when $n = 1$, as in the following example.

Example 2.6.5

Chrysanthemum Growth Suppose the chrysanthemum growth experiment of Example 2.6.2 had included only one plant, so that the sample consisted of the single observation

73

For this sample, $n = 1$ and $\bar{y} = 73$. However, the SD formula breaks down (giving $\frac{0}{0}$), so the SD cannot be computed. This is reasonable, because the sample gives no information about variability in chrysanthemum growth under the experimental conditions. If the formula for the SD said to divide by n, we would obtain an SD of zero, suggesting that there is little or no variability; such a conclusion hardly seems justified by observation of only one plant.

VISUALIZING MEASURES OF DISPERSION

The range and the interquartile range are easy to interpret. The range is the spread of all the observations, and the interquartile range is the spread of (roughly) the middle 50% of the observations. In terms of the histogram of a data set, the range can be visualized as (roughly) the width of the histogram. The quartiles are (roughly) the values that divide the area into four equal parts, and the interquartile range is the distance between the first and third quartiles. The following example illustrates these ideas.

Example 2.6.6

Daily Gain of Cattle The performance of beef cattle was evaluated by measuring their weight gain during a 140-day testing period on a standard diet. Table 2.6.2 gives the average daily gains (kg/day) for 39 bulls of the same breed (Charolais); the observations are listed in increasing order.[36] The values range from 1.18 kg/day to 1.92 kg/day. The quartiles are 1.29, 1.41, and 1.58 kg/day. Figure 2.6.3 shows a histogram of the data, the range, the quartiles, and the interquartile range (IQR). The shaded area represents the middle 50% (approximately) of the observations.

Table 2.6.2 Average daily gain (kg/day) of 39 Charolais bulls

1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.65	

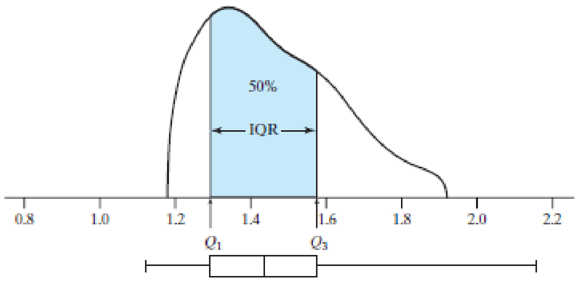


Figure 2.6.3 Smoothed histogram and boxplot of 39 daily gain measurements, showing the quartiles and the interquartile range (IQR). The shaded area represents about 50% of the observations.

VISUALIZING THE STANDARD DEVIATION

We have seen that the SD is a combined measure of the distances of the observations from their mean. It is natural to ask how many of the observations are within ±1 SD of the mean, within ±2 SDs of the mean, and so on. The following example explores this question.

Example 2.6.7

Daily Gain of Cattle For the daily-gain data of Example 2.6.6, the mean is $\bar{y} = 1.445$ kg/day and the SD is $s = 0.183$ kg/day. In Figure 2.6.4 the intervals $\bar{y} \pm s$, $\bar{y} \pm 2s$, and $\bar{y} \pm 3s$ have been marked on a histogram of the data. The interval $\bar{y} \pm s$ is

$$1.445 \pm 0.183 \text{ or } 1.262 \text{ to } 1.628$$

You can verify from Table 2.6.2 that this interval contains 25 of the 39 observations. Thus, $\frac{25}{39}$ or 64% of the observations are within ±1 SD of the mean; the corresponding area is shaded in Figure 2.6.4. The intervals $\bar{y} \pm 2s$ is

$$1.445 \pm 0.366 \text{ or } 1.079 \text{ to } 1.811$$

This interval contains $\frac{37}{39}$ or 95% of the observations. You may verify that the interval $y \pm 3s$ contains all the observations.

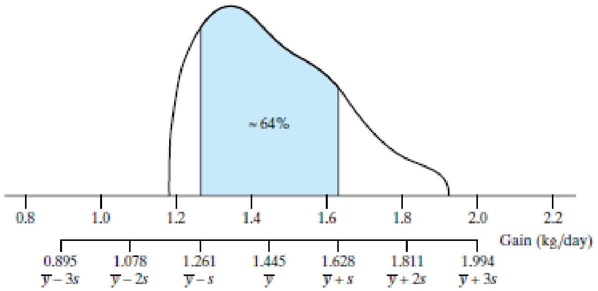


Figure 2.6.4 Histogram of daily-gain data showing intervals 1, 2, and 3 standard deviations from the mean. The shaded area represents about 64% of the observations.

It turns out that the percentages found in Example 2.6.7 are fairly typical of distributions that are observed in the life sciences.

Typical Percentages: The Empirical Rule

For "nicely shaped" distributions—that is, unimodal distributions that are not too skewed and whose tails are not overly long or short—we usually expect to find

about 68% of the observations within ±1 SD of the mean.

about 95% of the observations within ±2 SDs of the mean.

>99% of the observations within ±3 SDs of the mean.

The typical percentages enable us to construct a rough mental image of a frequency distribution if we know just the mean and SD. (The value 68% may seem to come from nowhere. Its origin will become clear in Chapter 4.)

ESTIMATING THE SD FROM A HISTOGRAM

The empirical rule gives us a way to construct a rough mental image of a frequency distribution if we know just the mean and SD: We can envision a histogram centered at the mean and extending out a bit more than 2 SDs in either direction. Of course, the actual distribution might not be symmetric, but our rough mental image will often be fairly accurate.

Thinking about this the other way around, we can look at a histogram and estimate the SD. To do this, we need to estimate the endpoints of an interval that is centered at the mean and that contains about 95% of the data. The empirical rule implies that this interval is roughly the same as $(\bar{y} - 2s, \bar{y} + 2s)$, so the length of the interval should be about 4 times the SD:

$$(\bar{y} - 2s, \bar{y} + 2s) \text{ has length of } 2s + 2s = 4s$$

This means

$$\text{length of interval} = 4s$$

so

$$\text{estimate of } s = \frac{\text{length of interval}}{4}$$

Of course, our visual estimate of the interval that covers the middle 95% of the data could be off. Moreover, the empirical rule works best for distributions that are symmetric. Thus, this method of estimating the SD will give only a general estimate. The method works best when the distribution is fairly symmetric, but it works reasonably well even if the distribution is somewhat skewed.

Example 2.6.8

Pulse after Exercise A group of 28 adults did some moderate exercise for 5 minutes and then measured their pulses. Figure 2.6.5 shows the distribution of the data.[37] We can see that about 95% of the observations are between about 75 and 125.* Thus, an interval of length 50(125 − 75) covers the middle 95% of the data. From this, we can estimate the SD to be $\frac{50}{4} = 12.5$. The actual SD is 13.4, which is not far off from our estimate.

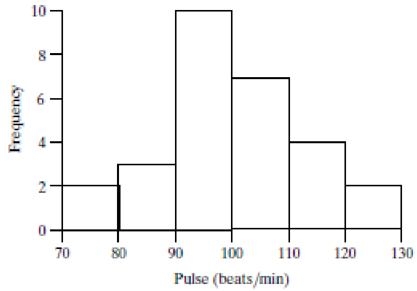


Figure 2.6.5 Pulse after moderate exercise for a group of adults

The typical percentages given by the empirical rule may be grossly wrong if the sample is small or if the shape of the frequency distribution is not "nice." For instance, the cricket singing time data (Table 2.3.1 and Figure 2.3.4) has $s = 4.4$ mm, and the interval $\bar{y} \pm s$ contains 90% of the observations. This is much higher than the "typical" 68% because the SD has been inflated by the long, straggly tail of the distribution.

COMPARISON OF MEASURES OF DISPERSION

The dispersion, or spread, of the data in a sample can be described by the standard deviation, the range, or the interquartile range.† The range is simple to understand, but it can be a poor descriptive measure because it depends only on the extreme tails of the distribution. The interquartile range, by contrast, describes the spread in the

*It is difficult to visually assess exactly where the middle 95% of the data lay using a histogram, but as this is only a visual estimate, we need not concern ourselves with producing an exact value. Our visual estimates of the SD might differ from one another, but they should all be relatively close.

†Another measure of dispersion is the **coefficient of variation**, which is the standard deviation expressed as a percentage of the mean: coefficient of variation $= \frac{s}{\bar{y}} * 100\%$. Because it is not affected by a change in scale (e.g., from inches to cm), the coefficient of variation is a useful measure for comparing the dispersions of two or more variables that are measured on different scales. See Exercises 2.6.13 and 2.6.14 for more information.

central "body" of the distribution. The standard deviation takes account of all the observations and can roughly be interpreted in terms of the spread of the observations around their mean. However, the SD can be inflated by observations in the extreme tails. The interquartile range is a robust measure, while the SD is not robust. Of course, the range is very highly nonrobust.

The descriptive interpretation of the SD is less straightforward than that of the range and the interquartile range. Nevertheless, the SD is the basis for most standard classical statistical methods. The SD enjoys this classic status for various technical reasons, including efficiency in certain situations.

The developments in later chapters will emphasize classical statistical methods, in which the mean and SD play a central role. Consequently, in this book we will rely primarily on the mean and SD rather than other descriptive measures.

Exercises 2.6.1–2.6.17

2.6.1 Calculate the SD of each of the following fictitious samples:

(a) 16, 13, 18, 13

(b) 38, 30, 34, 38, 35

(c) 1, −1, 5, −1

(d) 4, 6, −1, 4, 2

2.6.2 Calculate the SD of each of the following fictitious samples:

(a) 8, 6, 9, 4, 8

(b) 4, 7, 5, 4

(c) 9, 2, 6, 7, 6

2.6.3

(a) Invent a sample of size 5 for which the deviations $(y_i - \bar{y})$ are −3, −1, 0, 2, 2.

(b) Compute the SD of your sample.

(c) Should everyone get the same answer for part (b)? Why or why not?

2.6.4 Four plots of land, each 346 square feet, were planted with the same variety ("Beau") of wheat. The plot yields (lb) were as follows:[38]

35.1 30.6 36.9 29.8

Calculate the mean and the SD.

2.6.5 A plant physiologist grew birch seedlings in the greenhouse and measured the ATP content of their roots. (See Example 1.1.3.) The results (nmol ATP/mg tissue) were as follows for four seedlings that had been handled identically.[39]

1.45 1.19 1.05 1.07

Calculate the mean and the SD.

2.6.6 Ten patients with high blood pressure participated in a study to evaluate the effectiveness of the drug Timolol in reducing their blood pressure. The accompanying table shows systolic blood pressure measurements taken before and after 2 weeks of treatment with Timolol.[40] Calculate the mean and SD of the *change* in blood pressure (note that some values are negative).

	Blood pressure (mm HG)		
Patient	Before	After	Change
1	172	159	−13
2	186	157	−29
3	170	163	−7
4	205	207	2
5	174	164	−10
6	184	141	−43
7	178	182	4
8	156	171	15
9	190	177	−13
10	168	138	−30

2.6.7 Dopamine is a chemical that plays a role in the transmission of signals in the brain. A pharmacologist measured the amount of dopamine in the brain of each of seven rats. The dopamine levels (nmoles/g) were as follows:[41]

6.8 5.3 6.0 5.9 6.8 7.4 6.2

(a) Calculate the mean and SD.

(b) Determine the median and the interquartile range.

(c) Replace the observation 7.4 by 10.4 and repeat parts (a) and (b). Which of the descriptive measures display robustness and which do not?

2.6.8 In a study of the lizard *Sceloporus occidentalis*, biologists measured the distance (m) run in 2 minutes for each of 15 animals. The results (listed in increasing order) were as follows:[42]

18.4 22.2 24.5 26.4 27.5 28.7 30.6 32.9
32.9 34.0 34.8 37.5 42.1 45.5 45.5

(a) Determine the quartiles and the interquartile range.

(b) Determine the range.

2.6.9 Refer to the running-distance data of Exercise 2.6.8. The sample mean is 32.23 m and the SD is 8.07 m. What percentage of the observations are within

(a) 1 SD of the mean?

(b) 2 SDs of the mean?

2.6.10 Compare the results of Exercise 2.6.9 with the predictions of the empirical rule.

2.6.11 Listed in increasing order are the serum creatine phosphokinase (CK) levels (U/l) of 36 healthy men (these are the data of Example 2.2.6):

25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/l and the SD is 40.4 U/l. What percentage of the observations are within

(a) 1 SD of the mean?

(b) 2 SDs of the mean?

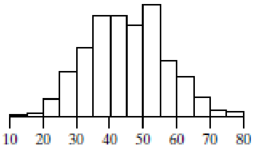
(c) 3 SDs of the mean?

2.6.12 Compare the results of Exercise 2.6.11 with the predictions of the empirical rule.

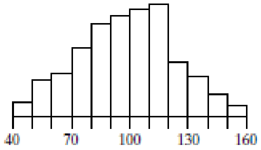
2.6.13 As part of the Berkeley Guidance Study[43] the heights (in cm) and weights (in kg) of 13 girls were measured at age 2 and again at age 9. Of course, the average height and weight were much greater at age 9 than at age 2. Likewise, the SDs of height and of weight were much greater at age 9 than they were at age 2. But what about the coefficient of variation, which gives the SD as a percentage of the mean? It turns out that the coefficient of variation for one of the variables (height or weight) went up only a moderate amount from age 2 to age 9, but for the other variable, the increase in the coefficient of variation was fairly large. For which variable, height or weight, would you expect the coefficient of variation to change more between age 2 and age 9? Why? [*Hint:* Think about how genetic and environmental factors each influence height and weight.]

2.6.14 Consider the 13 girls mentioned in Exercise 2.6.13. At age 18 their average height was 166.3 cm and the SD of their heights was 6.8 cm. Calculate the coefficient of variation.

2.6.15 Here is a histogram. Estimate the mean and the SD of the distribution.



2.6.16 Here is a histogram. Estimate the mean and the SD of the distribution.



2.6.17 For which sample (i or ii) would you expect the SD of heights to be larger? Or, would they be about the same?

(a) (i) A sample of 10 women ages 18–24, or (ii) a sample of 100 women ages 18–24.

(b) (i) A sample of 20 male college basketball players, or (ii) a sample of 20 college-age men.

(c) (i) A sample of 15 professional male jockeys, or (ii) a sample of 15 professional male biologists.

2.7 Effect of Transformation of Variables (Optional)

Sometimes when we are working with a data set, we find it convenient to transform a variable. For example, we might convert from inches to centimeters or from °F to °C. Transformation, or reexpression, of a variable Y means replacing Y by a new variable, say Y'. To be more comfortable working with data, it is helpful to know how the features of a distribution are affected if the observed variable is transformed.

The simplest transformations are **linear** transformations, so called because a graph of Y against Y' would be a straight line. A familiar reason for linear transformation is a change in the scale of measurement, as illustrated in the following two examples.

Example 2.7.1

Weight Suppose Y represents the weight of an animal in kg, and we decide to reexpress the weight in lb. Then

$$Y = \text{Weight in kg}$$
$$Y' = \text{Weight in lb}$$

so

$$Y' = 2.2Y$$

This is a **multiplicative** transformation, because Y' is calculated from Y by multiplying by the constant value 2.2. ■

Example 2.7.2

Body Temperature Measurements of basal body temperature (temperature on waking) were made on 47 women.[44]

Typical observations Y, in °C, were

$$Y\text{: } 36.23,\quad 36.41,\quad 36.77,\quad 36.15,\quad \ldots$$

Suppose we convert these data from °C to °F, and call the new variable Y':

$$Y'\text{: } 97.21,\quad 97.54,\quad 98.19,\quad 97.07,\quad \ldots$$

The relation between Y and Y' is

$$Y' = 1.8Y + 32$$

The combination of **additive** (+32) and multiplicative (×1.8) changes indicates a linear relationship. ■

As the foregoing examples illustrate, a linear transformation consists of (1) multiplying all the observations by a constant, or (2) adding a constant to all the observations, or (3) both.

HOW LINEAR TRANSFORMATIONS AFFECT THE FREQUENCY DISTRIBUTION

A linear transformation of the data does not change the essential shape of its frequency distribution; by suitably scaling the horizontal axis, you can make the transformed histogram identical to the original histogram. Example 2.7.3 illustrates this idea.

Example 2.7.3

Body Temperature Figure 2.7.1 shows the distribution of 47 temperature measurements that have been transformed by first subtracting 36 from each observation and then multiplying by 100 (as in Example 2.7.2). That is, $Y' = (Y - 36) \times 100$. The figure shows that the two distributions can be represented by the same histogram with different horizontal scales. ■

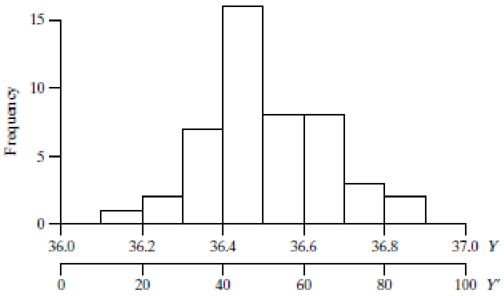


Figure 2.7.1 Distribution of 47 temperature measurements showing original and linearly transformed scales

HOW LINEAR TRANSFORMATIONS AFFECT $\bar{y}$ AND s

The effect of a linear transformation on $\bar{y}$ is "natural"; that is, **under a linear transformation, $\bar{y}$** changes like Y. For instance, if temperatures are converted from °C to °F, then the mean is similarly converted:

$$Y' = 1.8Y + 32 \qquad \text{so} \qquad \bar{y}' = 1.8\bar{y}' + 32$$

The effect of multiplying Y by a positive constant on s is "natural"; if $Y' = c \times Y$, with $c > 0$, then $s' = c \times s$. For instance, if weights are converted from kg to lb, the SD is similarly converted: $s' = 2.2s$. If $Y' = c \times Y$ and $c < 0$, then $s' = -c \times s$. In general, if $Y' = c \times Y$ then $s' = |c| \times s$.

However, an additive transformation does not affect s. If we add or subtract a constant, we do not change how spread out the distribution is, so s does not change. Thus, for example, we would *not* convert the SD of temperature data from °C to °F in the same way as we convert each observation; we would multiply the SD by 1.8, but we would *not* add 32. The fact that the SD is unchanged by additive transformation will appear less surprising if you recall (from the definition) that s depends only on the deviations $(y_i - \bar{y})$, and these are not changed by an additive transformation. The following example illustrates this idea.

Example 2.7.4

Additive Transformation Consider the simple set of fictitious data shown in Table 2.7.1. The data were then transformed by subtracting 20 from each observation.

The SD for the original observations is

$$s = \sqrt{\frac{(-1)^2 + (0)^2 + (2)^2 + (-1)^2}{3}}$$
$$= 1.4$$

Table 2.7.1 Effect of additive transformation

	Original observations (y)	Deviations $(y_i - \bar{y})$	Transformed observations (y')	Deviations $(y'_i - \bar{y}')$
	25	−1	5	−1
	26	0	6	0
	28	2	8	2
	25	−1	5	−1
Mean	26		6	

Because the deviations are unaffected by the transformation, the SD for the transformed observations is the same:

$$s' = 1.4$$

An additive transformation effectively picks up the histogram of a distribution and moves it to the left or to the right on the number line. The shape of the histogram does not change and the deviations do not change, so the SD does not change. A multiplicative transformation, on the other hand, stretches or shrinks the distribution, so the SD gets larger or smaller accordingly.

Other Statistics Under linear transformations, other measures of center (e.g., the median) change like $\bar{y}$, and other measures of dispersion (e.g., the interquartile range) change like s. The quartiles themselves change like $\bar{y}$.

NONLINEAR TRANSFORMATIONS

Data are sometimes reexpressed in a nonlinear way. Examples of nonlinear transformations are

$$Y' = \sqrt{Y}$$
$$Y' = \log(Y)$$
$$Y' = \frac{1}{Y}$$
$$Y' = Y^2$$

These transformations are termed "nonlinear" because a graph of Y' against Y would be a curve rather than a straight line. Computers make it easy to use nonlinear transformations. The logarithmic transformation is especially common in biology because many important relationships can be simply expressed in terms of logs. For instance, there is a phase in the growth of a bacterial colony when log(colony size) increases at a constant rate with time. [Note that logarithms are used in some familiar scales of measurement, such as pH measurement or earthquake magnitude (Richter scale).]

Nonlinear transformations can affect data in complex ways. For example, the mean does not change "naturally" under a log transformation; the log of the mean is *not* the same as the mean of the logs. Furthermore, nonlinear transformations (unlike linear ones) *do* change the essential shape of a frequency distribution.

In future chapters we will see that if a distribution is skewed to the right, such as the cricket singing-time distribution shown in Figure 2.7.2, then we may wish to apply a transformation that makes the distribution more symmetric, by pulling in the right-hand tail. Using $Y' = \sqrt{Y}$ will pull in the right-hand tail of a distribution and push out the left-hand tail. The transformation $Y' = \log(Y)$ is more severe than $\sqrt{Y}$ in this regard. The following example shows the effect of these transformations.

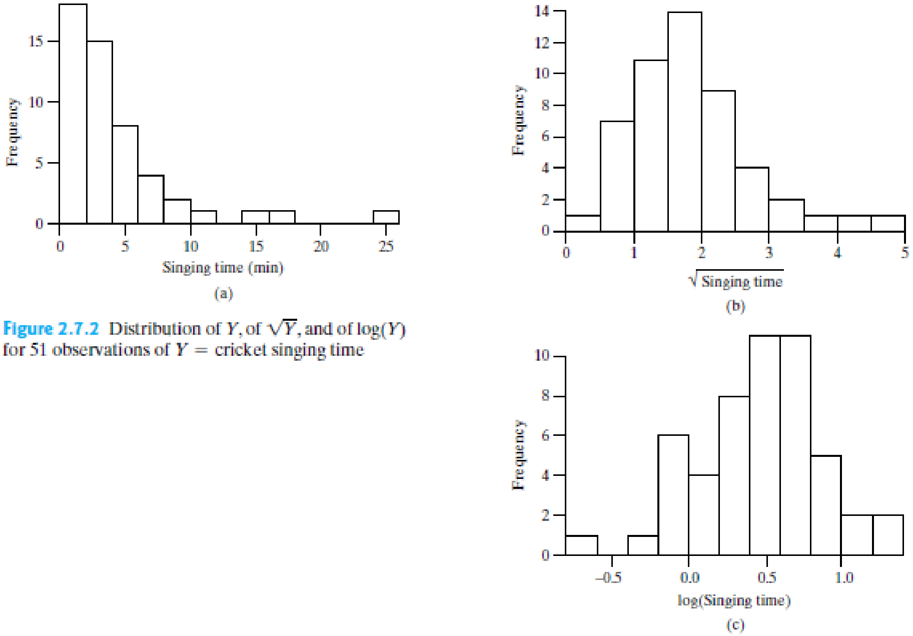


Figure 2.7.2 Distribution of Y, of $\sqrt{Y}$, and of $\log(Y)$ for 51 observations of Y = cricket singing time

Example 2.7.5 **Cricket Singing Times** Figure 2.7.2(a) shows the distribution of the cricket singing-time data of Table 2.3.1. If we transform these data by taking square roots, the transformed data have the distribution shown in Figure 2.7.2(b). Taking logs (base 10) yields the distribution shown in Figure 2.7.2(c). Notice that the transformations have the effect of "pulling in" the straggly upper tail and "stretching out" the clumped values on the lower end of the original distribution.

Exercises 2.7.1–2.7.6

2.7.1 A biologist made a certain pH measurement in each of 24 frogs; typical values were[45]

$$7.43, \quad 7.16, \quad 7.51, \quad \ldots$$

She calculated a mean of 7.373 and a SD of 0.129 for these original pH measurements. Next, she transformed the data by subtracting 7 from each observation and then multiplying by 100. For example, 7.43 was transformed to 43. The transformed data are

$$43, \quad 16, \quad 51, \ldots$$

What are the mean and SD of the transformed data?

2.7.2 The mean and SD of a set of 47 body temperature measurements were as follows:[46]

$$\bar{y} = 36.497\,^\circ\text{C} \qquad s = 0.172\,^\circ\text{C}$$

If the 47 measurements were converted to °F,

(a) What would be the new mean and SD?

(b) What would be the new coefficient of variation?

2.7.3 A researcher measured the average daily gains (in kg/day) of 20 beef cattle; typical values were[47]

1.39, 1.57, 1.44, ...

The mean of the data was 1.461 and the SD was 0.178.

(a) Express the mean and SD in lb/day. (*Hint:* 1 kg = 2.20 lb.)

(b) Calculate the coefficient of variation when the data are expressed (i) in kg/day; (ii) in lb/day.

2.7.4 Consider the data from Exercise 2.7.3. The mean and SD were 1.461 and 0.178. Suppose we transformed the data from

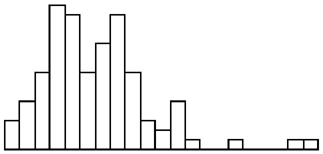
1.39, 1.57, 1.44, ...

to

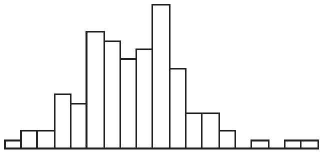
39, 57, 44, ...

What would be the mean and SD of the transformed data?

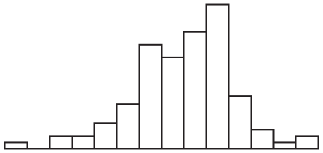
2.7.5 The following histogram shows the distribution for a sample of data:



One of the following histograms is the result of applying a square root transformation, and the other is the result of applying a log transformation. Which is which? How do you know?



(a)



(b)

2.7.6 (Computer problem) The file "Exercise 2.7.6.csv" is included on the data disk packaged with this text. This file contains 36 observations on the number of dendritic branch segments emanating from nerve cells taken from the brains of newborn guinea pigs. (These data were used in Exercise 2.2.4.) Open the file and enter the data into a statistics package. Make a histogram of the data, which are skewed to the right. Now consider the following possible transformations: sqrt(Y), log(Y), and 1/sqrt(Y). Which of these transformations does the best job of meeting the goal of making the resulting distribution reasonably symmetric?

2.8 Statistical Inference

The description of a data set is sometimes of interest for its own sake. Usually, however, the researcher hopes to generalize, to extend the findings beyond the limited scope of the particular group of animals, plants, or other units that were actually observed. Statistical theory provides a rational basis for this process of generalization, building on the random sampling model from Section 1.3 and taking into account the variability of the data. The key idea of the statistical approach is to view the particular data in a study as a sample from a larger population; the population is the real focus of scientific and/or practical interest. The following example illustrates this idea.

Example 2.8.1

Blood Types In an early study of the ABO blood-typing system, researchers determined blood types of 3,696 persons in England. The results are given in Table 2.8.1.[48]

Table 2.8.1 Blood types of 3,696 persons

Blood type	Frequency
A	1,634
B	327
AB	119
O	1,616
Total	3,696

These data were not collected for the purpose of learning about the blood types of those particular 3,696 people. Rather, they were collected for their scientific value as a source of information about the distribution of blood types in a larger population. For instance, one might presume that the blood type distribution of all English people should resemble the distribution for these 3,696 people. In particular, the observed relative frequency of type A blood was

$$\frac{1634}{3696} \text{ or } 44\% \text{ type A}$$

One might conclude from this that approximately 44% of the people in England have type A blood.

The process of drawing conclusions about a population, based on observations in a sample from that population, is called **statistical inference**. For instance, in Example 2.8.1 the conclusion that approximately 44% of the people in England have type A blood would be a statistical inference. The inference is shown schematically in Figure 2.8.1. Of course, such an inference might be entirely wrong—perhaps the 3,696 people are not at all representative of English people in general. We might be worried about two possible sources of difficulty: (1) the 3,696 people might have been selected in a way that was systematically biased for (or against) type A people, and (2) the number of people examined might have been too small to permit generalization to a population of many millions. In general, it turns out that the population size being in the millions is *not* a problem, but bias in the way people are selected is a big concern.

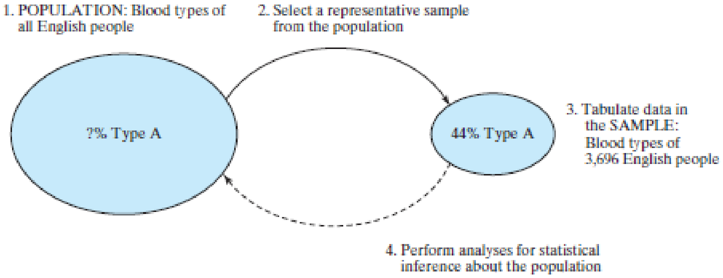

1. POPULATION: Blood types of all English people
2. Select a representative sample from the population
?% Type A
44% Type A
3. Tabulate data in the SAMPLE: Blood types of 3,696 English people
4. Perform analyses for statistical inference about the population

Figure 2.8.1 Schematic representation of inference from sample to population regarding prevalence of blood type A

In making a statistical inference, we hope that the sample resembles the population closely—that the sample is *representative* of the population. In Section 1.3 we saw how sampling bias can lead to nonrepresentative samples. However, even in the absence of bias, we must ask how likely it is that a particular sample will provide a good representation of the population. The important question is: *How representative (of the population) is a sample likely to be?* We will see in Chapter 5 how statistical theory can help to answer this question.

SPECIFYING THE POPULATION

In Section 1.3 we emphasized that the collection of individuals that comprise a sample should be representative of the population. In fact, this requirement is a bit stronger than what is actually necessary. Ultimately, what matters is that the measurements that we obtain on the variable of interest are representative of the values present in the population. The following provides an example of a case where the sample members might not be representative of the population, but one could argue that the measurements taken from this sample could be viewed as representative of the larger population.

Example 2.8.2

Blood Types How were the 3,696 English people of Example 2.8.1 actually chosen? It appears from the original paper that this was a "sample of convenience," that is, friends of the investigators, employees, and sundry unspecified sources. There is little basis for believing that the *people* themselves would be representative of the entire English population. Nevertheless, one might argue that their *blood types* might be (more or less) representative of the population. The argument would be that the biases that entered into the selection of those particular people were probably not related to blood type. [Nonetheless, an objection to this argument might be made on the basis of race. For example, the racial distribution of the sample could differ substantially from the racial distribution of England (the population), and there are known differences in blood type distributions among races.] The argument for representativeness would be much less plausible if the observed variable were blood pressure rather than blood type; we know that blood pressure tends to increase with age, and the selection procedure was undoubtedly biased against certain age groups (e.g., elderly people).

As Example 2.8.2 shows, whether the measurements obtained from a sample are likely to be representative of the measurements from a population depends not only on how the observational units (in this case people) were chosen, but also on the variable that was observed. Ideally we would always work with random samples, but we have noted that in some instances random samples are not possible or convenient. However, by turning our attention to the measurements themselves rather than the individuals from which they came, we can often make an argument for the generalizabiltity (or lack of generalizability) of our results to a larger population. We do this by thinking of the population as consisting of observations or a collection of values from a measurement process, rather than of people or other observational units. The following is another example.

Example 2.8.3

Alcohol and MOPEG The biochemical MOPEG plays a role in brain function. Seven healthy male volunteers participated in a study to determine whether drinking alcohol might elevate the concentration of MOPEG in the cerebrospinal fluid. The MOPEG concentration was measured twice for each man—once at the start of the experiment, and again after he drank 80 gm of ethanol. The results (in pmol/ml) are given in Table 2.8.2.[49]

Let us focus on the rightmost column, which shows the change in MOPEG concentration (i.e., the difference between the "after" and the "before" measurements). In thinking of these values as a sample from a population, we need to specify all the details of the experimental conditions—how the cerebrospinal specimens were obtained, the exact timing of the measurements and the alcohol consumption, and so on—as well as relevant characteristics of the volunteers themselves. Thus, the definition of the population might be something like this:

Table 2.8.2 Effect of alcohol on MOPEG

	MOPEG concentration		
Volunteer	Before	After	Change
1	46	56	10
2	47	52	5
3	41	47	6
4	45	48	3
5	37	37	0
6	48	51	3
7	58	62	4

> **Population** Change in cerebrospinal MOPEG concentration in healthy young men when measured before and after drinking 80 gm of ethanol, both measurements being made at 8:00 A.M., . . . (other relevant experimental conditions are specified here).

There is no single "correct" definition of a population for an experiment like this. A scientist reading a report of the experiment might find this definition too narrow (e.g., perhaps it does not matter that the volunteers were measured at 8:00 A.M.) or too broad. She might use her knowledge of alcohol and brain chemistry to formulate her own definition, and she would then use that definition as a basis for interpreting these seven observations.

DESCRIBING A POPULATION

Because observations are made only on a sample, characteristics of biological populations are almost never known exactly. Typically, our knowledge of a population characteristic comes from a sample. In statistical language, we say that the sample characteristic is an estimate of the corresponding population characteristic. Thus, estimation is a type of statistical inference.

Just as each sample has a distribution, a mean, and an SD, so also we can envision a population distribution, a population mean, and a population SD. In order to discuss inference from a sample to a population, we will need a language for describing the population. This language parallels the language that describes the sample. A sample characteristic is called a **statistic**; a population characteristic is called a **parameter**.

PROPORTIONS

For a categorical variable, we can describe a population by simply stating the proportion, or relative frequency, of the population in each category. The following is a simple example.

Example 2.8.4

Oat Plants In a certain population of oat plants, resistance to crown rust disease is distributed as shown in Table 2.8.3.[50]

Table 2.8.3 Disease resistance in oats

Resistance	Proportion of plants
Resistant	0.47
Intermediate	0.43
Susceptible	0.10
Total	1.00

Remark The population described in Example 2.8.4 is realistic, but it is not a specific real population; the exact proportions for any real population are not known. For similar reasons, we will use fictitious but realistic populations in several other examples, here and in Chapters 3, 4, and 5

For categorical data, the sample proportion of a category is an estimate of the corresponding population proportion. Because these two proportions are not necessarily the same, it is essential to have a notation that distinguishes between them. We denote the population proportion of a category by p and the sample proportion by $\hat{p}$ (read "p-hat"):

$$p = \text{Population proportion}$$

$$\hat{p} = \text{Sample proportion}$$

The symbol "^" can be interpreted as "estimate of." Thus,

$$\hat{p} \text{ is an estimate of } p$$

We illustrate this notation with an example.

Example 2.8.5 **Lung Cancer** Eleven patients suffering from adenocarcinoma (a type of lung cancer) were treated with the chemotherapeutic agent Mitomycin. Three of the patients showed a positive response (defined as shrinkage of the tumor by at least 50%).[51] Suppose we define the population for this study as "responses of all adenocarcinoma patients." Then we can represent the sample and population proportions of the category "positive response" as follows:

$$p = \text{Proportion of positive responders among all adenocarcinoma patients}$$

$$\hat{p} = \text{Proportion of positive responders among the 11 patients in the study}$$

$$\hat{p} = \frac{3}{11} = 0.27$$

Note that p is unknown, and $\hat{p}$, which is known, is an estimate of p.

We should emphasize that an "estimate," as we are using the term, may or may not be a *good* estimate. For instance, the estimate $\hat{p}$ in Example 2.8.5 is based on very few patients; estimates based on a small number of observations are subject to considerable uncertainty. Of course, the question of whether an estimation procedure is good or poor is an important one, and we will show in later chapters how this question can be answered.

OTHER DESCRIPTIVE MEASURES

If the observed variable is quantitative, one can consider descriptive measures other than proportions—the mean, the quartiles, the SD, and so on. Each of these quantities can be computed for a sample of data, and each is an estimate of its corresponding population analog. For instance, the sample median is an estimate of the population median. In later chapters, we will focus especially on the mean and the SD, and so we will need a special notation for the population mean and SD. **The population mean is denoted by μ (mu), and the population SD is denoted by σ (sigma).** We may define these as follows for a quantitative variable Y:

$$\mu = \text{Population average value of } Y$$

$$\sigma = \sqrt{\text{Population average value of } (Y - \mu)^2}$$

The following example illustrates this notation.

Example 2.8.6 **Tobacco Leaves** An agronomist counted the number of leaves on each of 150 tobacco plants of the same strain (Havana). The results are shown in Table 2.8.4.[52]

The sample mean is

$$\bar{y} = 19.78 = \text{Mean number of leaves on the 150 plants}$$

Table 2.8.4 Number of leaves on tobacco plants

Number of leaves	Frequency (number of plants)
17	3
18	22
19	44
20	42
21	22
22	10
23	6
24	1
Total	150

The population mean is

$$\mu = \text{Mean number of leaves on Havana tobacco plants grown under these conditions}$$

We do not know μ, but we can regard $\bar{y} = 19.78$ as an estimate of μ. The sample SD is

$$s = 1.38 = \text{SD of number of leaves on the 150 plants}$$

The population SD is

$$\sigma = \text{SD of number of leaves on Havana tobacco plants grown under these conditions}$$

We do not know σ, but we can regard as an estimate of σ.*

*You may wonder why we use $\bar{y}$ and s instead of $\hat{\mu}$ and $\hat{\sigma}$. One answer is tradition. Another answer is that since "^" means estimate, you might have other estimates in mind.

2.9 Perspective

In this chapter we have considered various ways of describing a set of data. We have also introduced the notion of regarding features of a sample as estimates of corresponding features of a suitably defined population.

PARAMETERS AND STATISTICS

Some features of a distribution—for instance, the mean—can be represented by a single number, while some—for instance, the shape—cannot. We have noted that a numerical measure that describes a sample is called a statistic. Correspondingly, a numerical measure that describes a population is called a parameter. For the most important numerical measures, we have defined notations to distinguish between the statistic and the parameter. These notations are summarized in Table 2.9.1 for convenient reference.

Table 2.9.1 Notation for some important statistics and parameters

Measure	Sample value (statistic)	Population value (parameter)
Proportion	$\hat{p}$	p
Mean	$\bar{y}$	μ
Standard deviation	s	σ

A LOOK AHEAD

It is natural to view a sample characteristic (e.g., $\bar{y}$) as an estimate of the corresponding population characteristic (e.g., μ). But in taking such a view, one must guard against unjustified optimism. Of course, if the sample were perfectly representative of the population, then the estimate would be perfectly accurate. But this raises the central question: How representative (of the population) is a sample likely to be? Intuition suggests that, if the observational units are appropriately selected, then the sample should be more or less representative of the population. Intuition also suggests that larger samples should tend to be more representative than smaller samples. These intuitions are basically correct, but they are too vague to provide practical guidance for research in the life sciences. Practical questions that need to be answered are

1. How can an investigator judge whether a sample can be viewed as "more or less" representative of a population?
2. How can an investigator quantify "more or less" in a specific case?

In Section 1.3 we described a theoretical probability model based on random sampling that provides a framework for the judgment in question (1), and in Chapter 6 we will see how this model can provide a concrete answer to question (2). Specifically, in Chapter 6 we will see how to analyze a set of data so as to quantify how closely the sample mean ($\bar{y}$) estimates the population mean (μ). But before returning to data analysis in Chapter 6, we will need to lay some groundwork in Chapters 3, 4, and 5; the developments in these chapters are an essential prelude to understanding the techniques of statistical inference.

Supplementary Exercises 2.S.1–2.S.24

2.S.1 If 2, 7, 10, and 1 are the number of students who weigh (in kg) 35, 30, 42, and 40, respectively, what is the mean weight of these 20 students?

2.S.2 A botanist grew 15 pepper plants on the same greenhouse bench. After 21 days, she measured the total stem length (cm) of each plant, and obtained the following values:[53]

12.4	12.2	13.4
10.9	12.2	12.1
11.8	13.5	12.0
14.1	12.7	13.2
12.6	11.9	13.1

(a) Calculate all three quartiles.

(b) Compute the lower fence and the upper fence of the distribution.

(c) How large would an observation in this data set have to be in order to be an outlier?

2.S.3 In a behavioral study of the fruitfly *Drosophila melanogaster*, a biologist measured, for individual flies, the total time spent preening during a 6-minute observation period. The following are the preening times (sec) for 20 flies:[54]

34	24	10	16	52
76	33	31	46	24
18	26	57	32	25
48	22	48	29	19

(a) Determine the mode (s).

(b) Calculate the range.

(c) Construct a dotplot of the data.

2.S.4 To calibrate a standard curve for assaying protein concentrations, a plant pathologist used a spectrophotometer to measure the absorbance of light (wavelength 500 nm) by a protein solution. The results of 27 replicate assays of a standard solution containing 60 μg protein per ml water were as follows:[55]

0.111	0.115	0.115	0.110	0.099
0.121	0.107	0.107	0.100	0.110
0.106	0.116	0.098	0.116	0.108
0.098	0.120	0.123	0.124	0.122
0.116	0.130	0.114	0.100	0.123
0.119	0.107			

Construct a frequency distribution and display it as a table and as a histogram.

2.S.5 Refer to the absorbance data of Exercise 2.S.4.

(a) Determine the median, the quartiles, and the interquartile range.

(b) How large must an observation be to be an outlier?

2.S.6 The median splits data into two equal halves. Is the median a robust statistic? Why or why not?

2.S.7 Twenty patients with severe epilepsy were observed for 8 weeks. The following are the numbers of major seizures suffered by each patient during the observation period:[56]

5	0	9	6	0	0	5	0	6	1
5	0	0	0	0	7	0	0	4	7

(a) Is the distribution of seizures bimodal? Justify your answer.

(b) Calculate the SD of seizures.

(c) What percentage of the seizures is within 1 SD of the mean?

(d) What percentage of the seizures is within 2 SDs of the mean?

2.S.8 Consider the histogram from Exercise 2.3.13. By "reading" the histogram, estimate the percentage of observations that are less than 45. Is this percentage closest to 10%, 30%, 50%, 70%, 90%? (*Note*: The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class. Rather, the percentage of observations that are less than 45 can be estimated by looking at area.)

2.S.9 Consider the histogram from Exercise 2.3.15. By "reading" the histogram, estimate the percentage of observations that are greater than 25. Is this percentage closest to 10%, 30%, 50%, 70%, 90%? (*Note*: The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class. Rather, the percentage of observations that are greater than 25 can be estimated by looking at area.)

2.S.10 Calculate the variance of each of the following fictitious samples:

(a) 12, 6, 7, 3

(b) 9, 13, 8, 10

(c) 21, 15, −10, 6

2.S.11 To study the spatial distribution of Japanese beetle larvae in the soil, researchers divided a 12- × 12-foot section of a cornfield into 144 one-foot squares. They counted the number of larvae Y in each square, with the results shown in the following table.[57]

Number of larvae	Frequency (Number of squares)
0	13
1	34
2	50
3	18
4	16
5	10
6	2
7	1
Total	144

(a) The mean and SD of Y are $\bar{y} = 2.23$ and $s = 1.47$. What percentage of the observations are within

(i) 1 SD of the mean?

(ii) 2 SDs of the mean?

(b) Determine the total number of larvae in all 144 squares. How is this number related to $\bar{y}$?

(c) Determine the median value of the distribution.

2.S.12 One measure of physical fitness is maximal oxygen uptake, which is the maximum rate at which a person can consume oxygen. A treadmill test was used to determine the maximal oxygen uptake of nine college women before and after participation in a 10-week program of vigorous exercise. The accompanying table shows the before and after measurements and the change (after–before); all values are in ml O_2 per mm per kg body weight.[58]

	Maximal oxygen uptake		
Participant	Before	After	Change
1	48.6	38.8	−9.8
2	38.0	40.7	2.7
3	31.2	32.0	0.8
4	45.5	45.4	−0.1
5	41.7	43.2	1.5
6	41.8	45.3	3.5
7	37.9	38.9	1.0
8	39.2	43.5	4.3
9	47.2	45.0	−2.2

The following computations are to be done on the *change* in maximal oxygen uptake (the right-hand column).

(a) Calculate the mean and the SD.

(b) Determine the median.

(c) Eliminate participant 1 from the data and repeat parts (a) and (b). Which of the descriptive measures display robustness and which do not?

2.S.13 A veterinary anatomist investigated the spatial arrangement of the nerve cells in the intestine of a pony. He removed a block of tissue from the intestinal wall, cut the block into many equal sections, and counted the number of nerve cells in each of 23 randomly selected sections. The counts were as follows.[59]

35	19	33	34	17	26	16	40
28	30	23	12	27	33	22	31
28	28	35	23	23	19	29	

(a) Determine the median, the quartiles, and the interquartile range.

(b) Construct a boxplot of the data.

2.S.14 Exercise 2.S.13 asks for a boxplot of the nerve-cell data. Does this graphic support the claim that the data came from a reasonably symmetric distribution?

2.S.15 A geneticist counted the number of bristles on a certain region of the abdomen of the fruitfly *Drosophila melanogaster*. The results for 119 individuals were as shown in the table.[60]

Number of bristles	Number of flies	Number of bristles	Number of flies
29	1	38	18
30	0	39	13
31	1	40	10
32	2	41	15
33	2	42	10
34	6	43	2
35	9	44	2
36	11	45	3
37	12	46	2

(a) Find the mean number of bristles.

(b) Find the SD of the sample.

(c) What percentage of the observations fall within 3 SDs of the mean?

(d) What is the coefficient of variation?

2.S.16 The carbon monoxide in cigarettes is thought to be hazardous to the fetus of a pregnant woman who smokes. In a study of this hypothesis, blood was drawn from pregnant women before and after smoking a cigarette. Measurements were made of the percent of blood hemoglobin bound to carbon monoxide as carboxyhemoglobin (COHb). The results for 10 women are shown in the table.[61]

	Blood COHb (%)		
Subject	Before	After	Increase
1	1.2	7.6	6.4
2	1.4	4.0	2.6
3	1.5	5.0	3.5
4	2.4	6.3	3.9
5	3.6	5.8	2.2
6	0.5	6.0	5.5
7	2.0	6.4	4.4
8	1.5	5.0	3.5
9	1.0	4.2	3.2
10	1.7	5.2	3.5

(a) Calculate the mean and SD of the *increase* in COHb.

(b) Calculate the mean COHb before and the mean after. Is the mean increase equal to the increase in means?

(c) Determine the median increase in COHb.

(d) Repeat part (c) for the before measurements and for the after measurements. Is the median increase equal to the increase in medians?

2.S.17 (Computer problem) A medical researcher in India obtained blood specimens from 31 young children, all of whom were infected with malaria. The following data, listed in increasing order, are the numbers of malarial parasites found in 1 ml of blood from each child.[62]

100	140	140	271	400	435	455	770
826	1,400	1,540	1,640	1,920	2,280	2,340	3,672
4,914	6,160	6,560	6,741	7,609	8,547	9,560	10,516
14,960	16,855	18,600	22,995	29,800	83,200	134,232	

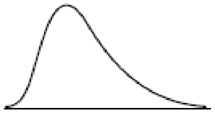
(a) Transform the data by taking the square root of each observation.

(b) Construct the frequency distribution of the data using a class width of 50.

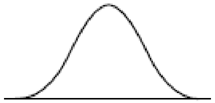
(c) Determine the IQR of the original data and the IQR of the log-transformed data. Is the IQR of the logs equal to the log of the IQR?

(d) Determine the SD of the original data and the SD of the log-transformed data. Is the SD of the logs equal to the log of the SD?

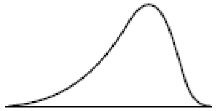
2.S.18 Rainfall, measured in inches, for the month of June in Cleveland, Ohio, was recorded for each of 41 years.[63] The values had a minimum of 1.2, an average of 3.6, and an SD of 1.6. Which of the following is a rough histogram for the data? How do you know?



I

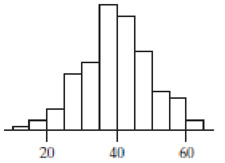


II

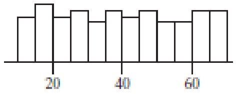


III

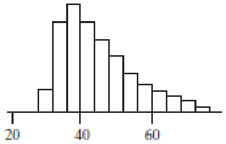
2.S.19 The following histograms (a), (b), and (c) show three distributions.



(a)



(b)



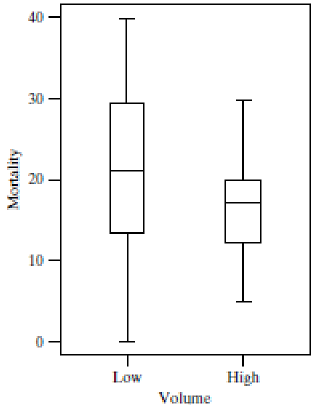
(c)

The accompanying computer output shows the mean, median, and SD of the three distributions, plus the mean, median, and SD for a fourth distribution. Match the histograms with the statistics. Explain your reasoning. (One set of statistics will not be used.)

```
1. Count    100          2. Count    100
   Mean     41.3522         Mean     39.6761
   Median   39.5585         Median   39.5377
   StdDev   13.0136         StdDev   10.0476

3. Count    100          4. Count    100
   Mean     37.7522         Mean     39.6493
   Median   39.5585         Median   39.5448
   StdDev   13.0136         StdDev   17.5126
```

2.S.20 The following boxplots show mortality rates (deaths within one year per 100 patients) for heart transplant patients at various hospitals. The low-volume hospitals are those that perform between 5 and 9 transplants per year. The high-volume hospitals perform 10 or more transplants per year.[64] Describe the distributions, paying special attention to how they compare to one another. Be sure to note the shape, center, and spread of each distribution.

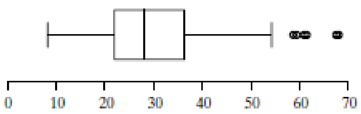


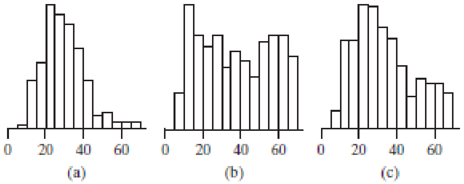
2.S.21 **(Computer problem)** Physicians measured the concentration of calcium (nM) in blood samples from 38 healthy persons. The data are listed as follows:[65]

95	110	135	120	88	125
112	100	130	107	86	130
122	122	127	107	107	107
88	126	125	112	78	115
78	102	103	93	88	110
104	122	112	80	121	126
90	96				

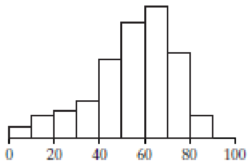
Calculate appropriate measures of the center and spread of the distribution. Describe the shape of the distribution and any unusual features in the data.

2.S.22 The following boxplot shows the same data that are shown in one of the three histograms. Which histogram goes with the boxplot? Explain your answer.





2.S.23 Here is a histogram.



Explain why the mean is less than the median of the distribution.

2.S.24 Consider the histogram from Exercise 2.S.23. By "reading" the histogram, estimate the mode of the distribution. *Note:* The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class.

Chapter 3

PROBABILITY AND THE BINOMIAL DISTRIBUTION

OBJECTIVES

In this chapter we will study the basic ideas of probability, including

- the "limiting frequency" definition of probability.
- the use of probability trees.
- the concept of a random variable.
- rules for finding means and standard deviations of random variables.
- the use of the binomial distribution.

3.1 Probability and the Life Sciences

Probability, or chance, plays an important role in scientific thinking about living systems. Some biological processes are affected directly by chance. A familiar example is the segregation of chromosomes in the formation of gametes; another example is the occurrence of mutations.

Even when the biological process itself does not involve chance, the results of an experiment are always somewhat affected by chance: chance fluctuations in environmental conditions, chance variation in the genetic makeup of experimental animals, and so on. Often, chance also enters directly through the design of an experiment; for instance, varieties of wheat may be randomly allocated to plots in a field. (Random allocation will be discussed in Chapter 11.)

The conclusions of a statistical data analysis are often stated in terms of probability. Probability enters statistical analysis not only because chance influences the results of an experiment, but also because probability models allow us to quantify how likely, or unlikely, an experimental result is, given certain modeling assumptions. In this chapter we will introduce the language of probability and develop some simple tools for calculating probabilities.

3.2 Introduction to Probability

In this section we introduce the language of probability and its interpretation.

BASIC CONCEPTS

A **probability** is a numerical quantity that expresses the likelihood of an event. The probability of an event E is written as

$$\Pr\{E\}$$

The probability $\Pr\{E\}$ is always a number between 0 and 1, inclusive.

We can speak meaningfully about a probability $\Pr\{E\}$ only in the context of a chance operation—that is, an operation whose outcome is not deterministic. The chance operation must be defined in such a way that *each time the chance operation is performed, the event E either occurs or does not occur.* The following two examples illustrate these ideas.

Example 3.2.1 **Coin Tossing** Consider the familiar chance operation of tossing a coin, and define the event

$$E\text{: Heads}$$

Each time the coin is tossed, either it falls heads or it does not. If the coin is equally likely to fall heads or tails, then

$$\Pr\{E\} = \frac{1}{2} = 0.5$$

Such an ideal coin is called a "fair" coin. If the coin is not fair (perhaps because it is slightly bent), then Pr{E} will be some value other than 0.5, for instance,

$$\Pr\{E\} = 0.6$$ ■

Example 3.2.2 **Coin Tossing** Consider the event

$$E\text{: 3 heads in a row}$$

The chance operation "toss a coin" is *not* adequate for this event, because we cannot tell from one toss whether E has occurred. A chance operation that would be adequate is

Chance operation: Toss a coin 3 times. ■

The language of probability can be used to describe the results of random sampling from a population. The simplest application of this idea is a sample of size $n = 1$; that is, choosing one member at random from a population. The following is an illustration.

Example 3.2.3 **Sampling Fruitflies** A large population of the fruitfly *Drosophila melanogaster* is maintained in a lab. In the population, 30% of the individuals are black because of a mutation, while 70% of the individuals have the normal gray body color. Suppose one fly is chosen at random from the population. Then the probability that a black fly is chosen is 0.3. More formally, define

$$E\text{: Sampled fly is black}$$

Then

$$\Pr\{E\} = 0.3$$ ■

The preceding example illustrates the basic relationship between probability and random sampling: *The probability that a randomly chosen individual has a certain characteristic is equal to the proportion of population members with the characteristic.*

FREQUENCY INTERPRETATION OF PROBABILITY

The **frequency interpretation** of probability provides a link between probability and the real world by relating the probability of an event to a measurable quantity, namely, the long-run relative frequency of occurrence of the event.*

*Some statisticians prefer a different view, namely that the probability of an event is a subjective quantity expressing a person's "degree of belief" that the event will happen. Statistical methods based on this "subjectivist" interpretation are rather different from those presented in this book.

According to the frequency interpretation, the probability of an event E is meaningful only in relation to a chance operation that can in principle be repeated indefinitely often. Each time the chance operation is repeated, the event E either occurs or does not occur. *The probability Pr{E} is interpreted as the relative frequency of occurrence of E in an indefinitely long series of repetitions of the chance operation.*

Specifically, suppose that the chance operation is repeated a large number of times, and that for each repetition the occurrence or nonoccurrence of E is noted. Then we may write

$$\Pr\{E\} \leftrightarrow \frac{\#\text{ of times } E \text{ occurs}}{\#\text{ of times chance operation is repeated}}$$

The arrow in the preceding expression indicates "equality in the long run"; that is, if the chance operation is repeated an unlimited number of times, the two sides of the expression will be approximately equal. Here is a simple example.

Example 3.2.4 **Coin Tossing** Consider again the chance operation of tossing a coin, and the event

$$E\text{: Heads}$$

If the coin is fair, then

$$\Pr\{E\} = 0.5 \leftrightarrow \frac{\#\text{ of heads}}{\#\text{ of tosses}}$$

The arrow in the preceding expression indicates that, in an infinitely long series of tosses of a fair coin, we expect to get heads about 50% of the time. ■

The following two examples illustrate the relative frequency interpretation for more complex events.

Example 3.2.5 **Coin Tossing** Suppose that a fair coin is tossed twice. For reasons that will be explained later in this section, the probability of getting heads both times is 0.25. This probability has the following relative frequency interpretation.

Chance operation: Toss a coin twice

$$E\text{: Both tosses are heads}$$

$$\Pr\{E\} = 0.25 \leftrightarrow \frac{\#\text{ of times both tosses are heads}}{\#\text{ of pairs of tosses}}$$ ■

Example 3.2.6 **Sampling Fruitflies** In the *Drosophila* population of Example 3.2.3, 30% of the flies are black and 70% are gray. Suppose that two flies are randomly chosen from the population. We will see later in this section that the probability that both flies are the same color is 0.58. This probability can be interpreted as follows:

Chance operation: Choose a random sample of size $n = 2$

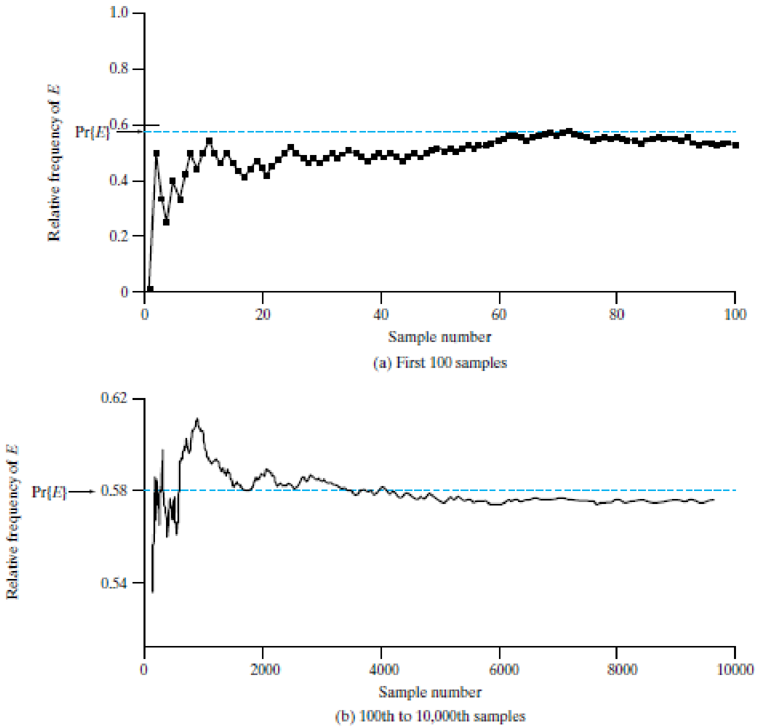
$$E\text{: Both flies in the sample are the same color}$$

$$\Pr\{E\} = 0.58 \leftrightarrow \frac{\#\text{ of times both flies are same color}}{\#\text{ of times a sample of } n = 2 \text{ is chosen}}$$

We can relate this interpretation to a concrete sampling experiment. Suppose that the *Drosophila* population is in a very large container, and that we have some mechanism for choosing a fly at random from the container. We choose one fly at random, and then another; these two constitute the first sample of $n = 2$. After recording their colors, we put the two flies back into the container, and we are ready to repeat the sampling operation once again. Such a sampling experiment would be tedious to carry out physically, but it can readily be simulated using a computer. Table 3.2.1 shows a partial record of the results of choosing 10,000 random samples of size $n = 2$ from a simulated *Drosophila* population. After each repetition of the chance operation (i.e., after each sample of $n = 2$), the cumulative relative frequency of occurrence of the event E was updated, as shown in the rightmost column of the table.

Table 3.2.1 Partial results of simulated sampling from a *Drosophila* population

Sample number	Color: 1st fly	Color: 2nd fly	Did E occur?	Relative frequency of E (cumulative)
1	G	B	No	0.000
2	B	B	Yes	0.500
3	B	G	No	0.333
4	G	B	No	0.250
5	G	G	Yes	0.400
6	G	B	No	0.333
7	B	B	Yes	0.429
8	G	G	Yes	0.500
9	G	B	No	0.444
10	B	B	Yes	0.500
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
20	G	B	No	0.450
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
100	G	B	No	0.540
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1,000	G	G	Yes	0.596
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10,000	B	B	Yes	0.577



(a) First 100 samples

(b) 100th to 10,000th samples

Figure 3.2.1 Results of sampling from fruitfly population. Note that the axes are scaled differently in (a) and (b).

Figure 3.2.1 shows the cumulative relative frequency plotted against the number of samples. Notice that, as the number of samples becomes large, the relative frequency of occurrence of E approaches 0.58 (which is Pr{E}). In other words, the percentage of color-homogeneous samples among all the samples approaches 58% as the number of samples increases. It should be emphasized, however, that the *absolute* number of color-homogeneous samples generally does *not* tend to get closer to 58% of the total number. For instance, if we compare the results shown in Table 3.2.1 for the first 100 samples and the first 1,000 samples, we find the following:

	Color-Homogeneous			**Deviation from 58% of Total**		
First 100 samples:	54	or	54.0%	–4	or	–4.0%
First 1,000 samples:	596	or	59.6%	+16	or	+1.6%

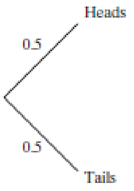
Note that the deviation from 58% is larger in absolute terms, but smaller in relative terms (i.e., in percentage terms), for 1,000 samples than for 100 samples. Likewise, for 10,000 samples the deviation from 58% is rather larger (a deviation of –30), but the percentage deviation is quite small (30/10,000 is 0.3%). The deficit of 4 color-homogeneous samples among the first 100 samples is not *canceled* by a corresponding excess in later samples but rather is *swamped*, or overwhelmed, by a larger denominator. ■

PROBABILITY TREES

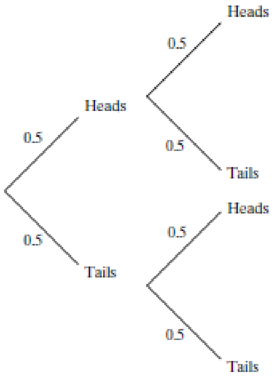
Often it is helpful to use a **probability tree** to analyze a probability problem. A probability tree provides a convenient way to break a problem into parts and to organize the information available. The following examples show some applications of this idea.

Example 3.2.7

Coin Tossing If a fair coin is tossed twice, then the probability of heads is 0.5 on each toss. The first part of a probability tree for this scenario shows that there are two possible outcomes for the first toss and that they have probability 0.5 each.



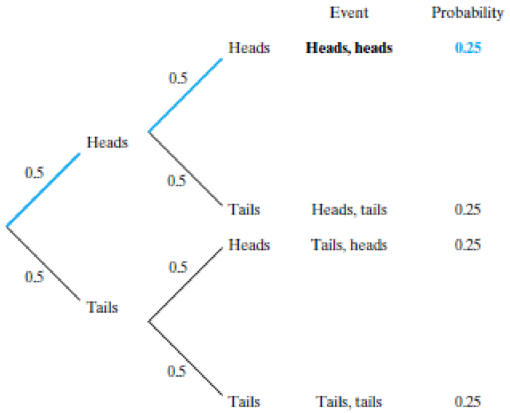
Then the tree shows that, for either outcome of the first toss, the second toss can be either heads or tails, again with probabilities 0.5 each.



To find the probability of getting heads on both tosses, we consider the path through the tree that produces this event. We multiply together the probabilities that we encounter along the path. Figure 3.2.2 summarizes this example and shows that

$$\Pr\{\text{heads on both tosses}\} = 0.5 \times 0.5 = 0.25$$ ■

Figure 3.2.2 Probability tree for two coin tosses



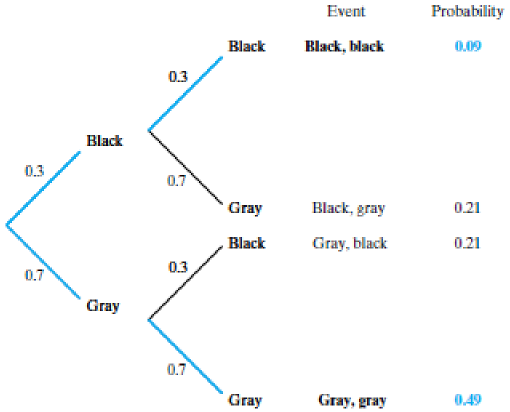
COMBINATION OF PROBABILITIES

If an event can happen in more than one way, the relative frequency interpretation of probability can be a guide to appropriate combinations of the probabilities of subevents. The following example illustrates this idea.

Example 3.2.8

Sampling Fruitflies In the *Drosophila* population of Examples 3.2.3 and 3.2.6, 30% of the flies are black and 70% are gray. Suppose that two flies are randomly chosen from the population. Suppose we wish to find the probability that both flies are the same color. The probability tree displayed in Figure 3.2.3 shows the four possible outcomes from sampling two flies. From the tree, we can see that the probability of getting two black flies is $0.3 \times 0.3 = 0.09$. Likewise, the probability of getting two gray flies is $0.7 \times 0.7 = 0.49$.

Figure 3.2.3 Probability tree for sampling two flies



To find the probability of the event

E: Both flies in the sample are the same color

we add the probability of black, black to the probability of gray, gray to get $0.09 + 0.49 = 0.58$. ■

In the coin tossing setting of Example 3.2.7, the second part of the probability tree had the same structure as the first part—namely, a 0.5 chance of heads and a 0.5 chance of tails—because the outcome of the first toss does not affect the probability of heads on the second toss. Likewise, in Example 3.2.8 the probability of the second fly being black was 0.3, regardless of the color of the first fly, because the population was assumed to be very large, so that removing one fly from the population would not affect the proportion of flies that are black. However, in some situations we need to treat the second part of the probability tree differently than the first part.

Example 3.2.9

Nitric Oxide Hypoxic respiratory failure is a serious condition that affects some newborns. If a newborn has this condition, it is often necessary to use extracorporeal membrane oxygenation (ECMO) to save the life of the child. However, ECMO is an invasive procedure that involves inserting a tube into a vein or artery near the heart, so physicians hope to avoid the need for it. One treatment for hypoxic respiratory failure is to have the newborn inhale nitric oxide. To test the effectiveness of this treatment, newborns suffering hypoxic respiratory failure were assigned at random to either be given nitric oxide or a control group.[1] In the treatment group 45.6% of the newborns had a negative outcome, meaning that either they needed ECMO or that they died. In the control group, 63.6% of the newborns had a negative outcome. Figure 3.2.4 shows a probability tree for this experiment.

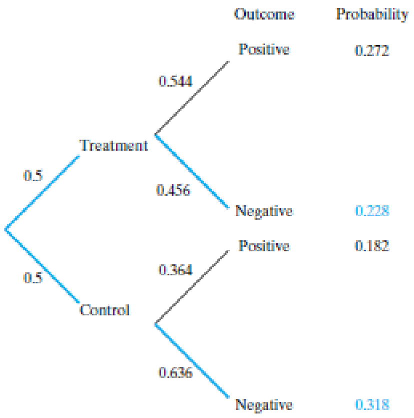


Figure 3.2.4 Probability tree for nitric oxide example

If we choose a newborn at random from this group, there is a 0.5 probability that the newborn will be in the treatment group and, if so, a probability of 0.456 of getting a negative outcome. Likewise, there is a 0.5 probability that the newborn will be in the control group and, if so, a probability of 0.636 of getting a negative outcome. Thus, the probability of a negative outcome is

$$0.5 \times 0.456 + 0.5 \times 0.636 = 0.228 + 0.318 = 0.546.$$ ■

HYPOTHETICAL 1,000

It is often helpful to think about a probability question in terms of what we would expect to see in 1,000 repetitions of the situation. The following example illustrates this mode of thinking.

Example 3.2.10

Nitric Oxide Suppose that 1,000 infants experiencing hypoxic respiratory failure were enrolled in a study like the one described in Example 3.2.9. We would expect 500 of them to be given the treatment (nitric oxide) and 500 of them to be in the control group. Based on Figure 3.2.4, of those given the treatment we expect 54.4% to have a positive outcome. Thus, we expect $500 \times 0.544 = 272$ positive outcomes in the treatment group; likewise, we expect $500 \times 0.456 = 228$ negative outcomes in the treatment group. For the control group, the corresponding numbers are $500 \times 0.364 = 182$ positive outcomes and $500 \times 0.636 = 318$ negative outcomes.

We can put these numbers together to get Table 3.2.2. From the table we can see that there are 546 negative outcomes out of the 1,000 total cases. This agrees with the probability of 0.546 found in Example 3.2.9. ■

Table 3.2.2 Nitric oxide outcomes

	Outcome		
	Positive	Negative	Total
Treatment	272	228	500
Control	182	318	500
Total	454	546	1,000

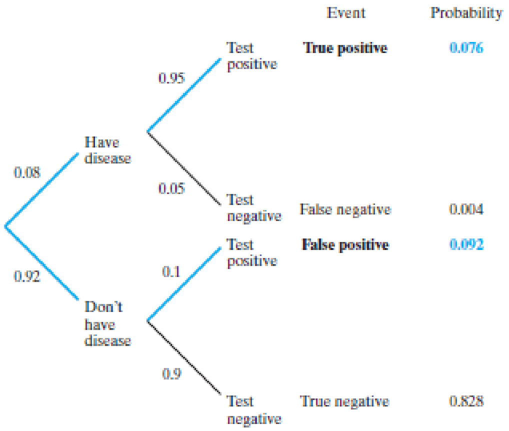
Example 3.2.11

Medical Testing Suppose a medical test is conducted on someone to try to determine whether or not the person has a particular disease. If the test indicates that the disease is present, we say the person has "tested positive." If the test indicates that the disease is not present, we say the person has "tested negative." However, there are two types of mistakes that can be made. It is possible that the test indicates that the disease is present, but the person does not really have the disease; this is known as a false positive. It is also possible that the person has the disease, but the test does not detect it; this is known as a false negative.

Suppose that a particular test has a 95% chance of detecting the disease if the person has it (this is called the sensitivity of the test) and a 90% chance of correctly indicating that the disease is absent if the person really does not have the disease (this is called the specificity of the test). Suppose 8% of the population has the disease. What is the probability that a randomly chosen person will test positive?

Figure 3.2.5 shows a probability tree for this situation. The first split in the tree shows the division between those who have the disease and those who don't. If someone has the disease, then we use 0.95 as the chance of the person testing positive. If the person doesn't have the disease, then we use 0.10 as the chance of the

Figure 3.2.5 Probability tree for medical testing example



person testing positive. Thus, the probability of a randomly chosen person testing positive is

$$0.08 \times 0.95 + 0.92 \times 0.10 = 0.076 + 0.092 = 0.168.$$ ■

We can apply the Hypothetical 1,000 idea to Example 3.2.11. Table 3.2.3 shows that we expect 80 out of 1,000 people to have the disease ($0.08 \times 1{,}000$) and 76 of them to test positive (0.95×80). We expect 920 people to not have the disease and 92 of them (0.1×920) to test positive. From the table we see that 168 out of 1,000 people test positive; this agrees with the probability of 0.168 found in Example 3.2.11.

Table 3.2.3 Medical testing outcomes

	Test result		
	Positive	Negative	Total
Disease	76	4	80
No disease	92	828	920
Total	168	832	1,000

Example 3.2.12 **False Positives** Consider the medical testing scenario of Example 3.2.11. If someone tests positive, what is the chance the person really has the disease? Table 3.2.3 shows that we would expect 168 out of 1,000 to test positive. The probability of a true positive is 0.076, so we would expect 76 "true positives" out of 1,000 persons tested. Thus, we expect 76 true positives out of 168 total positives, which is to say that the probability that someone really has the disease, given that the person tests positive, is $\frac{76}{168} = \frac{0.076}{0.168} \approx 0.452$. This probability is quite a bit smaller than most people expect it to be, given that the sensitivity and specificity of the test are 0.95 and 0.90. We also see that out of the 168 positive test results only 76 are true positives, giving a rate of $76/168 = 0.452$. ■

Exercises 3.2.1–3.2.8

3.2.1 In a certain population of the freshwater sculpin, *Cottus rotheus*, the distribution of the number of tail vertebrae is as shown in the table.[2]

No. of vertebrae	Percent of fish
20	3
21	51
22	40
23	6
Total	100

Find the probability that the number of tail vertebrae in a fish randomly chosen from the population

(a) equals 21.

(b) is less than or equal to 22.

(c) is greater than 21.

(d) is no more than 21.

3.2.2 The following table shows the distribution of ages of Americans.[3]

Age distribution in reference population

Age	Proportion
0–19	0.27
20–29	0.14
30–39	0.13
40–49	0.14
50–64	0.19
65+	0.13

Find the probability that the age of a randomly chosen American

(a) is less than 20.

(b) is between 20 and 49.

(c) is greater than 49.

(d) is greater than 29.

3.2.3 In a certain college, 55% of the students are women. Suppose we take a sample of two students. Use a probability tree to find the probability

(a) that both chosen students are women.

(b) that at least one of the two students is a woman.

3.2.4 Suppose that a disease is inherited via a sex-linked mode of inheritance so that a male offspring has a 50% chance of inheriting the disease, but a female offspring has no chance of inheriting the disease. Further suppose that 51.3% of births are male. What is the probability that a randomly chosen child will be affected by the disease?

3.2.5 Suppose that a student who is about to take a multiple choice test has only learned 40% of the material covered by the exam. Thus, there is a 40% chance that she will know the answer to a question. However, even if she does not know the answer to a question, she still has a 20% chance of getting the right answer by guessing. If we choose a question at random from the exam, what is the probability that she will get it right?

3.2.6 If a woman takes an early pregnancy test, she will either test positive, meaning that the test says she is pregnant, or test negative, meaning that the test says she is not pregnant. Suppose that if a woman really is pregnant, there is a 98% chance that she will test positive. Also, suppose that if a woman really is *not* pregnant, there is a 99% chance that she will test negative.

(a) Suppose that 1,000 women take early pregnancy tests and that 100 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive?

(b) Suppose that 1,000 women take early pregnancy tests and that 50 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive?

3.2.7

(a) Consider the setting of Exercise 3.2.6, part (a). Suppose that a woman tests positive. What is the probability that she really is pregnant?

(b) Consider the setting of Exercise 3.2.6, part (b). Suppose that a woman tests positive. What is the probability that she really is pregnant?

3.2.8 Suppose that a medical test has a 92% chance of detecting a disease if the person has it (i.e., 92% sensitivity) and a 94% chance of correctly indicating that the disease is absent if the person really does not have the disease (i.e., 94% specificity). Suppose 10% of the population has the disease.

(a) What is the probability that a randomly chosen person will test positive?

(b) Suppose that a randomly chosen person does test positive. What is the probability that this person really has the disease?

3.3 Probability Rules (Optional)

We have defined the probability of an event, $\Pr\{E\}$, as the long-run relative frequency with which the event occurs. In this section we will briefly consider a few rules that help determine probabilities. We begin with three basic rules.

BASIC RULES

Rule (1) The probability of an event E is always between 0 and 1. That is, $0 \leq \Pr\{E\} \leq 1$.

Rule (2) The sum of the probabilities of all possible events equals 1. That is, if the set of all possible events is $E_1, E_2, \ldots, E_k$, then $\sum_{i=1}^{k} \Pr\{E_i\} = 1$.

Rule (3) The probability that an event E does not happen, denoted by E^C, is one minus the probability that the event happens. That is, $\Pr\{E^C\} = 1 - \Pr\{E\}$. (We refer to E^C as the complement of E.)

We illustrate these rules with an example.

Example 3.3.1 **Blood Type** In the United States, 44% of the population has type O blood, 42% has type A, 10% has type B, and 4% has type AB.[4] Consider choosing someone at random and determining the person's blood type. The probability of a given blood type will correspond to the population percentage.

(a) The probability that the person will have type O blood $= \Pr\{\text{O}\} = 0.44$.

(b) $\Pr\{\text{O}\} + \Pr\{\text{A}\} + \Pr\{\text{B}\} + \Pr\{\text{AB}\} = 0.44 + 0.42 + 0.10 + 0.04 = 1$.

(c) The probability that the person will *not* have type O blood $= \Pr\{\text{O}^C\} = 1 - 0.44 = 0.56$. This could also be found by adding the probabilities of the other blood types: $\Pr\{\text{O}^C\} = \Pr\{\text{A}\} + \Pr\{\text{B}\} + \Pr\{\text{AB}\} = 0.42 + 0.10 + 0.04 = 0.56$.

We often want to discuss two or more events at once; to do this we will find some terminology to be helpful. We say that two events are *disjoint** if they cannot occur simultaneously. Figure 3.3.1 is a *Venn diagram* that depicts a *sample space* S of all possible outcomes as a rectangle with two disjoint events depicted as nonoverlapping regions.

The *union* of two events is the event that one or the other occurs or both occur. The *intersection* of two events is the event that they both occur. Figure 3.3.2 is a Venn diagram that shows the union of two events as the total shaded area, with the intersection of the events being the overlapping region in the middle.

If two events are disjoint, then the probability of their union is the sum of their individual probabilities. If the events are not disjoint, then to find the probability of their union we take the sum of their individual probabilities and subtract the probability of their intersection (the part that was "counted twice").

ADDITION RULES

Rule (4) If two events E_1 and E_2 are disjoint, then
$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$.

*Another term for disjoint events is "mutually exclusive" events.

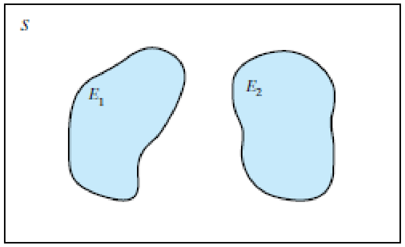


Figure 3.3.1 Venn diagram showing two disjoint events

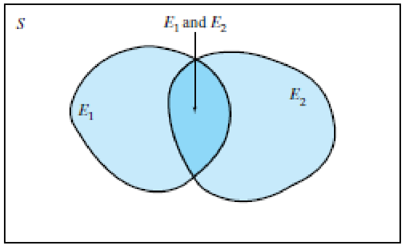


Figure 3.3.2 Venn diagram showing union (total shaded area) and intersection (middle area) of two events

Rule (5) For any two events E_1 and E_2,
$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \text{ and } E_2\}$.

We illustrate these rules with an example.

Example 3.3.2 **Hair Color and Eye Color** Table 3.3.1 shows the relationship between hair color and eye color for a group of 1,770 German men.[5]

Table 3.3.1 Hair color and eye color

		Hair color			
		Brown	Black	Red	Total
Eye color	Brown	400	300	20	720
	Blue	800	200	50	1,050
	Total	1,200	500	70	1,770

(a) Because events "black hair" and "red hair" are disjoint, if we choose someone at random from this group then Pr{black hair or red hair} = Pr{black hair} + Pr{red hair} = 500/1,770 + 70/1,770 = 570/1,770.

(b) If we choose someone at random from this group, then Pr{black hair} = 500/1,770.

(c) If we choose someone at random from this group, then Pr{blue eyes} = 1,050/1,770.

(d) The events "black hair" and "blue eyes" are not disjoint, since there are 200 men with both black hair and blue eyes. Thus, Pr{black hair or blue eyes} = Pr{black hair} + Pr{blue eyes} − Pr{black hair and blue eyes} = 500/1,770 + 1,050/1,770 − 200/1,770 = 1,350/1,770.

Two events are said to be *independent* if knowing that one of them occurred does not change the probability of the other one occurring. For example, if a coin is tossed twice, the outcome of the second toss is independent of the outcome of the first toss, since knowing whether the first toss resulted in heads or in tails does not change the probability of getting heads on the second toss.

Events that are not independent are said to be *dependent*. When events are dependent, we need to consider the *conditional probability* of one event, given that the other event has happened. We use the notation

$$\Pr\{E_2 | E_1\}$$

to represent the probability of E_2 happening, given that E_1 happened.

Example 3.3.3 **Hair Color and Eye Color** Consider choosing a man at random from the group shown in Table 3.3.1. Overall, the probability of blue eyes is 1,050/1,770, or about 59.3%. However, if the man has black hair, then the conditional probability of blue eyes is only 200/500, or 40%; that is, Pr{blue eyes|black hair} = 0.40. Because the probability of blue eyes depends on hair color, the events "black hair" and "blue eyes" are dependent. ■

Refer again to Figure 3.3.2, which shows the intersection of two regions (for E_1 and E_2). If we know that the event E_1 has happened, then we can restrict our attention to the E_1 region in the Venn diagram. If we now want to find the chance that E_2 will happen, we need to consider the intersection of E_1 and E_2 relative to the entire E_1 region. In the case of Example 3.3.3, this corresponds to knowing that a randomly chosen man has black hair, so that we restrict our attention to the 500 men (out of 1,770 total in the group) with black hair. Of these men, 200 have blue eyes. The 200 are in the intersection of "black hair" and "blue eyes." The fraction 200/500 is the conditional probability of having blue eyes, given that the man has black hair.

This leads to the following formal definition of the conditional probability of E_2 given E_1:

DEFINITION The conditional probability of E_2, given E_1, is

$$\Pr\{E_2 | E_1\} = \frac{\Pr\{E_1 \text{ and } E_2\}}{\Pr\{E_1\}}$$

provided that $\Pr\{E_1\} > 0$.

Example 3.3.4 **Hair Color and Eye Color** Consider choosing a man at random from the group shown in Table 3.3.1. The probability of the man having blue eyes given that he has black hair is

$$\begin{aligned}\Pr\{\text{blue eyes} | \text{black hair}\} &= \Pr\{\text{black hair and blue eyes}\}/\Pr\{\text{black hair}\}\\ &= \frac{200/1{,}770}{500/1{,}770} = \frac{200}{500} = 0.40.\end{aligned}$$ ■

In Section 3.2 we used probability trees to study compound events. In doing so, we implicitly used multiplication rules that we now make explicit.

MULTIPLICATION RULES

Rule (6) If two events E_1 and E_2 are independent, then
$\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2\}$.

Rule (7) For any two events E_1 and E_2, $\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2|E_1\}$.

Example 3.3.5 **Coin Tossing** If a fair coin is tossed twice, the two tosses are independent of each other. Thus, the probability of getting heads on both tosses is

$$\begin{aligned}\Pr\{\text{heads twice}\} &= \Pr\{\text{heads on first toss}\} \times \Pr\{\text{heads on second toss}\}\\ &= 0.5 \times 0.5 = 0.25.\end{aligned}$$ ■

Example 3.3.6 **Blood Type** In Example 3.3.1 we stated that 44% of the U.S. population has type O blood. It is also true that 15% of the population is Rh negative and that this is independent of blood group. Thus, if someone is chosen at random, the probability that the person has type O, Rh negative blood is

$$\begin{aligned}\Pr\{\text{group O and Rh negative}\} &= \Pr\{\text{group O}\} \times \Pr\{\text{Rh negative}\}\\ &= 0.44 \times 0.15 = 0.066.\end{aligned}$$ ■

Example 3.3.7 **Hair Color and Eye Color** Consider choosing a man at random from the group shown in Table 3.3.1. What is the probability that the man will have red hair and brown eyes? Hair color and eye color are dependent, so finding this probability involves using a conditional probability. The probability that the man will have red hair is 70/1,770. Given that the man has red hair, the conditional probability of brown eyes is 20/70. Thus,

$$\begin{aligned}\Pr\{\text{red hair and brown eyes}\} &= \Pr\{\text{red hair}\} \times \Pr\{\text{brown eyes} | \text{red hair}\}\\ &= 70/1{,}770 \times 20/70 = 20/1{,}770.\end{aligned}$$ ■

Sometimes a probability problem can be broken into two conditional "parts" that are solved separately and the answers combined.

Rule of Total Probability

Rule (8) For any two events E_1 and E_2,

$$\Pr\{E_1\} = \Pr\{E_2\} \times \Pr\{E_1 | E_2\} + \Pr\{E_2^C\} \times \Pr\{E_1 | E_2^C\}.$$

Example 3.3.8 **Hand Size** Consider choosing someone at random from a population that is 60% female and 40% male. Suppose that for a woman the probability of having a hand size smaller than 100 cm^2 is 0.31.[6] Suppose that for a man the probability of having a hand size smaller than 100 cm^2 is 0.08. What is the probability that the randomly chosen person will have a hand size smaller than 100 cm^2?

We are given that if the person is a woman, then the probability of a "small" hand size is 0.31 and that if the person is a man, then the probability of a "small" hand size is 0.08.

Thus,

$$\begin{aligned}\Pr\{\text{hand size} < 100\} &= \Pr\{\text{woman}\} \times \Pr\{\text{hand size} < 100 | \text{woman}\}\\ &\quad + \Pr\{\text{man}\} \times \Pr\{\text{hand size} < 100 | \text{man}\}\\ &= 0.6 \times 0.31 + 0.4 \times 0.08\\ &= 0.186 + 0.032\\ &= 0.218.\end{aligned}$$ ■

We can apply the Hypothetical 1,000 idea here. Table 3.3.2 shows 600 women, of whom 31% have small hands (0.31 × 600 = 186) and 69% don't. We also see

Table 3.3.2 Hand size

	Hand size		
	<100 cm^2	≥ 100 cm^2	Total
Woman	186	414	600
Man	32	368	400
Total	218	782	1,000

400 men, of whom 8% have small hands ($0.08 \times 400 = 32$) and 92% don't. The column "<100 cm^2" sums to 218; this agrees with the probability of 0.218.

Exercises 3.3.1–3.3.5

3.3.1 In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions.[7] Some of the results are shown in the following table.

	Income			
	Low	Medium	High	Total
Smoke	634	332	247	1,213
Don't smoke	1,846	1,622	1,868	5,336
Total	2,480	1,954	2,115	6,549

(a) What is the probability that someone in this study smokes?

(b) What is the conditional probability that someone in this study smokes, given that the person has high income?

(c) Is being a smoker independent of having a high income? Why or why not?

3.3.2 Consider the data table reported in Exercise 3.3.1.

(a) What is the probability that someone in this study is from the low income group and smokes?

(b) What is the probability that someone in this study is not from the low income group?

(c) What is the probability that someone in this study is from the medium income group?

(d) What is the probability that someone in this study is from the low income group or from the medium income group?

3.3.3 The following data table is taken from the study reported in Exercise 3.3.1. Here "stressed" means that the person reported that most days are extremely stressful or quite stressful; "not stressed" means that the person reported that most days are a bit stressful, not very stressful, or not at all stressful.

	Income			
	Low	Medium	High	Total
Stressed	526	274	216	1,016
Not stressed	1,954	1,680	1,899	5,533
Total	2,480	1,954	2,115	6,549

(a) What is the probability that someone in this study is stressed?

(b) Given that someone in this study is from the high income group, what is the probability that the person is stressed?

(c) Compare your answers to parts (a) and (b). Is being stressed independent of having high income? Why or why not?

3.3.4 Consider the data table reported in Exercise 3.3.3.

(a) What is the probability that someone in this study has low income?

(b) What is the probability that someone in this study either is stressed or has low income (or both)?

(c) What is the probability that someone in this study is stressed and has low income?

3.3.5 Suppose that in a certain population of married couples, 30% of the husbands smoke, 20% of the wives smoke, and in 8% of the couples both the husband and the wife smoke. Is the smoking status (smoker or nonsmoker) of the husband independent of that of the wife? Why or why not?

3.4 Density Curves

The examples presented in Section 3.2 dealt with probabilities for discrete variables. In this section we will consider probability when the variable is continuous.

RELATIVE FREQUENCY HISTOGRAMS AND DENSITY CURVES

In Chapter 2 we discussed the use of a histogram to represent a frequency distribution for a variable. A *relative frequency histogram* is a histogram in which we indicate the proportion (i.e., the relative frequency) of observations in each category, rather than the count of observations in the category. We can think of the relative frequency histogram as an approximation of the underlying true population distribution from which the data came.

It is often desirable, especially when the observed variable is continuous, to describe a population frequency distribution by a smooth curve. We may visualize the curve as an idealization of a relative frequency histogram with very narrow classes. The following example illustrates this idea.

Example 3.4.1 **Blood Glucose** A glucose tolerance test can be useful in diagnosing diabetes. The blood level of glucose is measured one hour after the subject has drunk 50 mg of glucose dissolved in water. Figure 3.4.1 shows the distribution of responses to this test for a certain population of women.[8] The distribution is represented by histograms with class widths equal to (a) 10 and (b) 5, and by (c) a smooth curve. ■

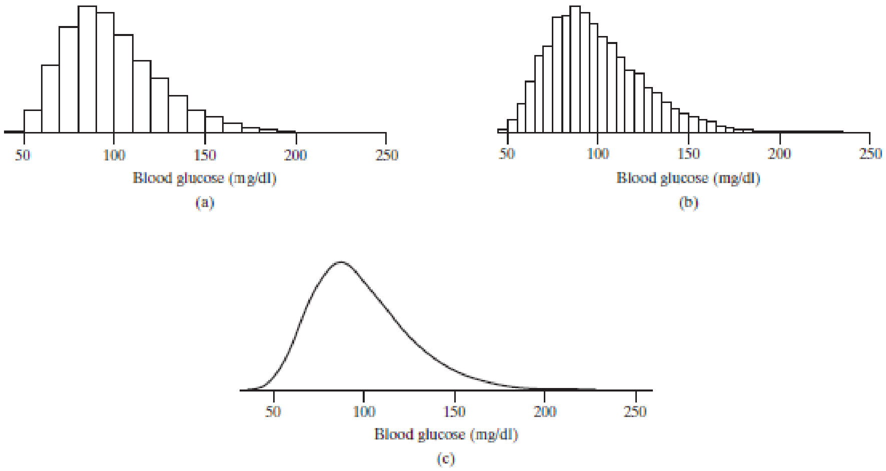


Figure 3.4.1 Different representations of the distribution of blood glucose levels in a population of women

A smooth curve representing a frequency distribution is called a **density curve**. The vertical coordinates of a density curve are plotted on a scale called a **density scale**. When the density scale is used, relative frequencies are represented as areas under the curve. Formally, the relation is as follows:

Interpretation of Density

For any two numbers a and b,

$$\begin{matrix}\text{Area under density curve} \\ \text{between } a \text{ and } b\end{matrix} = \begin{matrix}\text{Proportion of } Y \text{ values} \\ \text{between } a \text{ and } b\end{matrix}$$

This relation is indicated in Figure 3.4.2 for an arbitrary distribution

Because of the way the density curve is interpreted, the density curve is entirely above (or equal to) the x-axis and the area under the entire curve must be equal to 1, as shown in Figure 3.4.3.

The interpretation of density curves in terms of areas is illustrated concretely in the following example.

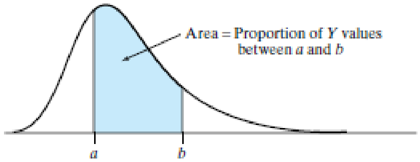


Figure 3.4.2 Interpretation of area under a density curve

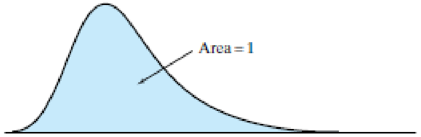
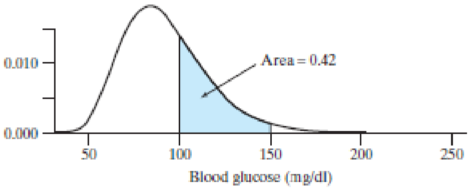


Figure 3.4.3 The area under an entire density curve must be 1

Example 3.4.2

Blood Glucose Figure 3.4.4 shows the density curve for the blood glucose distribution of Example 3.4.1, with the vertical scale explicitly shown. The shaded area is equal to 0.42, which indicates that about 42% of the glucose levels are between 100 mg/dl and 150 mg/dl. The area under the density curve to the left of 100 mg/dl is equal to 0.50; this indicates that the population median glucose level is 100 mg/dl. The area under the entire curve is 1.

Figure 3.4.4 Interpretation of an area under the blood glucose density curve



The Continuum Paradox The area interpretation of a density curve has a paradoxical element. If we ask for the relative frequency of a single specific Y value, the answer is zero. For example, suppose we want to determine from Figure 3.4.4 the relative frequency of blood glucose levels *equal* to 150. The area interpretation gives an answer of zero. This seems to be nonsense—how can every value of Y have a relative frequency of zero? Let us look more closely at the question. If blood glucose is measured to the nearest mg/dl, then we are really asking for the relative frequency of glucose levels between 149.5 and 150.5 mg/dl, and the corresponding area is not zero. On the other hand, if we are thinking of blood glucose as an *idealized* continuous variable, then the relative frequency of any particular value (e.g., 150) *is* zero. This is admittedly a paradoxical situation. It is similar to the paradoxical fact that an idealized straight line can be 1 centimeter long, and yet each of the idealized points of which the line is composed has length equal to zero. In practice, the continuum paradox does not cause any trouble; we simply do not discuss the relative frequency of a single Y value (just as we do not discuss the length of a single point).

PROBABILITIES AND DENSITY CURVES

If a variable has a continuous distribution, then we find probabilities by using the density curve for the variable. A probability for a continuous variable equals the area under the density curve for the variable between two points.

Example 3.4.3

Blood Glucose Consider the blood glucose level, in mg/dl, of a randomly chosen subject from the population described in Example 3.4.2. We saw in Example 3.4.2 that 42% of the population glucose levels are between 100 mg/dl and 150 mg/dl. Thus, $\Pr\{100 \leq \text{glucose level} \leq 150\} = 0.42$.

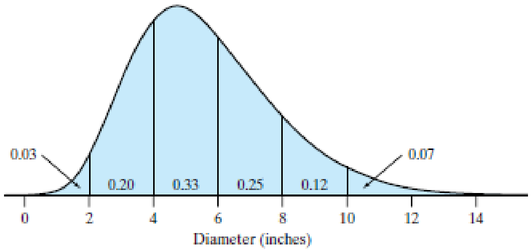
We are modeling blood glucose level as being a continuous variable, which means that $\Pr\{\text{glucose level} = 100\} = 0$ and $\Pr\{\text{glucose level} = 150\} = 0$, as we noted above. Thus,

$$\Pr\{100 \leq \text{glucose level} \leq 150\} = \Pr\{100 < \text{glucose level} < 150\} = 0.42.$$

Example 3.4.4

Tree Diameters The diameter of a tree trunk is an important variable in forestry. The density curve shown in Figure 3.4.5 represents the distribution of diameters (measured at breast height) in a population of 30-year-old Douglas fir trees; areas under the curve are shown in the figure.[9] Consider the diameter, in inches, of a randomly chosen tree. Then, for example, $\Pr\{4 < \text{diameter} < 6\} = 0.33$. If we want to find the probability that a randomly chosen tree has a diameter greater than 8 inches, we must add the last two areas under the curve in Figure 3.4.3: $\Pr\{\text{diameter} > 8\} = 0.12 + 0.07 = 0.19$.

Figure 3.4.5 Diameters of 30-year-old Douglas fir trees



Exercises 3.4.1–3.4.5

3.4.1 Consider the density curve shown in Figure 3.4.5, which represents the distribution of diameters (measured 4.5 feet above the ground) in a population of 30-year-old Douglas fir trees. Areas under the curve are shown in the figure. What percentage of the trees have diameters

(a) between 4 inches and 10 inches?

(b) less than 4 inches?

(c) more than 6 inches?

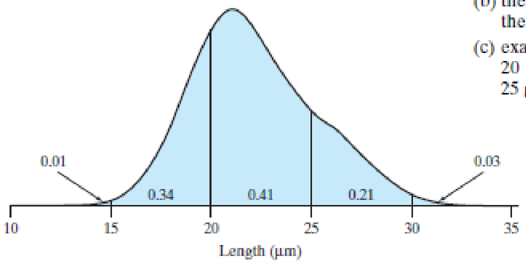
3.4.2 Consider the diameter of a Douglas fir tree drawn at random from the population that is represented by the density curve shown in Figure 3.4.5. Find

(a) $\Pr\{\text{diameter} < 10\}$

(b) $\Pr\{\text{diameter} > 4\}$

(c) $\Pr\{2 < \text{diameter} < 8\}$

3.4.3 In a certain population of the parasite *Trypanosoma*, the lengths of individuals are distributed as indicated by the density curve shown here. Areas under the curve are shown in the figure.[10]



Consider the length of an individual trypanosome chosen at random from the population. Find

(a) $\Pr\{20 < \text{length} < 30\}$

(b) $\Pr\{\text{length} > 20\}$

(c) $\Pr\{\text{length} < 20\}$

3.4.4 Consider the distribution of *Trypanosoma* lengths shown by the density curve in Exercise 3.4.3. Consider the length of an individual trypanosome chosen at random from the population. Find

(a) $\Pr\{\text{length} < 25\}$

(b) $\Pr\{\text{length} > 15\}$

(c) $\Pr\{15 < \text{length} < 30\}$

3.4.5 Consider the distribution of *Trypanosoma* lengths shown by the density curve in Exercise 3.4.3. Suppose we take a sample of two trypanosomes. What is the probability that

(a) both trypanosomes will be shorter than 20 μm?

(b) the first trypanosome will be shorter than 20 μm and the second trypanosome will be longer than 25 μm?

(c) exactly one of the trypanosomes will be shorter than 20 μm and one trypanosome will be longer than 25 μm?

3.5 Random Variables

A **random variable** is simply a variable that takes on numerical values that depend on the outcome of a chance operation. The following examples illustrate this idea.

Example 3.5.1

Dice Consider the chance operation of tossing a die. Let the random variable Y represent the number of spots showing. The possible values of Y are $Y = 1, 2, 3, 4, 5$, or 6. We do not know the value of Y until we have tossed the die. If we know how the die is weighted, then we can specify the probability that Y has a particular value, say $\Pr\{Y = 4\}$, or a particular set of values, say $\Pr\{2 \le Y \le 4\}$. For instance, if the die is perfectly balanced so that each of the six faces is equally likely, then

$$\Pr\{Y = 4\} = \frac{1}{6} \approx 0.17$$

and

$$\Pr\{2 \le Y \le 4\} = \frac{3}{6} = 0.5$$ ■

Example 3.5.2

Family Size Suppose a family is chosen at random from a certain population, and let the random variable Y denote the number of children in the chosen family. The possible values of Y are $0, 1, 2, 3, \ldots$. The probability that Y has a particular value is equal to the percentage of families with that many children. For instance, if 23% of the families have 2 children, then

$$\Pr\{Y = 2\} = 0.23$$ ■

Example 3.5.3

Medications After someone has heart surgery, the person is usually given several medications. Let the random variable Y denote the number of medications that a patient is given following cardiac surgery. If we know the distribution of the number of medications per patient for the entire population, then we can specify the probability that Y has a certain value or falls within a certain interval of values. For instance, if 52% of all patients are given 2, 3, 4, or 5 medications, then

$$\Pr\{2 \le Y \le 5\} = 0.52$$ ■

Example 3.5.4

Heights of Men Let the random variable Y denote the height of a man chosen at random from a certain population. If we know the distribution of heights in the population, then we can specify the probability that Y falls in a certain range. For instance, if 46% of the men are between 65.2 and 70.4 inches tall, then

$$\Pr\{65.2 \le Y \le 70.4\} = 0.46$$ ■

Each of the variables in Examples 3.5.1–3.5.3 is a *discrete random variable*, because in each case we can list the possible values that the variable can take on. In contrast, the variable in Example 3.5.4, height, is a *continuous random variable*: Height, at least in theory, can take on any of an infinite number of values in an interval. Of course, when we measure and record a person's height, we generally measure to the nearest inch or half inch. Nonetheless, we can think of true height as being a continuous variable. We use density curves to model the distributions of continuous random variables, such as blood glucose level or tree diameter, as discussed in Section 3.4.

MEAN AND VARIANCE OF A RANDOM VARIABLE

In Chapter 2 we briefly considered the concepts of population mean and population standard deviation. For the case of a discrete random variable, we can calculate the population mean and standard deviation if we know the probability distribution for the random variable. We begin with the mean.

The mean of a discrete random variable Y is defined as

$$\mu_Y = \Sigma y_i \Pr(Y = y_i)$$

where the y_i's are the values that the variable takes on and the sum is taken over all possible values.

The mean of a random variable is also known as the *expected value* and is often written as $E(Y)$; that is, $E(Y) = \mu_Y$.

Example 3.5.5

Fish Vertebrae In a certain population of the freshwater sculpin *Cottus rotheus*, the distribution of the number of tail vertebrae, Y, is as shown in Table 3.5.1.[2]

Table 3.5.1 Distribution of vertebrae

No. of vertebrae	Percent of fish
20	3
21	51
22	40
23	6
Total	100

The mean of Y is

$$\begin{aligned}\mu_Y &= 20 \times \Pr\{Y = 20\} + 21 \times \Pr\{Y = 21\} + 22 \times \Pr\{Y = 22\} + 23 \times \Pr\{Y = 23\}\\ &= 20 \times .03 \quad + 21 \times .51 \quad + 22 \times .40 \quad + 23 \times .06\\ &= 0.6 \quad + 10.71 \quad + 8.8 \quad + 1.38\\ &= 21.49.\end{aligned}$$ ■

Example 3.5.6 **Dice** Consider rolling a die that is perfectly balanced so that each of the six faces is equally likely to come up and let the random variable Y represent the number of spots showing. The expected value, or mean, of Y is

$$E(Y) = \mu_Y = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6} = 3.5.$$ ■

To find the standard deviation of a random variable, we first find the variance, σ^2, of the random variable and then take the square root of the variance to get the the standard deviation, σ.

The variance of a discrete random variable Y is defined as

$$\sigma_Y^2 = \Sigma(y_i - \mu_Y)^2\Pr(Y = y_i)$$

where the y_i's are the values that the variable takes on and the sum is taken over all possible values.

We often write VAR(Y) to denote the variance of Y.

Example 3.5.7 **Fish Vertebrae** Consider the distribution of vertebrae given in Table 3.5.1. In Example 3.5.5 we found that the mean of Y is $\mu_Y = 21.49$. The variance of Y is

$$\begin{aligned}\mathrm{VAR}(Y) = \sigma_Y^2 &= (20 - 21.49)^2 \times \Pr\{Y = 20\}\\ &\quad + (21 - 21.49)^2 \times \Pr\{Y = 21\}\\ &\quad + (22 - 21.49)^2 \times \Pr\{Y = 22\}\\ &\quad + (23 - 21.49)^2 \times \Pr\{Y = 23\}\\ &= (-1.49)^2 \times 0.03 + (-.49)^2 \times 0.51\\ &\quad + (0.51)^2 \times 0.40 + (1.51)^2 \times 0.06\\ &= 2.2201 \times 0.03 + .2401 \times 0.51 + .2601 \times 0.40 + 2.2801 \times 0.06\\ &= 0.066603 + 0.122451 + 0.10404 + 0.136806\\ &= 0.4299.\end{aligned}$$

The standard deviation of Y is $\sigma_Y = \sqrt{0.4299} \approx 0.6557$. ■

Example 3.5.8 **Dice** In Example 3.5.6 we found that the mean number obtained from rolling a fair die is 3.5 (i.e., $\mu_Y = 3.5$). The variance of the number obtained from rolling a fair die is

$$\begin{aligned}\sigma_Y^2 &= (1 - 3.5)^2 \times \Pr\{Y = 1\} + (2 - 3.5)^2 \times \Pr\{Y = 2\}\\ &\quad + (3 - 3.5)^2 \times \Pr\{Y = 3\} + (4 - 3.5)^2 \times \Pr\{Y = 4\}\\ &\quad + (5 - 3.5)^2 \times \Pr\{Y = 5\} + (6 - 3.5)^2 \times \Pr\{Y = 6\}\\ &= (-2.5)^2 \times \frac{1}{6} + (-1.5)^2 \times \frac{1}{6} + (-0.5)^2 \times \frac{1}{6} + (0.5)^2 \times \frac{1}{6}\\ &\quad + (1.5)^2 \times \frac{1}{6} + (2.5)^2 \times \frac{1}{6}\\ &= (6.25) \times \frac{1}{6} + (2.25) \times \frac{1}{6} + (0.25) \times \frac{1}{6} + (0.25) \times \frac{1}{6}\\ &\quad + (2.25) \times \frac{1}{6} + (6.25) \times \frac{1}{6}\\ &= 17.5 \times \frac{1}{6}\\ &\approx 2.9167.\end{aligned}$$

The standard deviation of Y is $\sigma_Y = \sqrt{2.9167} \approx 1.708$. ■

The preceding definitions are appropriate for discrete random variables. There are analogous definitions for continuous random variables, but they involve integral calculus and won't be presented here.

ADDING AND SUBTRACTING RANDOM VARIABLES (OPTIONAL)

If we add two random variables, it makes sense that we add their means. Likewise, if we create a new random variable by subtracting two random variables, then we subtract the individual means to get the mean of the new random variable. If we multiply a random variable by a constant (e.g., if we are converting feet to inches so that we are multiplying by 12), then we multiply the mean of the random variable by the same constant. If we add a constant to a random variable, then we add that constant to the mean.

The following rules summarize the situation:

Rules for Means of Random Variables

Rule (1) If X and Y are two random variables, then $\mu_{X+Y} = \mu_X + \mu_Y$.

$$\mu_{X-Y} = \mu_X - \mu_Y$$

Rule (2) If Y is a random variable and a and b constants, then $\mu_{a+bY} = a + b\mu_Y$.

Example 3.5.9 **Temperature** The average summer temperature, μ_Y, in a city is 81°F. To convert °F to °C, we use the formula °C = (°F − 32) × (5/9) or °C = (5/9) × °F − (5/9) × 32. Thus, the mean in degrees Celsius is (5/9) × (81) − (5/9) × 32 = 45 − 17.78 = 27.22. ■

Dealing with standard deviations of functions of random variables is a bit more complicated. We work with the variance first and then take the square root, at the end, to get the standard deviation we want. If we *multiply* a random variable by a constant (e.g., if we are converting inches to centimeters by multiplying by 2.54), then we multiply the variance by the square of the constant. This has the effect of multiplying the standard deviation by the absolute value of the constant. If we *add* a constant to a random variable, then we are not changing the relative spread of the distribution, so the variance does not change.

Example 3.5.10

Feet to Inches Let Y denote the height, in feet, of a person in a given population; suppose the standard deviation of Y is $\sigma_Y = 0.35$ (feet). If we wish to convert from feet to inches, we can define a new variable X as $X = 12Y$. The variance of Y is 0.35^2 (the square of the standard deviation). The variance of X is $12^2 \times 0.35^2$, which means that the standard deviation of X is $\sigma_X = 12 \times 0.35 = 4.2$ (inches). ■

If we add two random variables *that are independent of one another*, then we add their variances.* Moreover, if we subtract two random variables *that are independent of one another*, then we *add* their variances. If we want to find the standard deviation of the sum (or difference) of two independent random variables, we first find the variance of the sum (or difference) and then take its square root.

Example 3.5.11

Mass Consider finding the mass of a 10-ml graduated cylinder. If several measurements are made, using an analytical balance, then in theory we would expect the measurements to all be the same. In reality, however, the readings will vary from one measurement to the next. Suppose that a given balance produces readings that have a standard deviation of 0.03g; let X denote the value of a reading made using this balance. Suppose that a second balance produces readings that have a standard deviation of 0.04g; let Y denote denote the value of a reading made using this second balance.[11]

If we use each balance to measure the mass of a graduated cylinder, we might be interested in the difference, $X - Y$, of the two measurements. The standard deviation of $X - Y$ is positive. To find the standard deviation of $X - Y$, we first find the variance of the difference. The variance of X is 0.03^2 and the variance of Y is 0.04^2. The variance of the difference is $0.03^2 + 0.04^2 = 0.0025$. The standard deviation of $X - Y$ is the square root of 0.0025, which is 0.05. ■

The following rules summarize the situation for variances:

Rules for Variances of Random Variables

Rule (3) If Y is a random variable and a and b constants, then $\sigma^2_{a+bY} = b^2\sigma^2_Y$.

Rule (4) If X and Y are two *independent* random variables, then

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

*If we add two random variables that are not independent of one another, then the variance of the sum depends on the degree of dependence between the variables. To take an extreme case, suppose that one of the random variables is the negative of the other. Then the sum of the two random variables will always be zero, so the variance of the sum will be zero. This is quite different from what we would get by adding the two variances together. As another example, suppose Y is the number of questions correct on a 20-question exam and X is the number of questions wrong. Then $Y + X$ is always equal to 20, so there is no variability at all. Hence, the variance of $Y + X$ is zero, even though the variance of Y is positive, as is the variance of X.

Exercises 3.5.1–3.5.10

3.5.1 In a certain population of the European starling, there are 5,000 nests with young. The distribution of brood size (number of young in a nest) is given in the accompanying table.[12]

Brood Size	Frequency (No. of Broods)
1	90
2	230
3	610
4	1,400
5	1,760
6	750
7	130
8	26
9	3
10	1
Total	5,000

Suppose one of the 5,000 broods is to be chosen at random, and let Y be the size of the chosen brood. Find

(a) $\Pr\{Y = 3\}$

(b) $\Pr\{Y \geq 7\}$

(c) $\Pr\{4 \leq Y \leq 6\}$

3.5.2 In the starling population of Exercise 3.5.1, there are 22,435 young in all the broods taken together. (There are 90 young from broods of size 1, there are 460 from broods of size 2, etc.) Suppose one of the young is to be chosen at random, and let Y' be the size of the chosen individual's brood.

(a) Find $\Pr\{Y' = 3\}$.

(b) Find $\Pr\{Y' \geq 7\}$.

(c) Explain why choosing a young at random and then observing its brood is not equivalent to choosing a brood at random. Your explanation should show why the answer to part (b) is greater than the answer to part (b) of Exercise 3.5.1.

3.5.3 Calculate the mean, μ_Y, of the random variable Y from Exercise 3.5.1.

3.5.4 Consider a population of the fruitfly *Drosophila melanogaster* in which 30% of the individuals are black because of a mutation, while 70% of the individuals have the normal gray body color. Suppose three flies are chosen at random from the population; let Y denote the number of black flies out of the three. Then the probability distribution for Y is given by the following table:

Y (No. Black)	Probability
0	0.343
1	0.441
2	0.189
3	0.027
Total	1.000

(a) Find $\Pr\{Y \geq 2\}$

(b) Find $\Pr\{Y \leq 2\}$

3.5.5 Calculate the mean, μ_Y, of the random variable Y from Exercise 3.5.4.

3.5.6 Calculate the standard deviation, σ_Y, of the random variable Y from Exercise 3.5.4.

3.5.7 The prevalence of mild myopia (nearsightedness) in adults over age 40 is 25% in the U.S.[13] Suppose four adults over age 40 are chosen at random from the population; let Y denote the number with myopia out of the four. Then the probability distribution for Y is given by the following table:

Y (No. Myopic)	Probability
0	0.316
1	0.422
2	0.211
3	0.047
4	0.004
Total	1.000

(a) Find $\Pr\{Y \geq 3\}$

(b) Find $\Pr\{Y \leq 1\}$

(c) Find $\Pr\{Y \geq 1\}$

3.5.8 Calculate the mean, μ_Y, of the random variable Y from Exercise 3.5.7.

3.5.9 A group of college students were surveyed to learn how many times they had visited a dentist in the previous year.[14] The probability distribution for Y, the number of visits, is given by the following table:

Y (No. Visits)	Probability
0	0.15
1	0.50
2	0.35
Total	1.00

Calculate the mean, μ_Y, of the number of visits.

3.5.10 Calculate the standard deviation, σ_Y, of the random variable Y from Exercise 3.5.9.

3.6 The Binomial Distribution

To add some depth to the notion of probability and random variables, we now consider a special type of random variable, the **binomial**. The distribution of a binomial random variable is a probability distribution associated with a special kind of chance operation. The chance operation is defined in terms of a set of conditions called the independent-trials model.

THE INDEPENDENT-TRIALS MODEL

The **independent-trials model** relates to a sequence of chance "trials." Each trial is assumed to have two possible outcomes, which are arbitrarily labeled "success" and "failure." The probability of success on each individual trial is denoted by the letter p and is assumed to be constant from one trial to the next. In addition, the trials are required to be independent, which means that the chance of success or failure on each trial does not depend on the outcome of any other trial. The total number of trials is denoted by n. These conditions are summarized in the following definition of the model.

Independent-Trials Model

A series of n independent trials is conducted. Each trial results in success or failure. The probability of success is equal to the same quantity, p, for each trial, regardless of the outcomes of the other trials.

The following examples illustrate situations that can be described by the independent-trials model.

Example 3.6.1 **Albinism** If two carriers of the gene for albinism marry, each of their children has probability $1/4$ of being albino. The chance that the second child is albino is the same ($1/4$) whether or not the first child is albino; similarly, the outcome for the third child is independent of the first two, and so on. Using the labels "success" for albino and "failure" for nonalbino, the independent-trials model applies with $p = 1/4$ and n = the number of children in the family. ■

Example 3.6.2 **Mutant Cats** A study of cats in Omaha, Nebraska, found that 37% of them have a certain mutant trait.[15] Suppose that 37% of all cats have this mutant trait and that a random sample of cats is chosen from the population. As each cat is chosen for the sample, the probability is 0.37 that it will be mutant. This probability is the same as each cat is chosen, regardless of the results of the other cats, because the percentage of mutants in the large population remains equal to 0.37 even when a few individual cats have been removed. Using the labels "success" for mutant and "failure" for nonmutant, the independent-trials model applies with $p = 0.37$ and n = the sample size. ■

AN EXAMPLE OF THE BINOMIAL DISTRIBUTION

The binomial distribution specifies the probabilities of various numbers of successes and failures when the basic chance operation consists of n independent trials. Before giving the general formula for the binomial distribution, we consider a simple example.

Example 3.6.3 **Albinism** Suppose two carriers of the gene for albinism marry (see Example 3.6.1) and have two children. Then the probability that both of their children are albino is

$$\Pr\{\text{both children are albino}\} = \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}$$

The reason for this probability can be seen by considering the relative frequency interpretation of probability. Of a great many such families with two children, $\frac{1}{4}$ would have the first child albino; furthermore, $\frac{1}{4}$ *of these* would have the second child albino; thus, $\frac{1}{4}$ of $\frac{1}{4}$, or $\frac{1}{16}$ of all the couples would have both albino children. A similar kind of reasoning shows that the probability that both children are not albino is

$$\Pr\{\text{both children are not albino}\} = \left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{16}$$

A new twist enters if we consider the probability that one child is albino and the other is not. There are two possible ways this can happen:

$$\Pr\{\text{first child is albino, second is not}\} = \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{3}{16}$$

$$\Pr\{\text{first child is not albino, second is}\} = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16}$$

To see how to combine these possibilities, we again consider the relative frequency interpretation of probability. Of a great many such families with two children, the fraction of families with one albino and one nonalbino child would be the total of the two possibilities, or

$$\left(\frac{3}{16}\right) + \left(\frac{3}{16}\right) = \frac{6}{16}$$

Thus, the corresponding probability is

$$\Pr\{\text{one child is albino, the other is not}\} = \frac{6}{16}$$

Another way to see this is to consider a probability tree. The first split in the tree represents the birth of the first child; the second split represents the birth of the second child. The four possible outcomes and their associated probabilities are shown in Figure 3.6.1. These probabilities are collected in Table 3.6.1. ■

The probability distribution in Table 3.6.1 is called the binomial distribution with $p = \frac{1}{4}$ and $n = 2$. Note that the probabilities add to 1. This makes sense because all possibilities have been accounted for: We expect $\frac{9}{16}$ of the families to have no albino children, $\frac{6}{16}$ to have one albino child, and $\frac{1}{16}$ to have two albino children; there are no other possible compositions for a two-child family. The number of albino children, out of the two children, is an example of a binomial random variable. A **binomial random variable** is a random variable that satisfies the following four conditions, abbreviated as **BInS**:

Binary outcomes: There are two possible outcomes for each trial (success and failure).

Independent trials: The outcomes of the trials are independent of each other.

n is fixed: The number of trials, n, is fixed in advance.

Same value of p: The probability of a success on a single trial is the same for all trials.

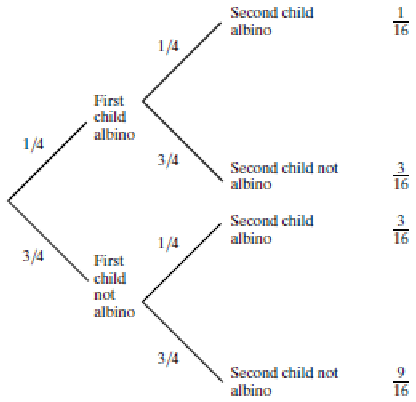


Figure 3.6.1 Probability tree for albinism among two children of carriers of the gene for albinism

Table 3.6.1 Probability distribution for number of albino children

Number of		
Albino	Nonalbino	Probability
0	2	$\frac{9}{16}$
1	1	$\frac{6}{16}$
2	0	$\frac{1}{16}$
	Total	1

THE BINOMIAL DISTRIBUTION FORMULA

A general formula is available that can be used to calculate probabilities associated with a binomial random variable for any values of n and p. This formula can be proved using logic similar to that in Example 3.6.3. (The formula is discussed further in Appendix 3.1.) The formula is given in the accompanying box.

The Binomial Distribution Formula

For a binomial random variable Y, the probability that the n trials result in j successes (and $n - j$ failures) is given by the following formula:

$$\Pr\{j \text{ successes}\} = \Pr\{Y = j\} = {}_nC_j p^j(1 - p)^{n-j}$$

The quantity ${}_nC_j$ appearing in the formula is called a **binomial coefficient**. Each binomial coefficient is an integer depending on n and on j. Values of binomial coefficients are given in Table 2 at the end of this book and can be found by the formula

$${}_nC_j = \frac{n!}{j!(n - j)!}$$

where $x!$ ("x-factorial") is defined for any positive integer x as

$$x! = x(x - 1)(x - 2) \ldots (2)(1)$$

and $0! = 1$. For more details, see Appendix 3.1.

For example, for $n = 5$ the binomial coefficients are as follows:

j:	0	1	2	3	4	5
${}_5C_j$:	1	5	10	10	5	1

Thus, for $n = 5$ the binomial probabilities are as indicated in Table 3.6.2. Notice the pattern in Table 3.6.2: The powers of p ascend (0, 1, 2, 3, 4, 5) and the powers of ($1 - p$) descend (5, 4, 3, 2, 1, 0). (In using the binomial distribution formula, remember that $x^0 = 1$ for any nonzero x.)

Table 3.6.2 Binomial probabilities for $n = 5$

Number of		
Successes j	Failures $n - j$	Probability
0	5	$1p^0(1-p)^5$
1	4	$5p^1(1-p)^4$
2	3	$10p^2(1-p)^3$
3	2	$10p^3(1-p)^2$
4	1	$5p^4(1-p)^1$
5	0	$1p^5(1-p)^0$

Notes on Table 2 The following features in Table 2 are worth noting:

(a) The first and last entries in each row are equal to 1. This will be true for any row; that is, ${}_nC_0 = 1$ and ${}_nC_n = 1$ for any value of n.

(b) Each row of the table is symmetric; that is ${}_nC_j$ and ${}_nC_{n-j}$ are equal.

(c) The bottom rows of the table are left incomplete to save space, but you can easily complete them using the symmetry of the ${}_nC_j$'s; if you need to know ${}_nC_j$, you can look up ${}_nC_{n-j}$ in Table 2. For instance, consider $n = 18$; if you want to know ${}_{18}C_{15}$, you just look up ${}_{18}C_3$; both ${}_{18}C_3$ and ${}_{18}C_{15}$ are equal to 816.

The following example shows a specific application of the binomial distribution with $n = 5$.

Example 3.6.4 **Mutant Cats** Suppose we draw a random sample of five individuals from a large population in which 37% of the individuals are mutants (as in Example 3.6.2). The probabilities of the various possible samples are then given by the binomial distribution formula with $n = 5$ and $p = 0.37$; the results are displayed in Table 3.6.3. For instance, the probability of a sample containing two mutants and three nonmutants is

$$10(0.37)^2(0.63)^3 \approx 0.34$$

Table 3.6.3 Binomial distribution with $n = 5$ and $p = 0.37$

Number of		
Mutants	Nonmutants	Probability
0	5	0.10
1	4	0.29
2	3	0.34
3	2	0.20
4	1	0.06
5	0	0.01
	Total	1.00

Thus, $\Pr\{Y = 2\} \approx 0.34$. This means that about 34% of random samples of size 5 will contain two mutants and three nonmutants.

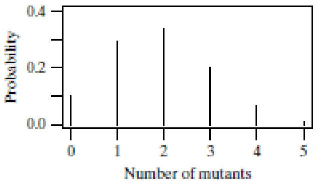


Figure 3.6.2 Binomial distribution with $n = 5$ and $p = 0.37$

Notice that the probabilities in Table 3.6.3 add to 1. The probabilities in a probability distribution must always add to 1, because they account for 100% of the possibilities.

The binomial distribution of Table 3.6.3 is pictured graphically in Figure 3.6.2. The spikes in the graph emphasize that the probability distribution is discrete.

Remark In applying the independent-trials model and the binomial distribution, we assign the labels "success" and "failure" arbitrarily. For instance, in Example 3.6.4, we could say "success" = "mutant" and $p = 0.37$; or, alternatively, we could say "success" = "nonmutant" and $p = 0.63$. Either assignment of labels is all right; it is only necessary to be consistent.

Computational Note Computer and calculator technology makes it fairly easy to handle the binomial distribution formula for small or moderate values of n. For large values of n, the use of the binomial formula gets to be tedious and even a computer will balk at being asked to calculate a binomial probability. However, the binomial formula can be approximated by other methods. One of these will be discussed in the optional Section 5.4.

Sometimes a binomial probability question involves combining two or more possible outcomes. The following example illustrates this idea.

Example 3.6.5

Sampling Fruitflies In a large *Drosophila* population, 30% of the flies are black (B) and 70% are gray (G). Suppose two flies are randomly chosen from the population (as in Example 3.2.3). The binomial distribution with $n = 2$ and $p = 0.3$ gives probabilities for the possible outcomes as shown in Table 3.6.4. (Using the binomial formula agrees with the results given by the probability tree shown in Figure 3.2.3.)

Table 3.6.4

Sample composition	Y	Probability
Both Gray	0	0.49
One Black, one Gray	1	0.42
Both Black	2	0.09
	Total	1.00

Let E be the event that both flies are the same color. Then E can happen in two ways: Both flies are gray or both are black. To find the probability of E, consider what would happen if we repeated the sampling procedure many times: Forty-nine percent of the samples would have both flies gray, and 9% would have both flies black. Consequently, the percentage of samples with both flies the same color would be 49% + 9% = 58%. Thus, we have shown that the probability of E is

$$\Pr\{E\} = 0.58$$

as we claimed in Example 3.2.3.

Whenever an event E can happen in two or more mutually exclusive ways, a rationale such as that of Example 3.6.5 can be used to find $\Pr\{E\}$.

Example 3.6.6

Blood Type In the United States, 85% of the population has Rh positive blood. Suppose we take a random sample of 6 persons and count the number with Rh positive blood. The binomial model can be applied here, since the BInS conditions are met: There is a binary outcome on each trial (Rh positive or Rh negative blood), the trials are independent (due to the random sampling), n is fixed at 6, and the same probability of Rh positive blood applies to each person ($p = 0.85$).

Let Y denote the number of persons, out of 6, with Rh positive blood. The probabilities of the possible values of Y are given by the binomial distribution formula with $n = 6$ and $p = 0.85$; the results are displayed in Table 3.6.5. For instance, the probability that $Y = 4$ is

$${}_6C_4(0.85)^4(0.15)^2 \approx 15(0.522)(0.0225) \approx 0.1762$$

If we want to find the probability that at least 4 persons (out of the 6 sampled) will have Rh positive blood, we need to find $\Pr\{Y \geq 4\} = \Pr\{Y = 4\} + \Pr\{Y = 5\} + \Pr\{Y = 6\} = 0.1762 + 0.3993 + 0.3771 = 0.9526$. This means that the probability of getting at least 4 persons with Rh positive blood in a sample of size 6 is 0.9526.

Table 3.6.5 Binomial distribution with $n = 6$ and $p = 0.85$

Number of successes	Probability
0	<0.0001
1	0.0004
2	0.0055
3	0.0415
4	0.1762
5	0.3993
6	0.3771
Total	1.0000

In some problems, it is easier to find the probability that an event *does not happen* rather than finding the probability of the event happening. To solve such problems we use the fact that the probability of an event happening is 1 minus the probability that the event does not happen: $\Pr\{E\} = 1 - \Pr\{E \text{ does not happen}\}$. The following is an example.

Example 3.6.7

Blood Type As in Example 3.6.6, let Y denote the number of persons, out of 6, with Rh positive blood. Suppose we want to find the probability that Y is less than 6 (i.e., the probability that there is *at least 1* person in the sample who has Rh *negative* blood). We could find this directly as $\Pr\{Y = 0\} + \Pr\{Y = 1\} + \cdots + \Pr\{Y = 5\}$. However, it is easier to find $\Pr\{Y = 6\}$ and subtract this from 1:

$$\Pr\{Y < 6\} = 1 - \Pr\{Y = 6\} = 1 - 0.3771 = 0.6229.$$

MEAN AND STANDARD DEVIATION OF A BINOMIAL

If we toss a fair coin 10 times, then we expect to get 5 heads, on average. This is an example of a general rule: *For a binomial random variable, the mean (i.e., the average number of successes) is equal to np.* This is an intuitive fact: The probability of success on each trial is p, so if we conduct n trials, then np is the expected number of successes. In Appendix 3.2 we show that this result is consistent with the rule given in Section 3.5 for finding the mean of the sum of random variables. *The standard deviation for a binomial random variable is given by $\sqrt{np(1-p)}$.* This formula is not intuitively clear; a derivation of the result is given in Appendix 3.2. For the example of tossing a coin 10 times, the standard deviation of the number of heads is

$$\sqrt{10 \times 0.5 \times 0.5} = \sqrt{2.5} \approx 1.58.$$

Example 3.6.8

Blood Type As discussed in Example 3.6.6, if Y denotes the number of persons with Rh positive blood in a sample of size 6, then a binomial model can be used to find probabilities associated with Y. The single most likely value of Y is 5 (which has probability 0.3993). The average value of Y is $6 \times 0.85 = 5.1$, which means that if we take many samples, each of size 6, and count the number of Rh positive persons in each sample, and then average those counts, we expect to get 5.1. The standard deviation of those counts is $\sqrt{6 \times 0.85 \times .015} \approx 0.87$. ■

APPLICABILITY OF THE BINOMIAL DISTRIBUTION

A number of statistical procedures are based on the binomial distribution. We will study some of these procedures in later chapters. Of course, the binomial distribution is applicable only in experiments where the BInS conditions are satisfied in the real biological situation. We briefly discuss some aspects of these conditions.

Application to Sampling The most important application of the independent-trials model and the binomial distribution is to describe random sampling from a population when the observed variable is dichotomous—that is, a categorical variable with two categories (e.g., black and gray in Example 3.6.5). This application is valid if the sample size is a negligible fraction of the population size so that the population composition is not altered appreciably by the removal of the individuals in the sample (so that the S part of BInS is satisfied: The probability of a success remains the same from trial to trial). However, if the sample is *not* a negligibly small part of the population, then the population composition may be altered by the sampling process so that the "trials" involved in composing the sample are not independent and the probability of a success changes as the sampling progresses. In this case, the probabilities given by the binomial formula are not correct. In most biological studies, the population is so large that this kind of difficulty does not arise.

Contagion In some applications the phenomenon of contagion can invalidate the condition of independence between trials. The following is an example.

Example 3.6.9

Chickenpox Consider the occurrence of chickenpox in children. Each child in a family can be categorized according to whether he had chickenpox during a certain year. One can say that each child constitutes a "trial" and that "success" is having chickenpox during the year, but the trials are *not* independent because the chance of a particular child catching chickenpox depends on whether his sibling caught chickenpox. As a specific example, consider a family with five children, and suppose that the chance

of an individual child catching chickenpox during the year is equal to 0.10. The binomial distribution gives the chance of all five children getting chickenpox as

$$\Pr\{5 \text{ children get chickenpox}\} = (0.10)^5 = 0.00001$$

However, this answer is not correct; because of contagion, the correct probability would be much larger. There would be many families in which one child caught chickenpox and then the other four children got chickenpox from the first child, so all five children would get chickenpox. ■

Exercises 3.6.1–3.6.12

3.6.1 The seeds of the garden pea *(Pisum sativum)* are either yellow or green. A certain cross between pea plants produces progeny in the ratio 3 yellow : 1 green.[16] If four randomly chosen progeny of such a cross are examined, what is the probability that

(a) three are yellow and one is green?

(b) all four are yellow?

(c) all four are the same color?

3.6.2 In Australia, 16% of the adult population is nearsighted.[17] If three Australians are chosen at random, what is the probability that

(a) two are nearsighted and one is not?

(b) exactly one is nearsighted?

(c) at most one is nearsighted?

(d) none of them are nearsighted?

3.6.3 In the United States, 44% of the population has type A blood. Consider taking a sample of size 4. Let Y denote the number of persons in the sample with type A blood. Find

(a) $\Pr\{Y = 0\}$.

(b) $\Pr\{Y = 1\}$.

(c) $\Pr\{Y = 2\}$.

(d) $\Pr\{0 \le Y \le 2\}$.

(e) $\Pr\{0 < Y \le 2\}$.

3.6.4 A certain drug treatment cures 90% of cases of hookworm in children.[18] Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

(a) all 20 will be cured.

(b) all but 1 will be cured.

(c) exactly 18 will be cured.

(d) exactly 90% will be cured.

3.6.5 The shell of the land snail *Limocolaria martensiana* has two possible color forms: streaked and pallid. In a certain population of these snails, 60% of the individuals have streaked shells.[19] Suppose that a random sample of 10 snails is to be chosen from this population. Find the probability that the percentage of streaked-shelled snails in the *sample* will be

(a) 50%. (b) 60%. (c) 70%.

3.6.6 Consider taking a sample of size 10 from the snail population in Exercise 3.6.5.

(a) What is the mean number of streaked-shelled snails?

(b) What is the standard deviation of the number of streaked-shelled snails?

3.6.7 In Europe, 8% of men are colorblind.[20] Consider taking repeated samples of 20 European men.

(a) What is the mean number of colorblind men?

(b) What is the standard deviation of the number of colorblind men?

3.6.8 The sex ratio of newborn human infants is about 105 males : 100 females.[21] If four infants are chosen at random, what is the probability that

(a) two are male and two are female?

(b) all four are male?

(c) all four are the same sex?

3.6.9 Construct a binomial setting (different from any examples presented in this book) and a problem for which the following is the answer: ${}_7C_3(0.8)^3(0.2)^4$.

3.6.10 Neuroblastoma is a rare, serious, but treatable disease. A urine test, the VMA test, has been developed that gives a positive diagnosis in about 70% of cases of neuroblastoma.[22] It has been proposed that this test be used for large-scale screening of children. Assume that 300,000 children are to be tested, of whom 8 have the disease. We are interested in whether or not the test detects the disease in the 8 children who have the disease. Find the probability that

(a) all eight cases will be detected.

(b) only one case will be missed.

(c) two or more cases will be missed. [*Hint:* Use parts (a) and (b) to answer part (c).]

3.6.11 If two carriers of the gene for albinism marry, each of their children has probability $\frac{1}{4}$ of being albino (see

Example 3.6.1). If such a couple has six children, what is the probability that

(a) none will be albino?

(b) at least one will be albino? [*Hint:* Use part (a) to answer part (b); note that "at least one" means "one or more."]

3.6.12 Childhood lead poisoning is a public health concern in the United States. In a certain population, 1 child in 8 has a high blood lead level (defined as 30 µg/dl or more).[23] In a randomly chosen group of 16 children from the population, what is the probability that

(a) none has high blood lead?

(b) 1 has high blood lead?

(c) 2 have high blood lead?

(d) 3 or more have high blood lead? [*Hint:* Use parts (a)–(c) to answer part (d).]

3.7 Fitting a Binomial Distribution to Data (Optional)

Occasionally it is possible to obtain data that permit a direct check of the applicability of the binomial distribution. One such case is described in the next example.

Example 3.7.1

Sexes of Children In a classic study of the human sex ratio, families were categorized according to the sexes of the children. The data were collected in Germany in the nineteenth century, when large families were common. Table 3.7.1 shows the results for 6,115 families with 12 children.[24]

Table 3.7.1 Sex ratios in 6,115 families with 12 children

Number of		Observed frequency
Boys	Girls	(number of families)
0	12	3
1	11	24
2	10	104
3	9	286
4	8	670
5	7	1,033
6	6	1,343
7	5	1,112
8	4	829
9	3	478
10	2	181
11	1	45
12	0	7
	Total	6,115

It is interesting to consider whether the observed variation among families can be explained by the independent-trials model. We will explore this question by fitting a binomial distribution to the data.

The first step in fitting the binomial distribution is to determine a value for $p = \Pr\{\text{boy}\}$. One possibility would be to assume that $p = 0.50$. However, since it is known that the human sex ratio at birth is not exactly 1 : 1 (in fact, it favors boys slightly), we will not make this assumption. Rather, we will "fit" p to the data; that is, we will determine a value for p that fits the data best. We observe that the total number of children in all the families is

$$(12)(6{,}115) = 73{,}380 \text{ children}$$

Among these children, the number of boys is

$$(3)(0) + (24)(1) + \cdots + (7)(12) = 38{,}100 \text{ boys}$$

Therefore, the value of p that fits the data best is

$$p = \frac{38{,}100}{73{,}380} = 0.519215$$

The next step is to compute probabilities from the binomial distribution formula with $n = 12$ and $p = 0.519215$. For instance, the probability of 3 boys and 9 girls is computed as

$$\begin{aligned} {}_{12}C_3(p)^3(1-p)^9 &= 220(0.519215)^3(0.480785)^9 \\ &\approx 0.042269 \end{aligned}$$

For comparison with the observed data, we convert each probability to a theoretical or "expected" frequency by multiplying by 6,115 (the total number of families). For instance, the expected number of families with 3 boys and 9 girls is

$$(6{,}115)(0.042269) \approx 258.5$$

The expected and observed frequencies are displayed together in Table 3.7.2. Table 3.7.2 shows reasonable agreement between the observed frequencies and the predictions of the binomial distribution. But a closer look reveals that the discrepancies, although not large, follow a definite pattern. The data contain more unisexual, or preponderantly unisexual, sibships than expected. In fact, the observed frequencies are higher than the expected frequencies for nine types of families in which one sex or the other predominates, while the observed frequencies are lower than the expected frequencies for four types of more "balanced" families. This pattern is

Table 3.7.2 Sex-ratio data and binomial expected frequencies

Number of		Observed	Expected	Sign of
Boys	Girls	frequency	frequency	(Obs. – Exp.)
0	12	3	0.9	+
1	11	24	12.1	+
2	10	104	71.8	+
3	9	286	258.5	+
4	8	670	628.1	+
5	7	1,033	1,085.2	–
6	6	1,343	1,367.3	–
7	5	1,112	1,265.6	–
8	4	829	854.2	–
9	3	478	410.0	+
10	2	181	132.8	+
11	1	45	26.1	+
12	0	7	2.3	+
	Total	6,115	6,115.0	

clearly revealed by the last column of Table 3.7.2, which shows the sign of the difference between the observed frequency and the expected frequency. Thus, the observed distribution of sex ratios has heavier "tails" and a lighter "middle" than the best-fitting binomial distribution.

The systematic pattern of deviations from the binomial distribution suggests that the observed variation among families cannot be entirely explained by the independent-trials model.* What factors might account for the discrepancy? This intriguing question has stimulated several researchers to undertake more detailed analysis of these data. We briefly discuss some of the issues.

One explanation for the excess of predominantly unisexual families is that the probability of producing a boy may vary among families. If p varies from one family to another, then sex will appear to "run" in families in the sense that the number of predominantly unisexual families will be inflated. In order to clearly visualize this effect, consider the fictitious data set shown in Table 3.7.3.

Table 3.7.3 Fictitious sex-ratio data and binomial expected frequencies

Number of		Observed	Expected	Sign of
Boys	Girls	frequency	frequency	(Obs. – Exp.)
0	12	2,940	0.9	+
1	11	0	12.1	–
2	10	0	71.8	–
3	9	0	258.5	–
4	8	0	628.1	–
5	7	0	1,085.2	–
6	6	0	1,367.3	–
7	5	0	1,265.6	–
8	4	0	854.3	–
9	3	0	410.0	–
10	2	0	132.8	–
11	1	0	26.1	–
12	0	3,175	2.3	+
	Total	6,115	6,115.0	

In the fictitious data set, there are (3,175)(12) = 38,100 males among 73,380 children, just as there are in the real data set. Consequently, the best-fitting p is the same ($p = 0.519215$) and the expected binomial frequencies are the same as in Table 3.7.2. The fictitious data set contains only unisexual sibships and so is an extreme example of sex "running" in families. The real data set exhibits the same phenomenon more weakly. One explanation of the fictitious data set would be that some families can have only boys ($p = 1$) and other families can have only girls ($p = 0$). In a parallel way, one explanation of the real data set would be that p varies slightly among families. Variation in p is biologically plausible, even though a mechanism causing the variation has not been discovered.

An alternative explanation for the inflated number of sexually homogeneous families would be that the sexes of the children in a family are literally dependent on one another, in the sense that the determination of an individual child's sex is somehow influenced by the sexes of the previous children. This explanation is implausible on biological grounds because it is difficult to imagine how the biological system could "remember" the sexes of previous offspring. ■

*A chi-square goodness-of-fit test of the binomial model shows that there is strong evidence that the differences between the observed and expected frequencies did not happen due to chance error in the sampling process. We will explore the topic of goodness-of-fit tests in Chapter 9.

Example 3.7.1 shows that poorness of fit to the independent-trials model can be biologically interesting. We should emphasize, however, that most statistical applications of the binomial distribution proceed from the assumption that the independent-trials model is applicable*. In a typical application, the data are regarded as resulting from a *single* set of n trials. Data such as the family sex-ratio data, which refer to *many* sets of $n = 12$ trials, are not often encountered.

*In Example 3.6.1 we asserted that occurrences of albinism among siblings are independent, which is consistent with current understandings of human genetics.

Exercises 3.7.1–3.7.3

3.7.1 The accompanying data on families with 6 children are taken from the same study as the families with 12 children in Example 3.7.1. Fit a binomial distribution to the data. (Round the expected frequencies to one decimal place.) Compare with the results in Example 3.7.1. What features do the two data sets share?

Number of		
Boys	Girls	Number of families
0	6	1,096
1	5	6,233
2	4	15,700
3	3	22,221
4	2	17,332
5	1	7,908
6	0	1,579
	Total	72,069

3.7.2 An important method for studying mutation-causing substances involves killing female mice 17 days after mating and examining their uteri for living and dead embryos. The classical method of analysis of such data assumes that the survival or death of each embryo constitutes an independent binomial trial. The accompanying table, which is extracted from a larger study, gives data for 310 females, all of whose uteri contained 9 embryos; all of the animals were treated alike (as controls).[25]

(a) Fit a binomial distribution to the observed data. (Round the expected frequencies to one decimal place.)

(b) Interpret the relationship between the observed and expected frequencies. Do the data cast suspicion on the classical assumption?

Number of embryos		Number of
Dead	Living	female mice
0	9	136
1	8	103
2	7	50
3	6	13
4	5	6
5	4	1
6	3	1
7	2	0
8	1	0
9	0	0
	Total	310

3.7.3 Students in a large botany class conducted an experiment on the germination of seeds of the Saguaro cactus. As part of the experiment, each student planted five seeds in a small cup, kept the cup near a window, and checked every day for germination (sprouting). The class results on the seventh day after planting were as displayed in the table.[26]

Number of seeds		Number of
Germinated	Not germinated	students
0	5	17
1	4	53
2	3	94
3	2	79
4	1	33
5	0	4
	Total	280

(a) Fit a binomial distribution to the data. (Round the expected frequencies to one decimal place.)

(b) Two students, Fran and Bob, were talking before class. All of Fran's seeds had germinated by the seventh day, whereas none of Bob's had. Bob wondered whether he had done something wrong. With the perspective gained from seeing all 280 students' results, what would you say to Bob? (*Hint*: Can the variation among the students be explained by the hypothesis that some of the seeds were good and some were poor, with each student receiving a randomly chosen five seeds?)

(c) Invent a fictitious set of data for 280 students, with the same overall percentage germination as the observed data given in the table, but with all the students getting either Fran's results (perfect) or Bob's results (nothing). How would your answer to Bob differ if the actual data had looked like this fictitious data set?

Supplementary Exercises 3.S.1–3.S.12

3.S.1 In the United States, 10% of adolescent girls have iron deficiency.[27] Suppose two adolescent girls are chosen at random. Find the probability that

(a) all three girls have iron deficiency.

(b) one girl has iron deficiency and the other two do not.

3.S.2 In preparation for an ecological study of centipedes, the floor of a beech woods is divided into a large number of 1-foot squares.[28] At a certain moment, the distribution of centipedes in the squares is as shown in the table.

Number of centipedes	Percent frequency (% of squares)
0	45
1	36
2	14
3	4
4	1
Total	100

Suppose that a square is chosen at random, and let Y be the number of centipedes in the chosen square. Find

(a) $\Pr\{Y = 2\}$

(b) $\Pr\{Y \geq 3\}$

3.S.3 Refer to the distribution of centipedes given in Exercise 3.S.2. Suppose four squares are chosen at random. Find the probability that two of the squares contain centipedes and two do not.

3.S.4 Refer to the distribution of centipedes given in Exercise 3.S.2. Suppose four squares are chosen at random. Find the expected value (i.e., the mean) of the number of squares that contain at least two centipedes.

3.S.5 Wavy hair in mice is a recessive genetic trait. If mice with wavy hair are mated with straight-haired (heterozygous) mice, each offspring has probability $\frac{1}{2}$ of having wavy hair.[29] Consider a large number of such matings, each producing a litter of five offspring. What percentage of the litters will consist of

(a) three wavy-haired and three straight-haired offspring?

(b) four or more straight-haired offspring?

(c) all the same type (either all wavy- or all straight-haired) offspring?

3.S.6 A certain drug causes liver damage in 5% of patients. Suppose the drug is to be tested on 60 patients. Find the probability that

(a) none of the patients will experience liver damage.

(b) one or more of the patients will experience liver damage. [*Hint*: Use part (a) to answer part (b).]

3.S.7 Refer to Exercise 3.S.6. Suppose now that the drug is to be tested on n patients, and let E represent the event that liver damage occurs in one or more of the patients. The probability $\Pr\{E\}$ is useful in establishing criteria for drug safety.

(a) Find $\Pr\{E\}$ for $n = 100$.

(b) How large must n be in order for $\Pr\{E\}$ to exceed 0.94?

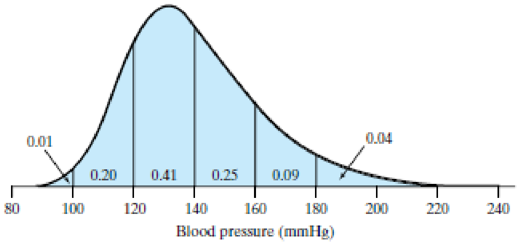
3.S.8 To study people's ability to deceive lie detectors, researchers sometimes use the "guilty knowledge" technique.[30] Certain subjects memorize six common words; other subjects memorize no words. Each subject is then tested on a polygraph machine (lie detector), as follows. The experimenter reads, in random order, 24 words: the six "critical" words (the memorized list) and, for each critical word, three "control" words with similar or related meanings. If the subject has memorized the six words, he or she tries to conceal that fact. The subject is scored a "failure" on a critical word if his or her electrodermal response is higher on the critical word than on any of the three control words. Thus, on each of the six critical words, even an innocent subject would have a 25% chance of failing. Suppose a subject is labeled "guilty" if the subject fails on five or more of the six critical words. If an innocent subject is tested, what is the probability that he or she will be labeled "guilty"?

3.S.9 The density curve shown here represents the distribution of systolic blood pressures in a population of middle-aged men.[31] Areas under the curve are shown in the figure. Suppose a man is selected at random from the population, and let Y be his blood pressure. Find

(a) $\Pr\{140 < Y < 180\}$.

(b) $\Pr\{Y < 140\}$.

(c) $\Pr\{\mathrm{Y} > 160\}$.



3.S.10 Refer to the blood pressure distribution of Exercise 3.S.9. Suppose four men are selected at random from the population. Find the probability that

(a) all four have blood pressures higher than 160 mm Hg.

(b) three have blood pressures higher than 160, and one has blood pressure 160 or less.

3.S.11 In the United States 9% of all people are left-handed.[32] If we take a random sample of six Americans what is the probability that

(a) none of them is left-handed?

(b) all six are left-handed?

(c) at least one is left-handed?

3.S.12 Refer to the information about left-handedness in Exercise 3.S.11. Consider taking repeated samples of 100 Americans.

(a) What is the mean number of left-handed persons?

(b) What is the standard deviation of the number of left-handed persons?