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 manipulating 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 determined at least partially by chance. 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.

Another chance operation that would be adequate is

Chance operation: Toss a coin 100 times

with the understanding that E occurs if there is a run of 3 heads anywhere in the 100 tosses. Intuition suggests that E would be more likely with the second definition of the chance operation (100 tosses) than with the first (3 tosses). This intuition is correct and serves to underscore the importance of the chance operation in interpreting a probability. ■

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.*

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 "approximate equality in the long run"; that is, if the chance operation is repeated many 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 a 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.

*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.

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 (that is, 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.

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 %	– 4	or	–4 %
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),

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

Sample	Color			Relative frequency
number	1st Fly	2nd Fly	Did *E* occur?	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

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.

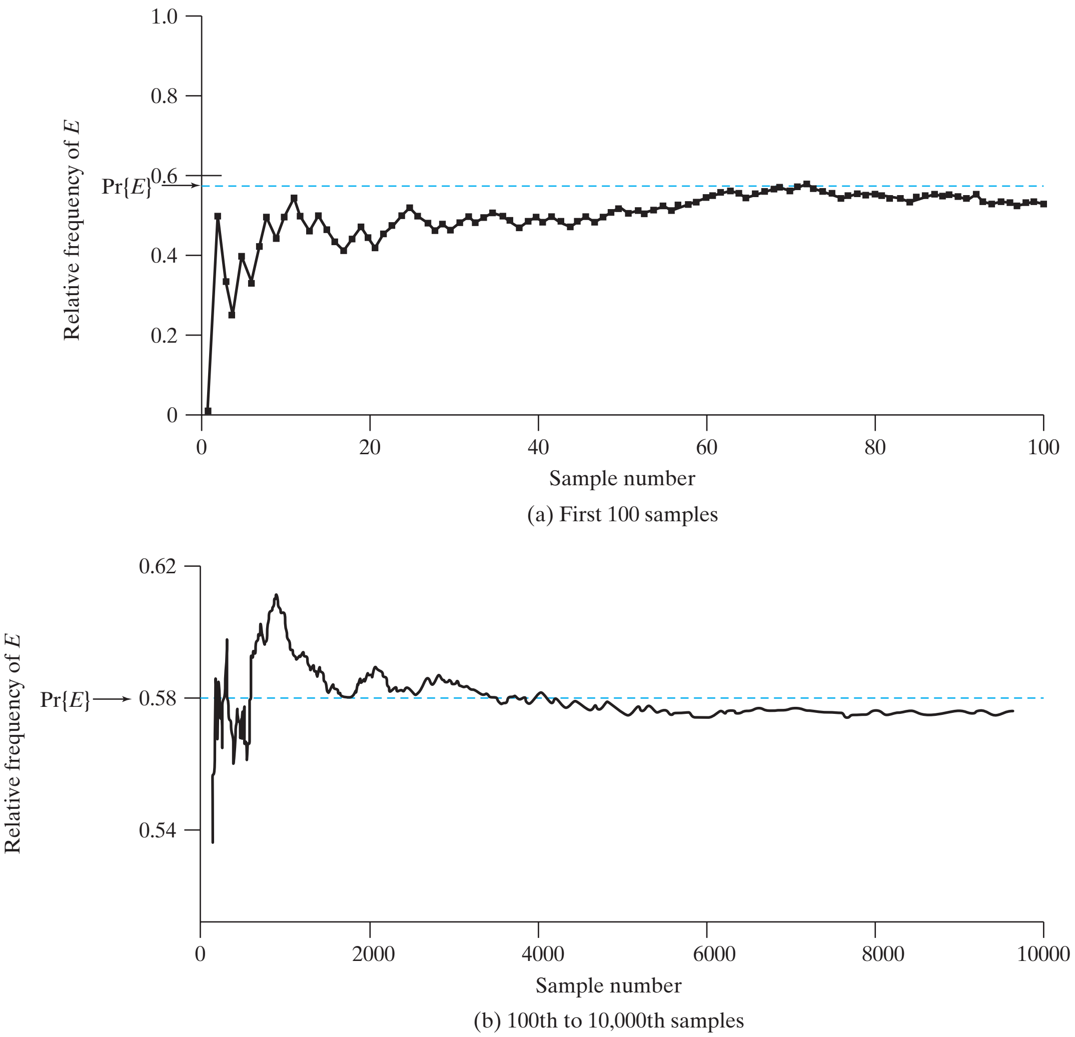
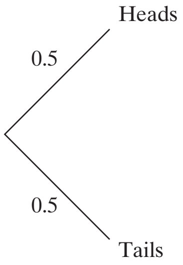


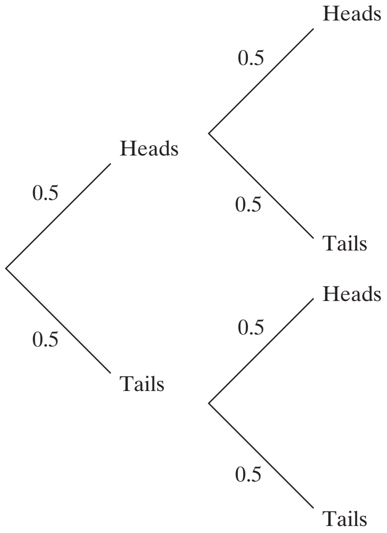
Figure 3.2.1 Results of sampling from fruitfly population. Note that the axes are scaled differently in (a) and (b).

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

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

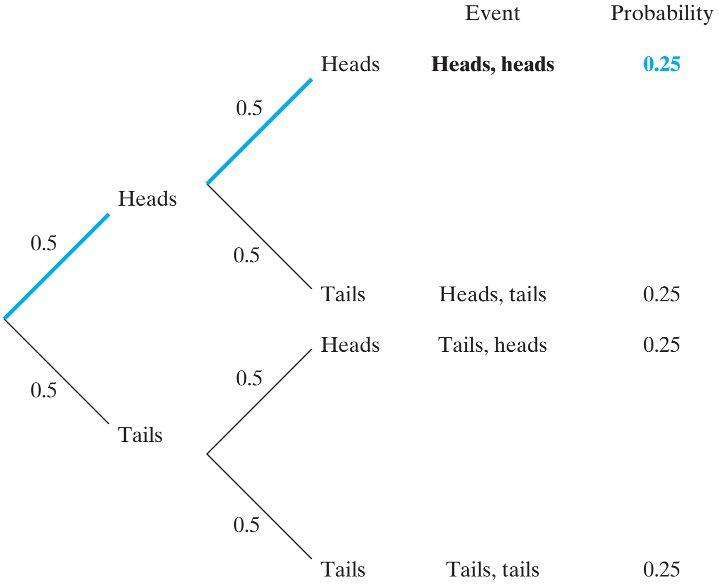


Figure 3.2.2 Probability tree for two coin tosses

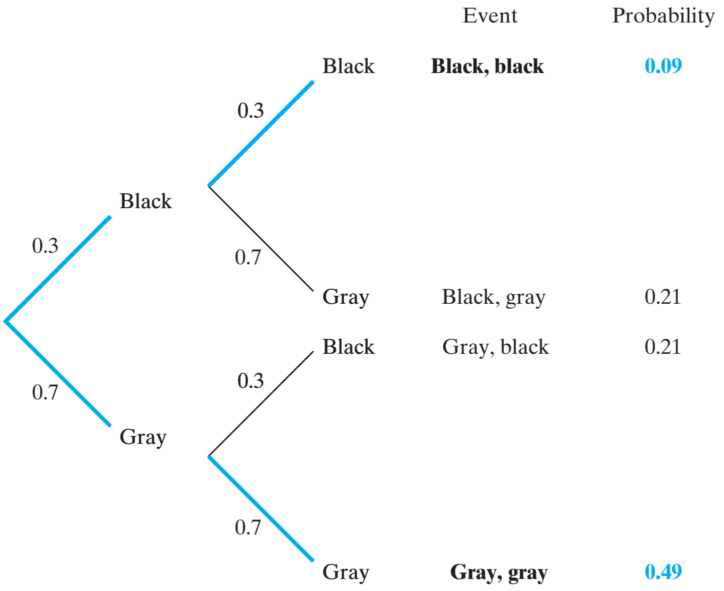
Combination of Probabilities

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

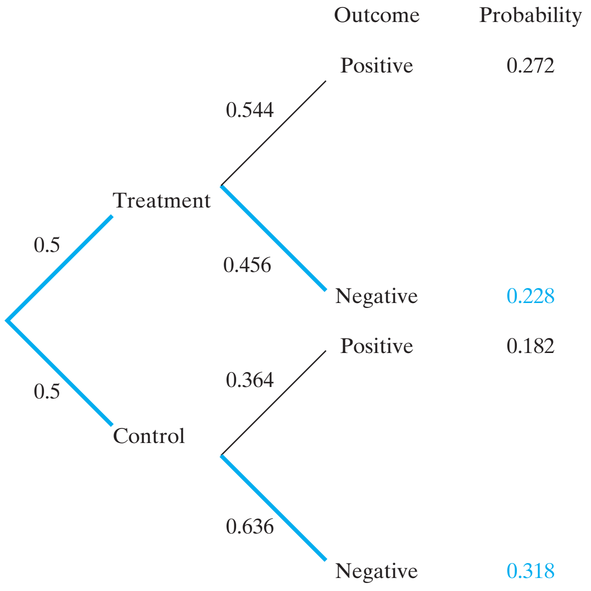


Figure 3.2.4 Probability tree for nitric oxide example

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.

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

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

Example 3.2.10

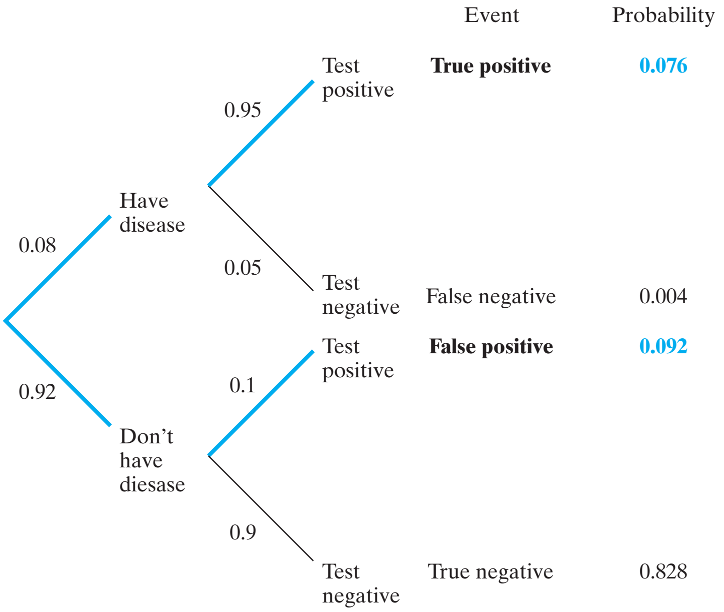
Medical Testing Suppose a medical test is conducted on someone to try to determine whether or not the person has a particular disease. If the test indicates that the disease is present, we say the person has "tested positive." If the test indicates that the disease is not present, we say the person has "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 person testing positive. Thus, the probability of a randomly chosen person testing positive is

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

Figure 3.2.5 Probability tree for medical testing example



Example 3.2.11 False Positives Consider the medical testing scenario of Example 3.2.10. If someone tests positive, what is the chance the person really has the disease? In Example 3.2.10 we found that 0.168 (16.8%) of the population will test positive, so if 1,000 persons are tested, we would expect 168 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. ■

Exercises 3.2.1–3.2.7

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 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.

3.2.3 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?

3.2.4 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?

3.2.5 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?

3.2.6

(a) Consider the setting of Exercise 3.2.5, part (a). Suppose that a woman tests positive. What is the probability that she really is pregnant?

(b) Consider the setting of Exercise 3.2.5, part (b). Suppose that a woman tests positive. What is the probability that she really is pregnant?

3.2.7 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 possible events is $E_1, E_2, \ldots, E_k$, then $\Sigma_{i=1}^{k}\Pr\{E_i\} = 1$.

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.[3] 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) $\Pr\{\text{O}\} + \Pr\{\text{A}\} + \Pr\{\text{B}\} + \Pr\{\text{AB}\} = 0.44 + 0.42 + 0.10 + 0.04 = 1$.

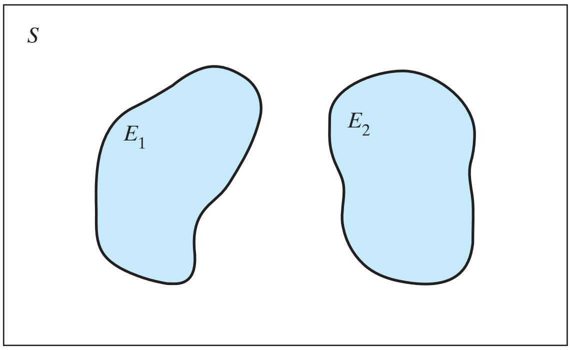


Figure 3.3.1 Venn diagram showing two disjoint events

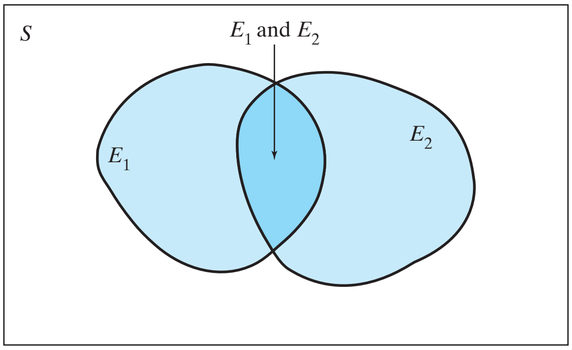


Figure 3.3.2 Venn diagram showing union (total shaded area) and intersection (middle area) of two events

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

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
$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$.

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.[4]

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

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:

Defintion 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.[5] 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,

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

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.[6] 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 either 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.[7] The distribution is represented by histograms with class widths equal to (a) 10 and (b) 5, and by (c) a smooth curve.

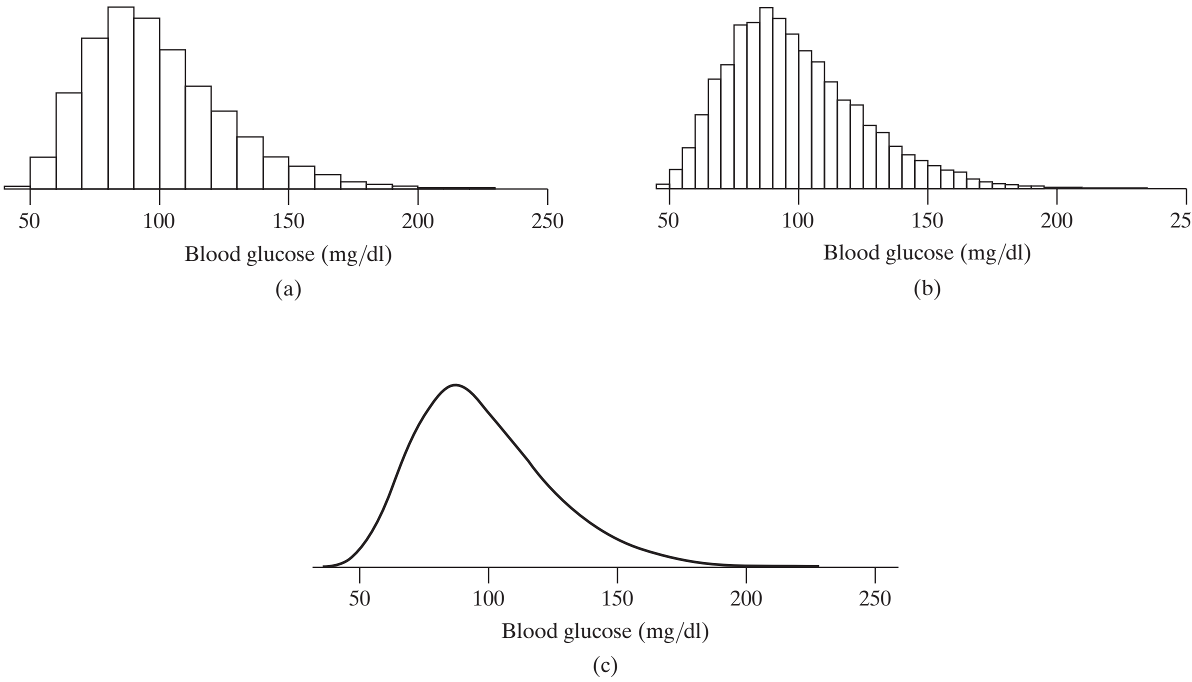


Figure 3.4.1 Different representations of the distribution of blood glucose levels in a population of women

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,

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

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.

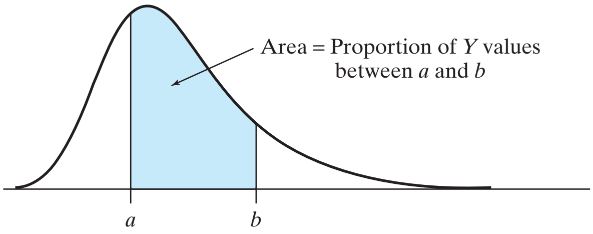


Figure 3.4.2 Interpretation of area under a density curve

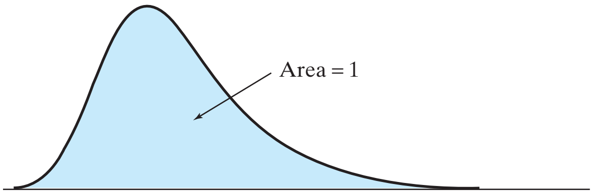
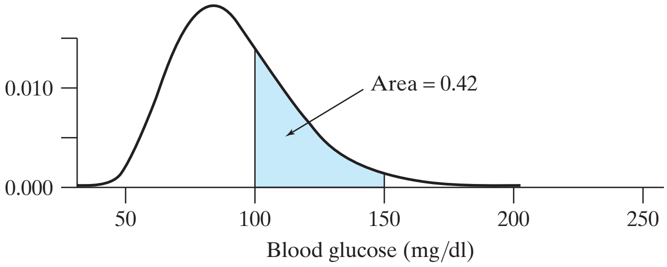


Figure 3.4.3 The area under an entire density curve must be 1

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 (such as 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, $\Pr\{100 \leq \text{glucose level} \leq 150\} = 0.42$.

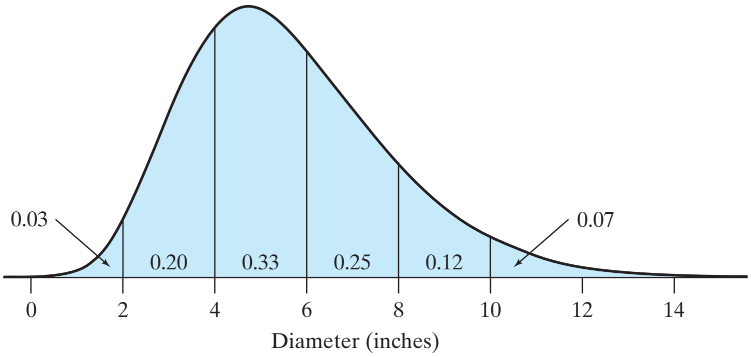
We are modeling blood glucose level as being a continuous variable, which means that $\Pr\{\text{glucose level} = 100\} = 0$, as we noted above. Thus,

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

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 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.[8] 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 3.4.1–3.4.4

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?

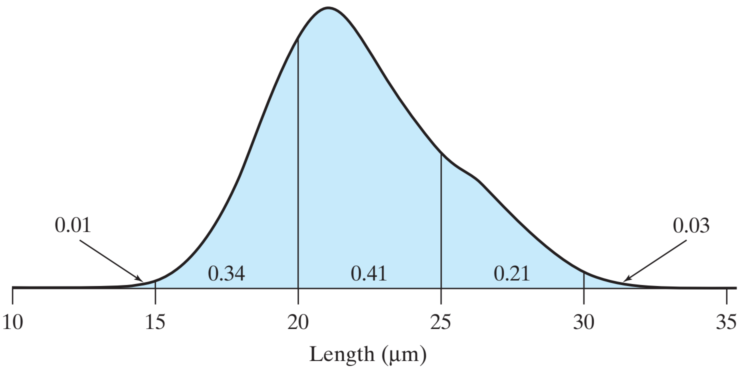
3.4.2 Consider the diameter of a Douglas fir tree drawn at random from the population that is represented by the density curve shown in Figure 3.4.5. Find

(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.[9]



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. 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, ... 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

$$\Pr\{2 \leq Y \leq 5\} = 0.52$$

■

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

$$\Pr\{65.2 \leq Y \leq 70.4\} = 0.46$$

■

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

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

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

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

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 (for example, 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 (for example, 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 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 the square root to get the standard deviation of the sum (or difference).

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.[10]

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 that 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 that 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 3.5.1–3.5.8

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.[11]

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 A group of college students were surveyed to learn how many times they had visited a dentist in the previous year.[12] 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.

3.5.8 Calculate the standard deviation, σ_Y, of the random variable Y from Exercise 3.5.7.

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.[13] 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**:

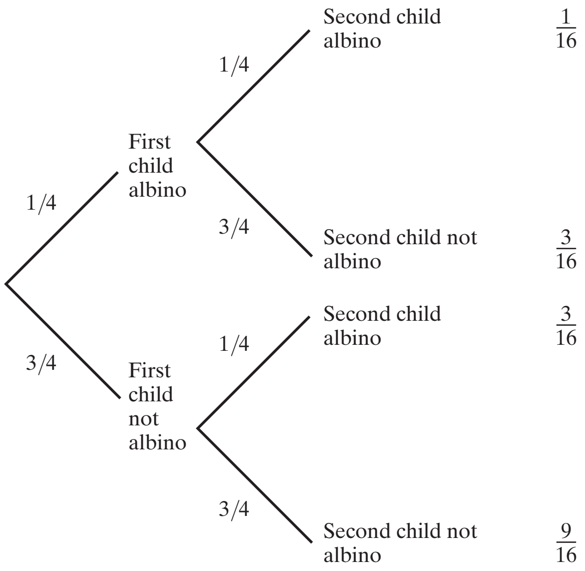


Figure 3.6.1 Probability tree for albinism among two children of carriers of the gene for albinism

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	$\frac{1}{16}$
		1

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.

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	$1p^5(1 - p)^0$

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 2 mutants and 3 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
		1.00

Thus, $\Pr\{Y = 3\} \approx 0.34$. This means that about 34% of random samples of size 5 will contain two mutants and three nonmutants.

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. ■

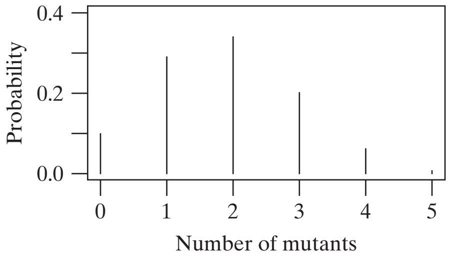


Figure 3.6.2 Binomial distribution with $n = 5$ and $p = 0.37$

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.

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.

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 5.5.

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 G	0	0.49
One B, one G	1	0.42
Both B	2	0.09
		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
	1

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 $\Pr\{Y \neq 6\}$ 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 (that is, 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 (for instance, 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 that all five children would get chickenpox. ■

Exercises 3.6.1–3.6.10

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.[14] 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 the United States, 42% 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) $\Pr\{0 \leq Y \leq 2\}$.

(e) $\Pr\{0 < Y \leq 2\}$.

3.6.3 A certain drug treatment cures 90% of cases of hookworm in children.[15] 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.

3.6.4 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.[16] 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%.

3.6.5 Consider taking a sample of size 10 from the snail population in Exercise 3.6.4.

(a) What is the mean number of streaked-shelled snails?

(b) What is the standard deviation of the number of streaked-shelled snails?

3.6.6 The sex ratio of newborn human infants is about 105 males : 100 females.[17] 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?

3.6.7 Construct a binomial setting (different from any examples presented in this book) and a problem for which the following is the answer: ${}_7C_3(0.8)^3(0.2)^5$.

3.6.8 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.[18] 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).]

3.6.9 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."]

3.6.10 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).[19] 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.[20]

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.

Table 3.7.1 Sex ratios in 6,115 families with twelve 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
		6,115

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

$$(3)(0) + (24)(1) + \cdots + (12)(7) = 38{,}100 \text{ boys}$$

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 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?

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	854.3	−
9	3	478	410.0	+
10	2	181	132.8	+
11	1	45	26.1	+
12	0	7	2.3	+
		6,115	6,115.0	

*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.

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 frequency	Expected frequency	Sign of (Obs. − Exp.)
Boys	Girls			
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	+
		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 the mechanism causing the variation has not yet 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. ■

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.

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
		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).[21]

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
		310

(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?

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.[22]

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
		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.10

3.S.1 In the United States, 10% of adolescent girls have iron deficiency.[23] Suppose two adolescent girls are chosen at random. Find the probability that

(a) both girls have iron deficiency.

(b) one girl has iron deficiency and the other does 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.[24] 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
	100

Suppose that a square is chosen at random, and let Y be the number of centipedes in the chosen square. Find

(a) $\Pr\{Y = 1\}$ (b) $\Pr\{Y \geq 2\}$

3.S.3 Refer to the distribution of centipedes given in Exercise 3.S.2. Suppose five squares are chosen at random. Find the probability that three 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 five squares are chosen at random. Find the expected value (i.e., the mean) of the number of squares that contain at least one centipede.

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.[25] Consider a large number of such matings, each producing a litter of five offspring. What percentage of the litters will consist of

(a) two wavy-haired and three straight-haired offspring?

(b) three 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 kidney damage in 1% of patients. Suppose the drug is to be tested on 50 patients. Find the probability that

(a) none of the patients will experience kidney damage.

(b) one or more of the patients will experience kidney 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 kidney 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 0.95?

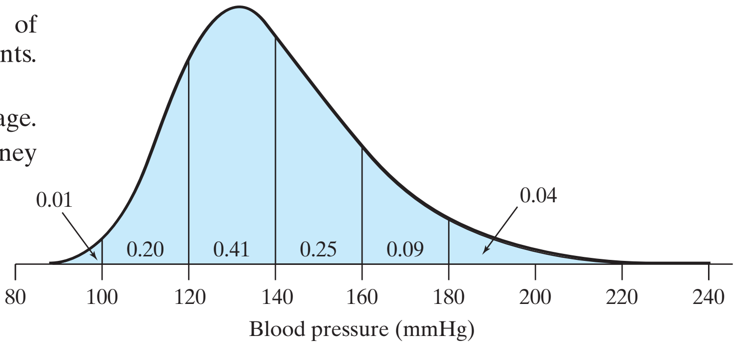
3.S.8 To study people's ability to deceive lie detectors, researchers sometimes use the "guilty knowledge" technique.[26] Certain subjects memorize six common words; other subjects memorize no words. Each subject is then tested on a polygraph machine (lie detector), as follows. The experimenter reads, 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 four 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.[27] 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) $\Pr\{120 < Y < 160\}$.

(b) $\Pr\{Y < 120\}$.

(c) $\Pr\{Y > 140\}$.



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 140 mm Hg.

(b) three have blood pressures higher than 140, and one has blood pressure 140 or less.