CHAPTER 6

Probability

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Proportions (math review, Appendix A)
  - Fractions
  - Decimals
  - Percentages
- Basic algebra (math review, Appendix A)
  - *z*-Scores (Chapter 5)

Preview

Background: If you open a dictionary and randomly pick one word, which are you more likely to select:

1. A word beginning with the letter K?
2. A word with a *K* as its third letter?

If you think about this question and answer honestly, you probably will decide that words beginning with a *K* are more probable.

A similar question was asked a group of participants in an experiment reported by Tversky and Kahneman (1973). Their participants estimated that words beginning with *K* are twice as likely as words with a *K* as the third letter. In truth, the relationship is just the opposite. There are more than twice as many words with a *K* in the third position as there are words beginning with a *K*. How can people be so wrong? Do they completely misunderstand probability?

When you were deciding which type of *K* words are more likely, you probably searched your memory and tried to estimate which words are more common. How many words can you think of that start with the letter *K*? How many words can you think of that have a *K* as the third letter? Because you have had years of practice alphabetizing words according to their first letter, you should find it much easier to search your memory for words beginning with a *K* than to search for words with a *K* in the third position. Consequently, you are likely to conclude that first-letter *K* words are more common.

If you had searched for words in a dictionary (instead of those in your memory), you would have found more third-letter *K* words, and you would have concluded (correctly) that these words are more common.

The Problem: If you open a dictionary and randomly pick one word, it is impossible to predict exactly which word you will get. In the same way, when researchers recruit people to participate in research studies, it is impossible to predict exactly which individuals will be obtained.

The Solution: Although it is impossible to predict exactly which word will be picked from a dictionary, or which person will participate in a research study, you can use *probability* to demonstrate that some outcomes are more likely than others. For example, it is more likely that you will pick a third-letter *K* word than a first-letter *K* word. Similarly, it is more likely that you will obtain a person with an IQ around 100 than a person with an IQ around 150.

6.1 INTRODUCTION TO PROBABILITY

In Chapter 1, we introduced the idea that research studies begin with a general question about an entire population, but the actual research is conducted using a sample. In this situation, the role of inferential statistics is to use the sample data as the basis for answering questions about the population. To accomplish this goal, inferential procedures are typically built around the concept of probability. Specifically, the relationships between samples and populations are usually defined in terms of probability.

Suppose, for example, that you are selecting a single marble from a jar that contains 50 black and 50 white marbles. (In this example, the jar of marbles is the *population* and the single marble to be selected is the *sample*.) Although you cannot guarantee the exact outcome of your sample, it is possible to talk about the potential outcomes in terms of probabilities. In this case, you have a 50-50 chance of getting either color. Now consider another jar (population) that has 90 black and only 10 white marbles. Again, you cannot predict the exact outcome of a sample, but now you know that the sample probably will be a black marble. By knowing the makeup of a population, we can determine the probability of obtaining specific samples. In this way, probability gives us a connection between populations and samples, and this connection is the foundation for the inferential statistics that are presented in the chapters that follow.

You may have noticed that the preceding examples begin with a population and then use probability to describe the samples that could be obtained. This is exactly backward from what we want to do with inferential statistics. Remember that the goal of inferential statistics is to begin with a sample and then answer a general question about the population. We reach this goal in a two-stage process. In the first stage, we develop probability as a bridge from populations to samples. This stage involves identifying the types of samples that probably would be obtained from a specific population. Once this bridge is established, we simply reverse the probability rules to allow us to move from samples to populations (Figure 6.1). The process of reversing the probability relationship can be demonstrated by considering again the two jars of marbles we looked at earlier. (Jar 1 has 50 black and 50 white marbles; jar 2 has 90 black and only 10 white marbles.) This time, suppose you are blindfolded when the sample is selected, so you do not know which jar is being used. Your task is to look at the sample that you obtain and then decide which jar is most likely. If you select a sample of $n = 4$ marbles and all are black, which jar would you choose? It should be clear that it would be relatively unlikely (low probability) to obtain this sample from jar 1; in four draws, you almost certainly would get at least 1 white marble. On the other hand, this sample would have a high probability of coming from jar 2, where nearly all of the marbles are black. Your decision, therefore, is that the sample probably came from jar 2. Note that you now are using the sample to make an inference about the population.

DEFINING PROBABILITY

Probability is a huge topic that extends far beyond the limits of introductory statistics, and we do not attempt to examine it all here. Instead, we concentrate on the few concepts and definitions that are needed for an introduction to inferential statistics. We begin with a relatively simple definition of probability.

DEFINITION

For a situation in which several different outcomes are possible, the **probability** for any specific outcome is defined as a fraction or a proportion of all the possible outcomes. If the possible outcomes are identified as A, B, C, D, and so on, then

$$\text{probability of } A = \frac{\text{number of outcomes classified as } A}{\text{total number of possible outcomes}}$$

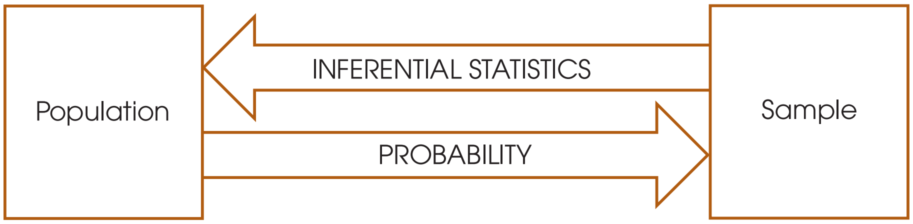


FIGURE 6.1

The role of probability in inferential statistics. Probability is used to predict what kind of samples are likely to be obtained from a population. Thus, probability establishes a connection between samples and populations. Inferential statistics rely on this connection when they use sample data as the basis for making conclusions about populations.

For example, if you are selecting a card from a complete deck, there are 52 possible outcomes. The probability of selecting the king of hearts is $p = \frac{1}{52}$. The probability of selecting an ace is $p = \frac{4}{52}$ because there are 4 aces in the deck.

To simplify the discussion of probability, we use a notation system that eliminates a lot of the words. The probability of a specific outcome is expressed with a p (for probability) followed by the specific outcome in parentheses. For example, the probability of selecting a king from a deck of cards is written as p(king). The probability of obtaining heads for a coin toss is written as p(heads).

Note that probability is defined as a proportion, or a part of the whole. This definition makes it possible to restate any probability problem as a proportion problem. For example, the probability problem "What is the probability of selecting a king from a deck of cards?" can be restated as "What proportion of the whole deck consists of kings?" In each case, the answer is $\frac{4}{52}$, or "4 out of 52." This translation from probability to proportion may seem trivial now, but it is a great aid when the probability problems become more complex. In most situations, we are concerned with the probability of obtaining a particular sample from a population. The terminology of *sample* and *population* do change the basic definition of probability. For example, the whole deck of cards can be considered as a population, and the single card we select is the sample.

Probability values The definition we are using identifies probability as a fraction or a proportion. If you work directly from this definition, the probability values you obtain are expressed as fractions. For example, if you are selecting a card at random,

$$p(\text{spade}) = \frac{13}{52} = \frac{1}{4}$$

Of if you are tossing a coin,

$$p(\text{heads}) = \frac{1}{2}$$

If you are unsure how to convert from fractions to decimals or percentages, you should review the section on proportions in the math review, Appendix A.

You should be aware that these fractions can be expressed equally well as either decimals or percentages:

$$p=\frac{1}{4}=0.25=25\%$$

$$p=\frac{1}{2}=0.50=50\%$$

By convention, probability values most often are expressed as decimal values. But you should realize that any of these three forms is acceptable.

You also should note that all of the possible probability values are contained in a limited range. At one extreme, when an event never occurs, the probability is zero, or 0% (Box 6.1). At the other extreme, when an event always occurs, the probability is 1, or 100%. Thus, all probability values are contained in a range from 0 to 1. For example, suppose that you have a jar containing 10 white marbles. The probability of randomly selecting a black marble is

$$p(\text{black}) = \frac{0}{10} = 0$$

The probability of selecting a white marble is

$$p(\text{white}) = \frac{10}{10} = 1$$

RANDOM SAMPLING

For the preceding definition of probability to be accurate, it is necessary that the outcomes be obtained by a process called *random sampling*.

DEFINITION

A **random sample** requires that each individual in the population has an *equal chance* of being selected.

A second requirement, necessary for many statistical formulas, states that if more than one individual is being selected, the probabilities must *stay constant* from one selection to the next. Adding this second requirement produces what is called *independent random sampling*. The term *independent* refers to the fact that the probability of selecting any particular individual is independent of those individuals who have already been selected for the sample. For example, the probability that you will be selected is constant and does not change even when other individuals are selected before you are.

DEFINITION

An **independent random sample** requires that each individual has an equal chance of being selected and that the probability of being selected stays constant from one selection to the next if more than one individual is selected.

Because independent random sample is a required component for most statistical applications, we always assume that this is the sampling method being used. To simplify discussion, we typically omit the word “independent” and simply refer to this sampling technique as *random sampling*. However, you should always assume that both requirements (equal chance and constant probability) are part of the process.

Each of the two requirements for random sampling has some interesting consequences. The first assures that there is no bias in the selection process. For a population with N individuals, each individual must have the same probability, $p = \frac{1}{N}$, of being selected. This means, for example, that you would not get a random sample of people in your city by selecting names from a yacht-club membership list. Similarly, you would not get a random sample of college students by selecting individuals from your psychology classes. You also should note that the first requirement of random sampling prohibits you from applying the definition of probability to situations in which the possible outcomes are not equally likely. Consider, for example, the question of whether you will win a million dollars in the lottery tomorrow. There are only two possible alternatives.

1. You will win.
2. You will not win.

According to our simple definition, the probability of winning would be one out of two, or $p = \frac{1}{2}$. However, the two alternatives are not equally likely, so the simple definition of probability does not apply.

The second requirement also is more interesting than may be apparent at first glance. Consider, for example, the selection of $n = 2$ cards from a complete deck. For the first draw, the probability of obtaining the jack of diamonds is

$$p(\text{jack of diamonds}) = \frac{1}{52}$$

After selecting one card for the sample, you are ready to draw the second card. What is the probability of obtaining the jack of diamonds this time? Assuming that you still are holding the first card, there are two possibilities:

$$p(\text{jack of diamonds}) = \frac{1}{51} \text{ if the first card was not the jack of diamonds}$$

or

$$p(\text{jack of diamonds}) = 0 \text{ if the first card was the jack of diamonds}$$

In either case, the probability is different from its value for the first draw. This contradicts the requirement for random sampling, which says that the probability must stay constant. To keep the probabilities from changing from one selection to the next, it is necessary to return each individual to the population before you make the next selection. This process is called *sampling with replacement*. The second requirement for random samples (constant probability) demands that you sample with replacement.

(*Note:* We are using a definition of random sampling that requires equal chance of selection and constant probabilities. This kind of sampling is also known as independent random sampling, and often is called *random sampling with replacement*. Many of the statistics we encounter later are founded on this kind of sampling. However, you should realize that other definitions exist for the concept of random sampling. In particular, it is very common to define random sampling without the requirement of constant probabilities—that is, *random sampling without replacement*. In addition, there are many different sampling techniques that are used when researchers are selecting individuals to participate in research studies.)

PROBABILITY AND FREQUENCY DISTRIBUTIONS

The situations in which we are concerned with probability usually involve a population of scores that can be displayed in a frequency distribution graph. If you think of the graph as representing the entire population, then different proportions of the graph represent different proportions of the population. Because probabilities and proportions are equivalent, a particular proportion of the graph corresponds to a particular probability in the population. Thus, whenever a population is presented in a frequency distribution graph, it is possible to represent probabilities as proportions of the graph. The relationship between graphs and probabilities is demonstrated in the following example.

EXAMPLE 6.1

We use a very simple population that contains only $N = 10$ scores with values 1, 1, 2, 3, 3, 4, 4, 4, 5, 6. This population is shown in the frequency distribution graph in Figure 6.2. If you are taking a random sample of $n = 1$ score from this population,

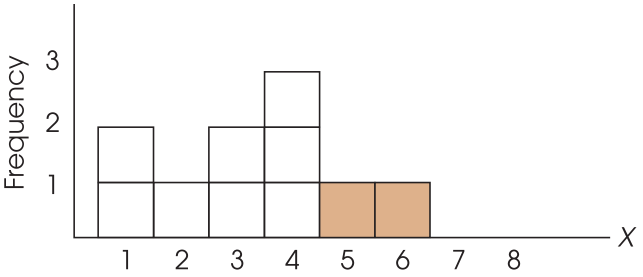


FIGURE 6.2

A frequency distribution histogram for a population that consists of $N = 10$ scores. The shaded part of the figure indicates the portion of the whole population that corresponds to scores greater than $X = 4$. The shaded portion is two-tenths ($p = \frac{2}{10}$) of the whole distribution.

what is the probability of obtaining an individual with a score greater than 4? In probability notation,

$$p(X > 4) = ?$$

Using the definition of probability, there are 2 scores that meet this criterion out of the total group of $N = 10$ scores, so the answer would be $p = \frac{2}{10}$. This answer can be obtained directly from the frequency distribution graph if you recall that probability and proportion measure the same thing. Looking at the graph (see Figure 6.2), what proportion of the population consists of scores greater than 4? The answer is the shaded part of the distribution—that is, 2 squares out of the total of 10 squares in the distribution. Notice that we now are defining probability as a proportion of *area* in the frequency distribution graph. This provides a very concrete and graphic way of representing probability.

Using the same population once again, what is the probability of selecting an individual with a score less than 5? In symbols,

$$p(X < 5) = ?$$

Going directly to the distribution in Figure 6.2, we now want to know what part of the graph is not shaded. The unshaded portion consists of 8 out of the 10 blocks (eight-tenths of the area of the graph), so the answer is $p = \frac{8}{10}$.

LEARNING CHECK

1. A survey of the students in a psychology class revealed that there were 19 females and 8 males. Of the 19 females, only 4 had no brothers or sisters, and 3 of the males were also the only child in the household. If a student is randomly selected from this class,
   - **a.** What is the probability of obtaining a male?
   - **b.** What is the probability of selecting a student who has at least one brother or sister?
   - **c.** What is the probability of selecting a female who has no siblings?

2. A jar contains 10 red marbles and 30 blue marbles.
   - **a.** If you randomly select 1 marble from the jar, what is the probability of obtaining a red marble?
   - **b.** If you take a *random sample* of $n = 3$ marbles from the jar and the first two marbles are both blue, what is the probability that the third marble will be red?

3. Suppose that you are going to select a random sample of $n = 1$ score from the distribution in Figure 6.2. Find the following probabilities:
   - **a.** $p(X > 2)$
   - **b.** $p(X > 5)$
   - **c.** $p(X < 3)$

ANSWERS

1. **a.** $p = \frac{8}{27}$
   b. $p = \frac{20}{27}$
   c. $p = \frac{4}{27}$

2. **a.** $p = \frac{10}{40} = 0.25$

b. $p = \frac{10}{40} = 0.25$. Remember that random sampling requires sampling with replacement.

3. **a.** $p = \frac{7}{10} = 0.70$

b. $p = \frac{1}{10} = 0.10$

c. $p = \frac{3}{10} = 0.30$

6.2 PROBABILITY AND THE NORMAL DISTRIBUTION

The normal distribution was first introduced in Chapter 2 as an example of a commonly occurring shape for population distributions. An example of a normal distribution is shown in Figure 6.3.

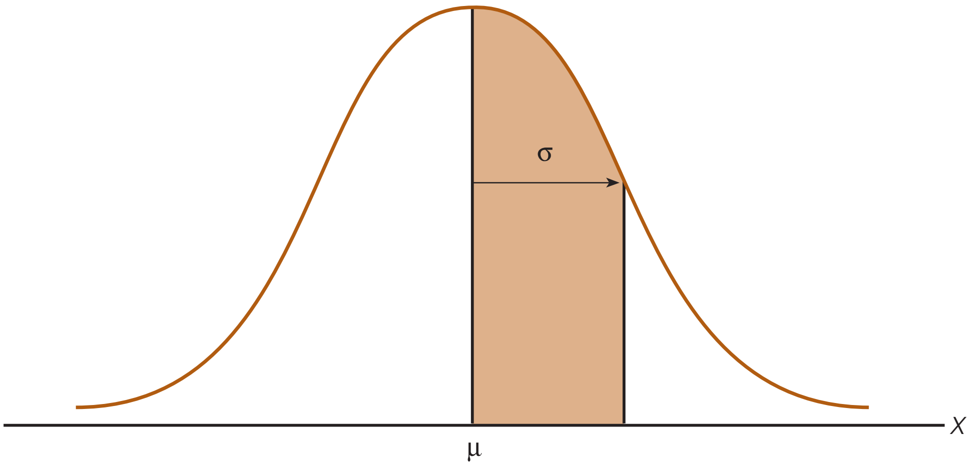
Note that the normal distribution is symmetrical, with the highest frequency in the middle and frequencies tapering off as you move toward either extreme. Although the exact shape for the normal distribution is defined by an equation (see Figure 6.3), the normal shape can also be described by the proportions of area contained in each section of the distribution. Statisticians often identify sections of a normal distribution by using *z*-scores. Figure 6.4 shows a normal distribution with several sections marked in *z*-score units. You should recall that *z*-scores measure positions in a distribution in terms of standard deviations from the mean. (Thus, $z = +1$ is 1 standard deviation above the mean, $z = +2$ is 2 standard deviations above the mean, and so on.) The graph shows the percentage of scores that fall in each of these sections. For example, the section between the mean ($z = 0$) and the point that is 1 standard deviation above the mean ($z = 1$) contains 34.13% of the scores. Similarly, 13.59% of the scores are located in the section between

FIGURE 6.3

The normal distribution. The exact shape of the normal distribution is specified by an equation relating each *X* value (score) with each *Y* value (frequency). The equation is

$$Y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$$

(π and e are mathematical constants). In simpler terms, the normal distribution is symmetrical with a single mode in the middle. The frequency tapers off as you move farther from the middle in either direction.



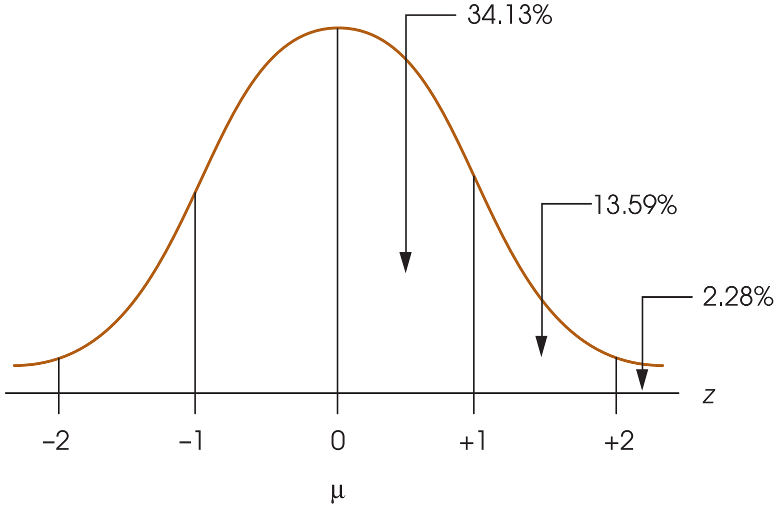


FIGURE 6.4
The normal distribution following a *z*-score transformation.

1 and 2 standard deviations above the mean. In this way it is possible to define a normal distribution in terms of its proportions; that is, a distribution is normal if and only if it has all the right proportions.

There are two additional points to be made about the distribution shown in Figure 6.4. First, you should realize that the sections on the left side of the distribution have exactly the same areas as the corresponding sections on the right side because the normal distribution is symmetrical. Second, because the locations in the distribution are identified by *z*-scores, the percentages shown in the figure apply to *any normal distribution* regardless of the values for the mean and the standard deviation. Remember: When any distribution is transformed into *z*-scores, the mean becomes zero and the standard deviation becomes one.

Because the normal distribution is a good model for many naturally occurring distributions and because this shape is guaranteed in some circumstances (as we see in Chapter 7), we devote considerable attention to this particular distribution. The process of answering probability questions about a normal distribution is introduced in the following example.

EXAMPLE 6.2

The population distribution of SAT scores is normal with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$. Given this information about the population and the known proportions for a normal distribution (see Figure 6.4), we can determine the probabilities associated with specific samples. For example, what is the probability of randomly selecting an individual from this population who has an SAT score greater than 700?

Restating this question in probability notation, we get

$$p(X > 700) = ?$$

We follow a step-by-step process to find the answer to this question.

1. First, the probability question is translated into a proportion question: Out of all possible SAT scores, what proportion is greater than 700?

2. The set of "all possible SAT scores" is simply the population distribution. This population is shown in Figure 6.5. The mean is $\mu = 500$, so the score $X = 700$ is to the right of the mean. Because we are interested in all scores greater than 700, we shade in the area to the right of 700. This area represents the proportion we are trying to determine.
3. Identify the exact position of $X = 700$ by computing a z-score. For this example,

$$z = \frac{X - \mu}{\sigma} = \frac{700 - 500}{100} = \frac{200}{100} = 2.00$$

That is, an SAT score of $X = 700$ is exactly 2 standard deviations above the mean and corresponds to a z-score of $z = +2.00$. We have also located this z-score in Figure 6.5.

4. The proportion we are trying to determine may now be expressed in terms of its z-score:

$$p(z > 2.00) = ?$$

According to the proportions shown in Figure 6.4, all normal distributions, regardless of the values for μ and σ, have 2.28% of the scores in the tail beyond $z = +2.00$. Thus, for the population of SAT scores,

$$p(X > 700) = p(z > +2.00) = 2.28\%$$

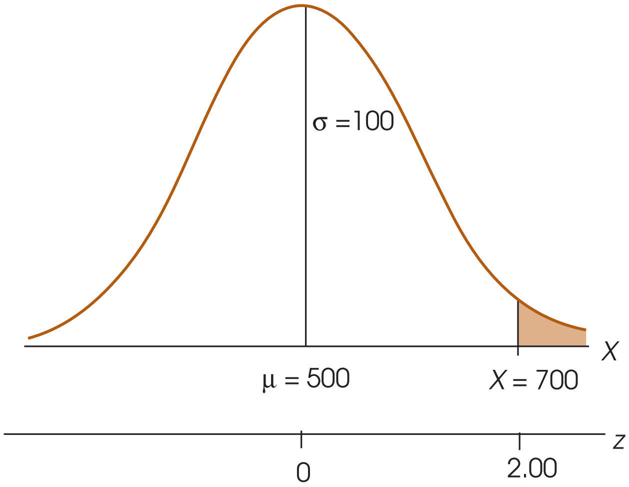
THE UNIT NORMAL TABLE

Before we attempt any more probability questions, we must introduce a more useful tool than the graph of the normal distribution shown in Figure 6.4. The graph shows proportions for only a few selected z-score values. A more complete listing of z-scores and proportions is provided in the *unit normal table*. This table lists proportions of the normal distribution for a full range of possible z-score values.

The complete unit normal table is provided in Appendix B Table B.1, and part of the table is reproduced in Figure 6.6. Notice that the table is structured in a four-column

FIGURE 6.5

The distribution of SAT scores described in Example 6.2.



(A) *z*	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and *z*
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331

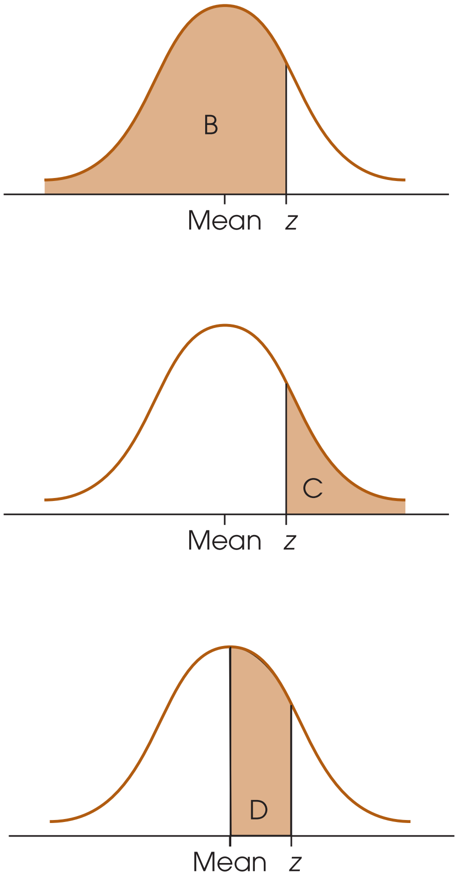


FIGURE 6.6

A portion of the unit normal table. This table lists proportions of the normal distribution corresponding to each *z*-score value. Column A of the table lists *z*-scores. Column B lists the proportion in the body of the normal distribution up to the *z*-score value. Column C lists the proportion of the normal distribution that is located in the tail of the distribution beyond the *z*-score value. Column D lists the proportion between the mean and the *z*-score value.

format. The first column (A) lists *z*-score values corresponding to different positions in a normal distribution. If you imagine a vertical line drawn through a normal distribution, then the exact location of the line can be described by one of the *z*-score values listed in column A. You should also realize that a vertical line separates the distribution into two sections: a larger section called the *body* and a smaller section called the *tail*. Columns B and C in the table identify the proportion of the distribution in each of the two sections. Column B presents the proportion in the body (the larger portion), and column C presents the proportion in the tail. Finally, we have added a fourth column, column D, that identifies the proportion of the distribution that is located *between* the mean and the *z*-score.

We use the distribution in Figure 6.7(a) to help introduce the unit normal table. The figure shows a normal distribution with a vertical line drawn at $z = +0.25$. Using the portion of the table shown in Figure 6.6, find the row in the table that contains $z = 0.25$ in column A. Reading across the row, you should find that the line drawn $z = +\,0.25$

separates the distribution into two sections with the larger section containing 0.5987 (59.87%) of the distribution and the smaller section containing 0.4013 (40.13%) of the distribution. Also, there is exactly 0.0987 (9.87%) of the distribution between the mean and $z = +0.25$.

To make full use of the unit normal table, there are a few facts to keep in mind:

1. The *body* always corresponds to the larger part of the distribution whether it is on the right-hand side or the left-hand side. Similarly, the *tail* is always the smaller section whether it is on the right or the left.
2. Because the normal distribution is symmetrical, the proportions on the right-hand side are exactly the same as the corresponding proportions on the left-hand side. Earlier, for example, we used the unit normal table to obtain proportions for $z = +0.25$. Figure 6.7(b) shows the same proportions for $z = -0.25$. For a negative *z*-score, however, notice that the tail of the distribution is on the left side and the body is on the right. For a positive *z*-score [Figure 6.7(a)], the positions are reversed. However, the proportions in each section are exactly the same, with 0.55987 in the body and 0.4013 in the tail. Once again, the table does not list negative *z*-score values. To find proportions for negative *z*-scores, you must look up the corresponding proportions for the positive value of *z*.
3. Although the *z*-score values change signs (+ and –) from one side to the other, the proportions are always positive. Thus, column C in the table always lists the proportion in the tail whether it is the right-hand tail or the left-hand tail.

PROBABILITIES, PROPORTIONS, AND *Z*-SCORES

The unit normal table lists relationships between *z*-score locations and proportions in a normal distribution. For any *z*-score location, you can use the table to look up the corresponding proportions. Similarly, if you know the proportions, you can use the table to find the specific *z*-score location. Because we have defined probability as equivalent to proportion, you can also use the unit normal table to look up probabilities for normal distributions. The following examples demonstrate a variety of different ways that the unit normal table can be used.

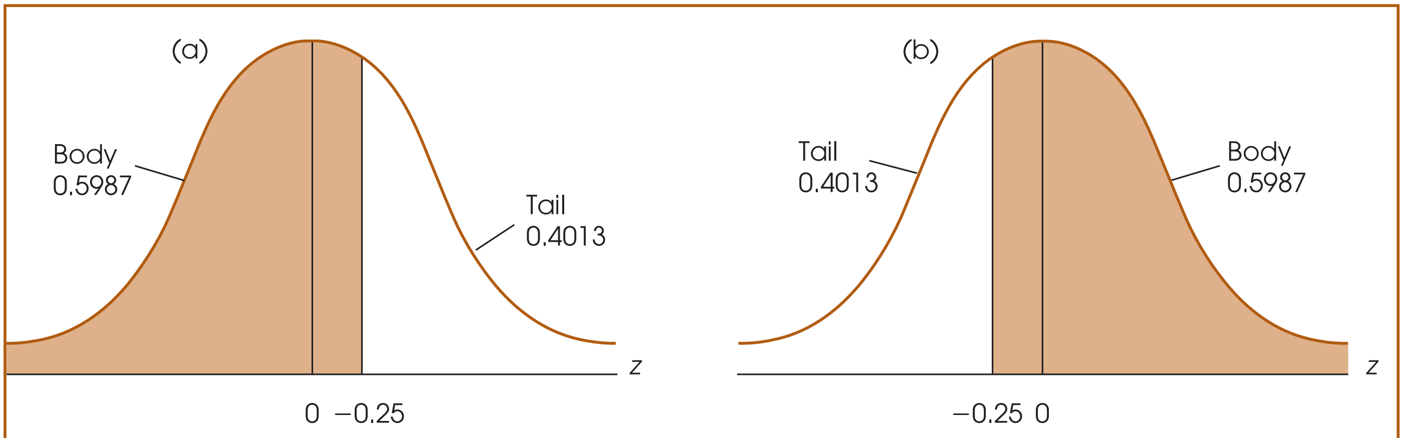


FIGURE 6.7

Proportions of a normal distribution corresponding to $z = +0.25$ (a) and –0.25 (b).

Finding proportions or probabilities for specific *z*-score values For each of the following examples, we begin with a specific *z*-score value and then use the unit normal table to find probabilities or proportions associated with the *z*-score.

EXAMPLE 6.3A

What proportion of the normal distribution corresponds to *z*-score values greater than $z = 1.00$? First, you should sketch the distribution and shade in the area you are trying to determine. This is shown in Figure 6.8(a). In this case, the shaded portion is the tail of the distribution beyond $z = 1.00$. To find this shaded area, you simply look for $z = 1.00$ in column A to find the appropriate row in the unit normal table. Then scan across the row to column C (tail) to find the proportion. Using the table in Appendix B, you should find that the answer is 0.1587.

You also should notice that this same problem could have been phrased as a probability question. Specifically, we could have asked, "For a normal distribution, what is the probability of selecting a *z*-score value greater than $z = +1.00$?" Again, the answer is $p(z > 1.00) = 0.1587$ (or 15.87%).

EXAMPLE 6.3B

For a normal distribution, what is the probability of selecting a *z*-score less than $z = 1.50$? In symbols, $p(z < 1.50) = ?$ Our goal is to determine what proportion of the normal distribution corresponds to *z*-scores less than 1.50. A normal distribution is shown in Figure 6.8(b) and $z = 1.50$ is marked in the distribution. Notice that we have shaded all the values to the left of (less than) $z = 1.50$. This is the portion we are trying to find. Clearly the shaded portion is more than 50%, so it corresponds to the body of the distribution. Therefore, find $z = 1.50$ in column A of the unit normal table and read across the row to obtain the proportion from column B. The answer is $p(z < 1.50) = 0.9332$ (or 93.32%).

EXAMPLE 6.3C

Moving to the left on the *X*-axis results in smaller *X* values and smaller *z*-scores. Thus, a *z*-score of –3.00 reflects a smaller value than a *z*-score of –1.

Many problems require that you find proportions for negative *z*-scores. For example, what proportion of the normal distribution is contained in the tail beyond $z = –0.50$? That is, $p(z < –0.50)$. This portion has been shaded in Figure 6.8(c). To answer questions with negative *z*-scores, simply remember that the normal distribution is symmetrical with a *z*-score of zero at the mean, positive values to the right, and negative values to the left. The proportion in the left tail beyond $z = –0.50$ is identical to the proportion

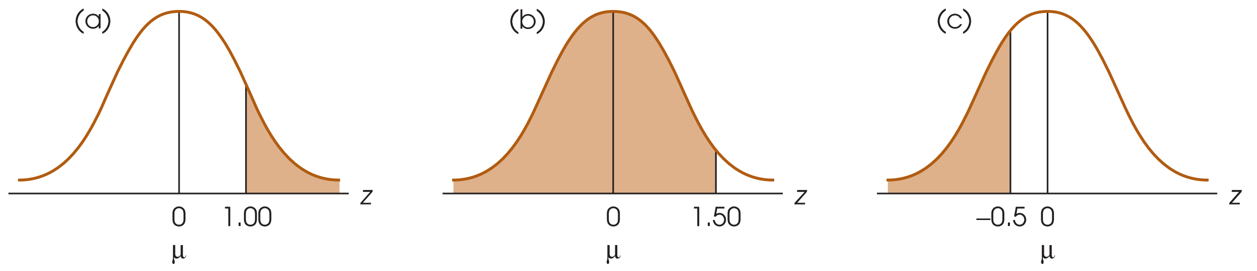


FIGURE 6.8
The distribution for Examples 6.3A to 6.3C.

in the right tail beyond $z = +0.50$. To find this proportion, look up $z = 0.50$ in column A, and read across the row to find the proportion in column C (tail). You should get an answer of 0.3085 (30.85%).

Finding the *z*-score location that corresponds to specific proportions The preceding examples all involved using a *z*-score value in column A to look up proportions in column B or C. You should realize, however, that the table also allows you to begin with a known proportion and then look up the corresponding *z*-score. The following examples demonstrate this process.

EXAMPLE 6.4A

For a normal distribution, what *z*-score separates the top 10% from the remainder of the distribution? To answer this question, we have sketched a normal distribution [Figure 6.9(a)] and drawn a vertical line that separates the highest 10% (approximately) from the rest. The problem is to locate the exact position of this line. For this distribution, we know that the tail contains 0.1000 (10%) and the body contains 0.9000 (90%). To find the *z*-score value, you simply locate the row in the unit normal table that has 0.1000 in column C or 0.9000 in column B. For example, you can scan down the values in column C (tail) until you find a proportion of 0.1000. Note that you probably will not find the exact proportion, but you can use the closest value listed in the table. For this example, a proportion of 0.1000 is not listed in column C but you can use 0.1003, which is listed. Once you have found the correct proportion in the table, simply read across the row to find the corresponding z-score value in column A.

For this example, the *z*-score that separates the extreme 10% in the tail is $z = 1.28$. At this point you must be careful because the table does not differentiate between the right-hand tail and the left-hand tail of the distribution. Specifically, the final answer could be either $z = +1.28$, which separates 10% in the right-hand tail, or $z = –1.28$, which separates 10% in the left-hand tail. For this problem we want the right-hand tail (the highest 10%), so the *z*-score value is $z = +1.28$.

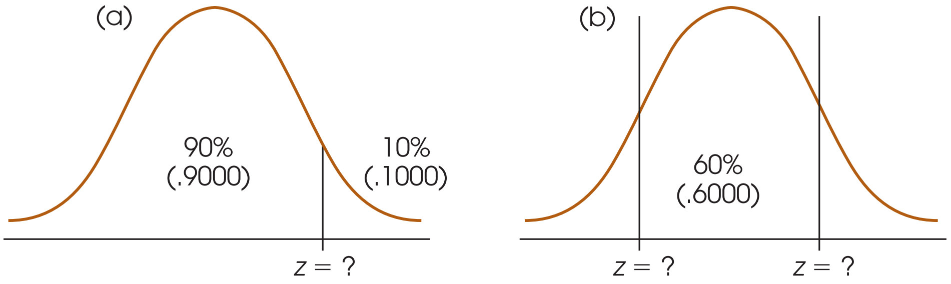
EXAMPLE 6.4B

For a normal distribution, what *z*-score values form the boundaries that separate the middle 60% of the distribution from the rest of the scores?

Again, we have sketched a normal distribution [Figure 6.9(b)] and drawn vertical lines so that roughly 60% of the distribution in the central section, with the remainder

FIGURE 6.9

The distribution for Examples 6.4A and 6.4B.



split equally between the two tails. The problem is to find the z-score values that define the exact locations for the lines. To find the z-score values, we begin with the known proportions: 0.6000 in the center and 0.4000 divided equally between the two tails. Although these proportions can be used in several different ways, this example provides an opportunity to demonstrate how column D in the table can be used to solve problems. For this problem, the 0.6000 in the center can be divided in half with exactly 0.3000 to the right of the mean and exactly 0.3000 to the left. Each of these sections corresponds to the proportion listed in column D. Begin by scanning down column D, looking for a value of 0.3000. Again, this exact proportion is not in the table, but the closest value is 0.2995. Reading across the row to column A, you should find a z-score value of $z = 0.84$. Looking again at the sketch [Figure 6.9(b)], the right-hand line is located at $z = +0.84$ and the left-hand line is located at $z = -0.84$.

You may have noticed that we have sketched distributions for each of the preceding problems. As a general rule, you should always sketch a distribution, locate the mean with a vertical line, and shade in the portion that you are trying to determine. Look at your sketch. It will help you to determine which columns to use in the unit normal table. If you make a habit of drawing sketches, you will avoid careless errors when using the table.

LEARNING CHECK

1. Find the proportion of a normal distribution that corresponds to each of the following sections:

a. $z < 0.25$
b. $z > 0.80$
c. $z < -1.50$
d. $z > -0.75$

2. For a normal distribution, find the z-score location that divides the distribution as follows:

a. Separate the top 20% from the rest.
b. Separate the top 60% from the rest.
c. Separate the middle 70% from the rest.

3. The tail will be on the right-hand side of a normal distribution for any positive z-score. (True or false?)

ANSWERS

1. a. $p = 0.5987$
b. $p = 0.2119$
c. $p = 0.0668$
d. $p = 0.7734$

2. a. $z = 0.84$
b. $z = -0.25$
c. $z = -1.04$ and $+ 1.04$

3. True

6.3 PROBABILITIES AND PROPORTIONS FOR SCORES FROM A NORMAL DISTRIBUTION

In the preceding section, we used the unit normal table to find probabilities and proportions corresponding to specific *z*-score values. In most situations, however, it is necessary to find probabilities for specific *X* values. Consider the following example:

> It is known that IQ scores form a normal distribution with $\mu = 100$ and $\sigma = 15$. Given this information, what is the probability of randomly selecting an individual with an IQ score less than 120?

This problem is asking for a specific probability or proportion of a normal distribution. However, before we can look up the answer in the unit normal table, we must first transform the IQ scores (*X* values) into *z*-scores. Thus, to solve this new kind of probability problem, we must add one new step to the process. Specifically, to answer probability questions about scores (*X* values) from a normal distribution, you must use the following two-step procedure:

Caution: The unit normal table can be used only with normal-shaped distributions. If a distribution is not normal, transforming to *z*-scores does not make it normal.

1. Transform the *X* values into *z*-scores.
2. Use the unit normal table to look up the proportions corresponding to the *z*-score values.

This process is demonstrated in the following examples. Once again, we suggest that you sketch the distribution and shade the portion you are trying to find to avoid careless mistakes.

EXAMPLE 6.5

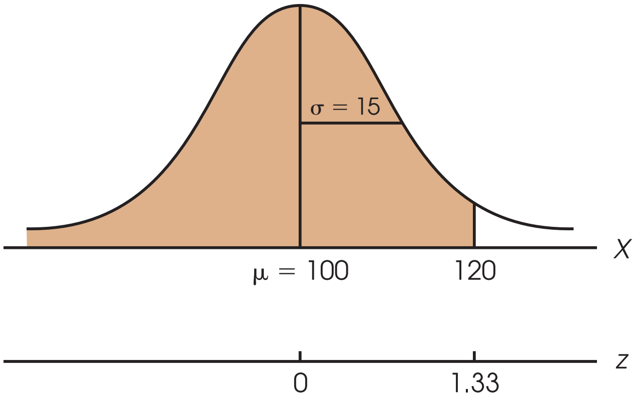
We now answer the probability question about IQ scores that we presented earlier. Specifically, what is the probability of randomly selecting an individual with an IQ score less than 120? Restated in terms of proportions, we want to find the proportion of the IQ distribution that corresponds to scores less than 120. The distribution is drawn in Figure 6.10, and the portion we want has been shaded.

The first step is to change the *X* values into *z*-scores. In particular, the score of $X = 120$ is changed to

$$z = \frac{X - \mu}{\sigma} = \frac{120 - 100}{15} = \frac{20}{15} = 1.33$$

FIGURE 6.10

The distribution of IQ scores. The problem is to find the probability or proportion of the distribution corresponding to scores less than 120.



Thus, an IQ score of $X = 120$ corresponds to a z-score of $z = 1.33$, and IQ scores less than 120 correspond to z-scores less than 1.33.

Next, look up the z-score value in the unit normal table. Because we want the proportion of the distribution in the body to the left of $X = 120$ (see Figure 6.10), the answer is in column B. Consulting the table, we see that a z-score of 1.33 corresponds to a proportion of 0.9082. The probability of randomly selecting an individual with an IQ less than 120 is $p = 0.9082$. In symbols,

$$p(X < 120) = p(z < 1.33) = 0.9082 \text{ (or } 90.82\%\text{)}$$

Finally, notice that we phrased this question in terms of a *probability*. Specifically, we asked, "What is the probability of selecting an individual with an IQ less than 120?" However, the same question can be phrased in terms of a *proportion:* "What proportion of all of the individuals in the population have IQ scores less than 120?" Both versions ask exactly the same question and produce exactly the same answer. A third alternative for presenting the same question is introduced in Box 6.1.

Finding proportions/probabilities located between two scores The next example demonstrates the process of finding the probability of selecting a score that is located *between* two specific values. Although these problems can be solved using the proportions of columns B and C (body and tail), they are often easier to solve with the proportions listed in column D.

EXAMPLE 6.6 The highway department conducted a study measuring driving speeds on a local section of interstate highway. They found an average speed of $\mu = 58$ miles per hour with a standard deviation of $\sigma = 10$. The distribution was approximately normal.

BOX 6.1 PROBABILITIES, PROPORTIONS, AND PERCENTILE RANKS

Thus far we have discussed parts of distributions in terms of proportions and probabilities. However, there is another set of terminology that deals with many of the same concepts. Specifically, in Chapter 2 we defined the *percentile rank* for a specific score as the percentage of the individuals in the distribution who have scores that are less than or equal to the specific score. For example, if 70% of the individuals have scores of $X = 45$ or lower, then $X = 45$ has a percentile rank of 70%. When a score is referred to by its percentile rank, the score is called a *percentile*. For example, a score with a percentile rank of 70% is called the 70th percentile.

Using this terminology, it is possible to rephrase some of the probability problems that we have been working. In Example 6.5, the problem is presented as "What is the probability of randomly selecting an individual with an IQ of less than 120?" Exactly the same question could be phrased as "What is the percentile rank for an IQ score of 120?" In each case, we are drawing a line at $X = 120$ and looking for the proportion of the distribution on the left-hand side of the line. Similarly, Example 6.8 asks "How much time do you have to spend commuting each day to be in the highest 10% nationwide?" Because this score separates the top 10% from the bottom 90%, the same question could be rephrased as "What is the 90th percentile for the distribution of commuting times?"

Given this information, what proportion of the cars are traveling between 55 and 65 miles per hour? Using probability notation, we can express the problem as

$$p(55 < X < 65) = ?$$

The distribution of driving speeds is shown in Figure 6.11 with the appropriate area shaded. The first step is to determine the *z*-score corresponding to the *X* value at each end of the interval.

$$\text{For } X = 55\text{: } z = \frac{X - \mu}{\sigma} = \frac{55 - 58}{10} = \frac{-3}{10} = -0.30$$

$$\text{For } X = 65\text{: } z = \frac{X - \mu}{\sigma} = \frac{65 - 58}{10} = \frac{7}{10} = 0.70$$

Looking again at Figure 6.11, we see that the proportion we are seeking can be divided into two sections: (1) the area left of the mean, and (2) the area right of the mean. The first area is the proportion between the mean and $z = –0.30$, and the second is the proportion between the mean and $z = +0.70$. Using column D of the unit normal table, these two proportions are 0.1179 and 0.2580. The total proportion is obtained by adding these two sections:

$$p(55 < X < 65) = p(–0.30 < z < +0.70) = 0.1179 + 0.2580 = 0.3759$$

EXAMPLE 6.7 Using the same distribution of driving speeds from the previous example, what proportion of cars are traveling between 65 and 75 miles per hour?

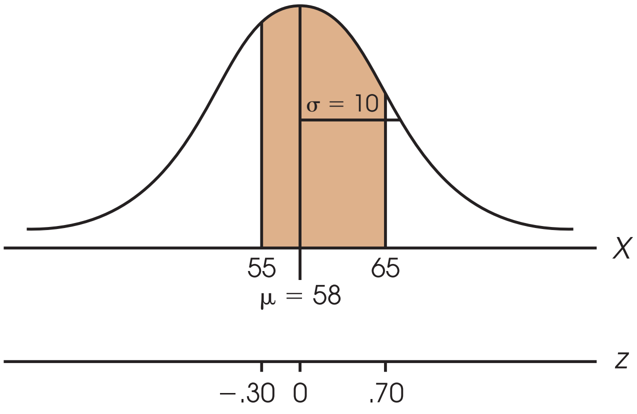
$$p(65 < X < 75) = ?$$

The distribution is shown in Figure 6.12 with the appropriate area shaded. Again, we start by determining the *z*-score corresponding to each end of the interval.

$$\text{For } X = 75\text{:} \quad z = \frac{X - \mu}{\sigma} = \frac{75 - 58}{10} = \frac{17}{10} = 1.70$$

$$\text{For } X = 65\text{:} \quad z = \frac{X - \mu}{\sigma} = \frac{65 - 58}{10} = \frac{7}{10} = 0.70$$

FIGURE 6.11
The distribution for Example 6.6.



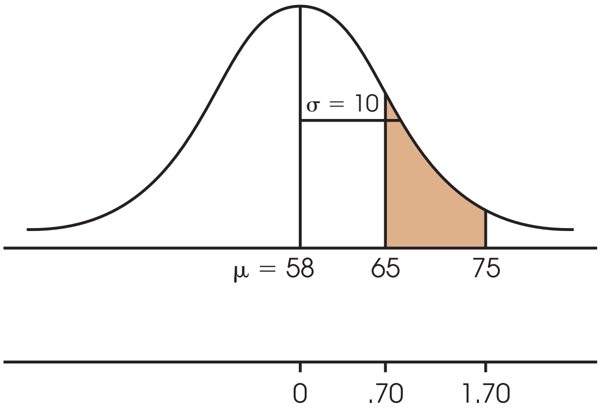


FIGURE 6.12
The distribution for Example 6.7.

There are several different ways to use the unit normal table to find the proportion between these two *z*-scores. For this example, we use the proportions in the tail of the distribution (column C). According to column C in the unit normal table, the proportion in the tail beyond $z = 0.70$ is $p = 0.2420$. Note that this proportion includes the section that we want, but it also includes an extra, unwanted section located in the tail beyond $z = 1.70$. Locating $z = 1.70$ in the table, and reading across the row to column C, we see that the unwanted section is $p = 0.0446$. To obtain the correct answer, we subtract the unwanted portion from the total proportion in the tail beyond $z = 0.70$.

$$p(65 < X < 75) = p(0.70 < z < 1.70) = 0.2420 - 0.0446 = 0.1974$$

Finding scores corresponding to specific proportions or probabilities In the previous three examples, the problem was to find the proportion or probability corresponding to specific *X* values. The two-step process for finding these proportions is shown in Figure 6.13. Thus far, we have only considered examples that move in a clockwise direction around the triangle shown in the figure; that is, we start with an *X* value that is transformed into a *z*-score, and then we use the unit normal table to look up the

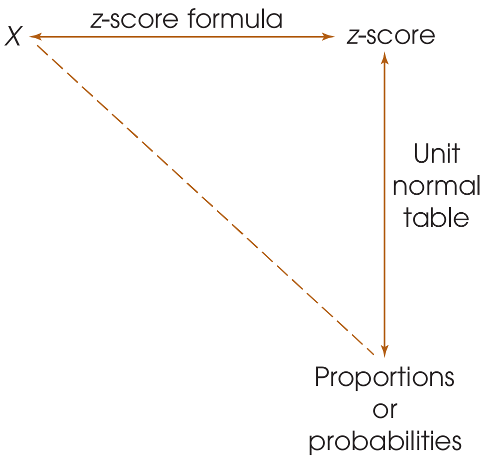


FIGURE 6.13
Determining probabilities or proportions for a normal distribution is shown as a two-step process with *z*-scores as an intermediate stop along the way. Note that you cannot move directly along the dashed line between *X* values and probabilities and proportions. Instead, you must follow the solid lines around the corner.

appropriate proportion. You should realize, however, that it is possible to reverse this two-step process so that we move backward, or counterclockwise, around the triangle. This reverse process allows us to find the score (*X* value) corresponding to a specific proportion in the distribution. Following the lines in Figure 6.13, we begin with a specific proportion, use the unit normal table to look up the corresponding *z*-score, and then transform the *z*-score into an *X* value. The following example demonstrates this process.

EXAMPLE 6.8

The U.S. Census Bureau (2005) reports that Americans spend an average of $\mu = 24.3$ minutes commuting to work each day. Assuming that the distribution of commuting times is normal with a standard deviation of $\sigma = 10$ minutes, how much time do you have to spend commuting each day to be in the highest 10% nationwide? (An alternative form of the same question is presented in Box 6.1.) The distribution is shown in Figure 6.14 with a portion representing approximately 10% shaded in the right-hand tail.

In this problem, we begin with a proportion (10% or 0.10), and we are looking for a score. According to the map in Figure 6.13, we can move from *p* (proportion) to *X* (score) via *z*-scores. The first step is to use the unit normal table to find the *z*-score that corresponds to a proportion of 0.10 in the tail. First, scan the values in column C to locate the row that has a proportion of 0.10 in the tail of the distribution. Note that you will not find 0.1000 exactly, but locate the closest value possible. In this case, the closest value is 0.1003. Reading across the row, we find $z = 1.28$ in column A.

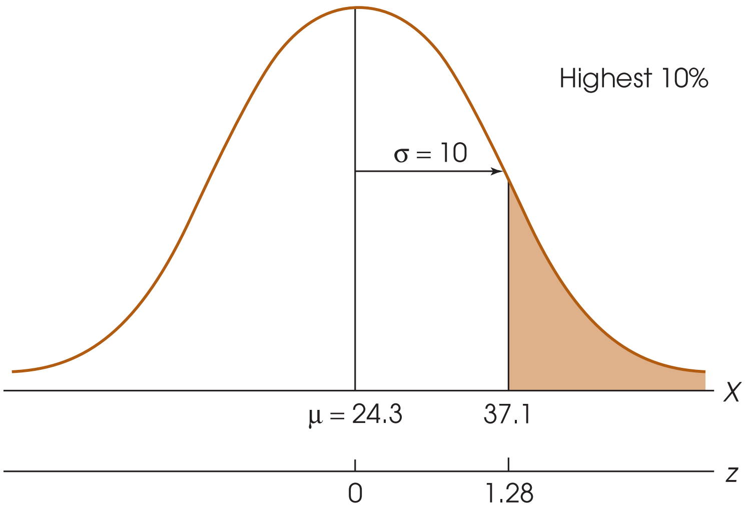
The next step is to determine whether the *z*-score is positive or negative. Remember that the table does not specify the sign of the *z*-score. Looking at the distribution in Figure 6.14, you should realize that the score we want is above the mean, so the *z*-score is positive, $z = +1.28$.

The final step is to transform the *z*-score into an *X* value. By definition, a *z*-score of $+1.28$ corresponds to a score that is located above the mean by 1.28 standard deviations. One standard deviation is equal to 10 points ($\sigma = 10$), so 1.28 standard deviations is

$$1.28\sigma = 1.28(10) = 12.8 \text{ points}$$

FIGURE 6.14

The distribution of commuting times for American workers. The problem is to find the score that separates the highest 10% of commuting times from the rest.



Thus, our score is located above the mean ($\mu = 24.3$) by a distance of 12.8 points. Therefore,

$$X = 24.3 + 12.8 = 37.1$$

The answer for our original question is that you must commute at least 37.1 minutes a day to be in the top 10% of American commuters.

EXAMPLE 6.9 Again, the distribution of commuting time for American workers is normal with a mean of $\mu = 24.3$ minutes and a standard deviation of $\sigma = 10$ minutes. For this example, we find the range of values that defines the middle 90% of the distribution. The entire distribution is shown in Figure 6.15 with the middle portion shaded.

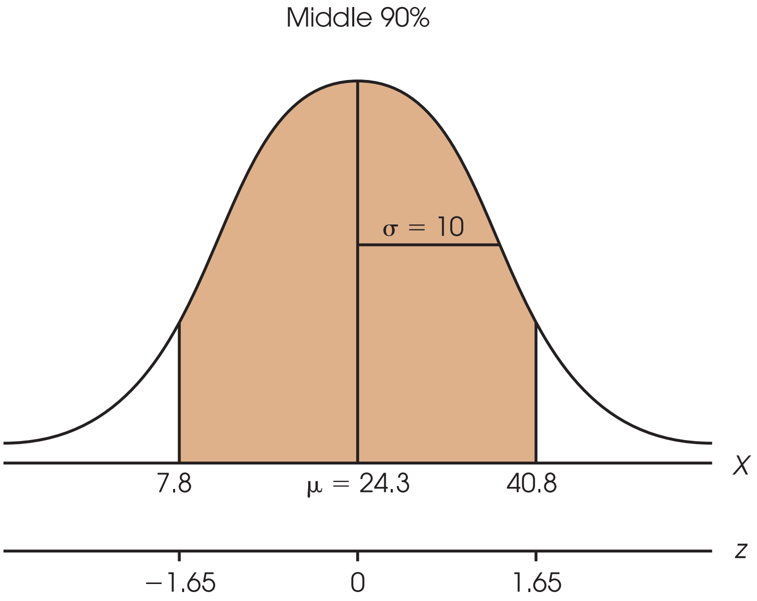
The 90% (0.9000) in the middle of the distribution can be split in half with 45% (0.4500) on each side of the mean. Looking up 0.4500, in column D of the unit normal table, you will find that the exact proportion is not listed. However, you will find 0.4495 and 0.4505, which are equally close. Technically, either value is acceptable, but we use 0.4505 so that the total area in the middle is at least 90%. Reading across the row, you should find a *z*-score of $z = 1.65$ in column A. Thus, the *z*-score at the right boundary is $z = +1.65$ and the *z*-score at the left boundary is $z = –1.65$. In either case, a *z*-score of 1.65 indicates a location that is 1.65 standard deviations away from the mean. For the distribution of commuting times, one standard deviation is $\sigma = 10$, so 1.65 standard deviations is a distance of

$$1.65\sigma = 1.65(10) = 16.5 \text{ points}$$

Therefore, the score at the right-hand boundary is located above the mean by 16.5 points and corresponds to $X = 24.3 + 16.5 = 40.8$. Similarly, the score at the left-hand boundary is below the mean by 16.5 points and corresponds to $X = 24.3 – 16.5 = 7.8$. The middle 90% of the distribution corresponds to values between 7.8 and 40.8. Thus, 90% of American commuters spend between 7.8 and 40.8 minutes commuting to work each day. Only 10% of commuters spend either more time or less time.

FIGURE 6.15

The distribution of commuting times for American workers. The problem is to find the middle 90% of the distribution.



LEARNING CHECK

1. For a normal distribution with a mean of $\mu = 60$ and a standard deviation of $\sigma = 12$, find each probability value requested.
   a. $p(X > 66)$
   b. $p(X < 75)$
   c. $p(X < 57)$
   d. $p(48 < X < 72)$
2. Scores on the Mathematics section of the SAT Reasoning Test form a normal distribution with a mean of $\mu = 500$ and a standard deviation of $\sigma = 100$.
   a. If the state college only accepts students who score in the top 60% on this test, what is the minimum score needed for admission?
   b. What is the minimum score necessary to be in the top 10% of the distribution?
   c. What scores form the boundaries for the middle 50% of the distribution?
3. What is the probability of selecting a score greater than 45 from a positively skewed distribution with $\mu = 40$ and $\sigma = 10$? (Be careful.)

ANSWERS

1. **a.** $p = 0.3085$
   b. $p = 0.8944$
   c. $p = 0.4013$
   d. $p = 0.6826$
2. **a.** $z = –0.25$; $X = 475$
   b. $z = 1.28$; $X = 628$
   c. $z = \pm 0.67$; $X = 433$ and $X = 567$
3. You cannot obtain the answer. The unit normal table cannot be used to answer this question because the distribution is not normal.

6.4 PROBABILITY AND THE BINOMIAL DISTRIBUTION

When a variable is measured on a scale consisting of exactly two categories, the resulting data are called binomial. The term *binomial* can be loosely translated as "two names," referring to the two categories on the measurement scale.

Binomial data can occur when a variable naturally exists with only two categories. For example, people can be classified as male or female, and a coin toss results in either heads or tails. It also is common for a researcher to simplify data by collapsing the scores into two categories. For example, a psychologist may use personality scores to classify people as either high or low in aggression.

In binomial situations, the researcher often knows the probabilities associated with each of the two categories. With a balanced coin, for example, $p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$. The question of interest is the number of times each category occurs in a series of trials or in a sample of individuals. For example:

> What is the probability of obtaining 15 heads in 20 tosses of a balanced coin?
> What is the probability of obtaining more than 40 introverts in a sampling of 50 college freshmen?

As we shall see, the normal distribution serves as an excellent model for computing probabilities with binomial data.

THE BINOMIAL DISTRIBUTION

To answer probability questions about binomial data, we must examine the binomial distribution. To define and describe this distribution, we first introduce some notation.

1. The two categories are identified as A and B.
2. The probabilities (or proportions) associated with each category are identified as

   $p = p(A) =$ the probability of A

   $q = p(B) =$ the probability of B

   Notice that $p + q = 1.00$ because A and B are the only two possible outcomes.
3. The number of individuals or observations in the sample is identified by n.
4. The variable X refers to the number of times category A occurs in the sample.

Notice that X can have any value from 0 (none of the sample is in category A) to n (all of the sample is in category A).

DEFINITION

Using the notation presented here, the **binomial distribution** shows the probability associated with each value of X from $X = 0$ to $X = n$.

A simple example of a binomial distribution is presented next.

EXAMPLE 6.10

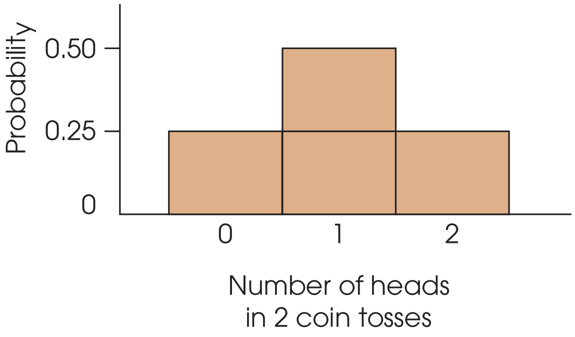
Figure 6.16 shows the binomial distribution for the number of heads obtained in 2 tosses of a balanced coin. This distribution shows that it is possible to obtain as many as 2 heads or as few as 0 heads in 2 tosses. The most likely outcome (highest probability) is to obtain exactly 1 head in 2 tosses. The construction of this binomial distribution is discussed in detail next.

For this example, the event we are considering is a coin toss. There are two possible outcomes, heads and tails. We assume the coin is balanced, so

$$p = p(\text{heads}) = \frac{1}{2}$$

FIGURE 6.16

The binomial distribution showing the probability for the number of heads in 2 tosses of a balanced coin.



$$q = p(\text{tails}) = \frac{1}{2}$$

We are looking at a sample of $n = 2$ tosses, and the variable of interest is

X = the number of heads

To construct the binomial distribution, we look at all of the possible outcomes from tossing a coin 2 times. The complete set of 4 outcomes is listed in the following table.

1st Toss	2nd Toss	
Heads	Heads	(Both heads)
Heads	Tails	(Each sequence has exactly 1 head)
Tails	Heads	
Tails	Tails	(No heads)

Notice that there are 4 possible outcomes when you toss a coin 2 times. Only 1 of the 4 outcomes has 2 heads, so the probability of obtaining 2 heads is $p = \frac{1}{4}$. Similarly, 2 of the 4 outcomes have exactly 1 head, so the probability of 1 head is $p = \frac{2}{4} = \frac{1}{2}$. Finally, the probability of no heads ($X = 0$) is $p = \frac{1}{4}$. These are the probabilities shown in Figure 6.16.

Note that this binomial distribution can be used to answer probability questions. For example, what is the probability of obtaining at least 1 head in 2 tosses? According to the distribution shown in Figure 6.16, the answer is $\frac{3}{4}$.

Similar binomial distributions have been constructed for the number of heads in 4 tosses of a balanced coin and in 6 tosses of a coin (Figure 6.17). It should be obvious from the binomial distributions shown in Figures 6.16 and 6.17 that the binomial

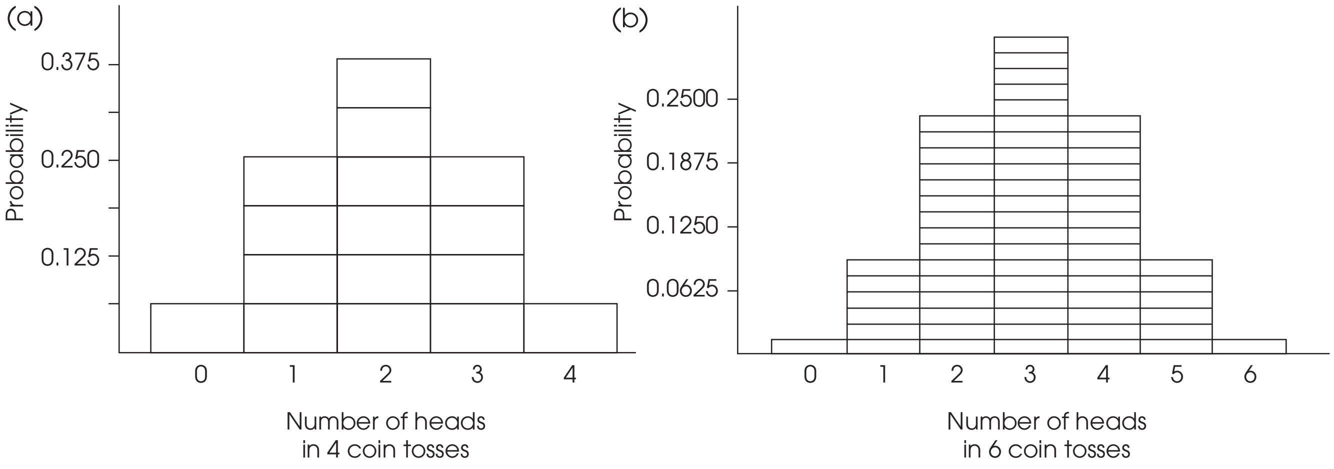


FIGURE 6.17

Binomial distributions showing probabilities for the number of heads (a) in 4 tosses of a balanced coin and (b) in 6 tosses of a balanced coin.

distribution tends toward a normal shape, especially when the sample size (n) is relatively large.

It should not be surprising that the binomial distribution tends to be normal. With $n = 10$ coin tosses, for example, the most likely outcome would be to obtain around $X = 5$ heads. On the other hand, values far from 5 would be very unlikely—you would not expect to get all 10 heads or all 10 tails (0 heads) in 10 tosses. Notice that we have described a normal-shaped distribution: The probabilities are highest in the middle (around $X = 5$), and they taper off as you move toward either extreme.

THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

We have stated that the binomial distribution tends to approximate a normal distribution, particularly when n is large. To be more specific, the binomial distribution is a nearly perfect normal distribution when pn and qn are both equal to or greater than 10. Under these circumstances, the binomial distribution approximates a normal distribution with the following parameters:

The value of 10 for *pn* or *qn* is a general guide, not an absolute cutoff. Values slightly less than 10 still provide a good approximation. However, with smaller values the normal approximation becomes less accurate as a substitute for the binomial distribution.

$$\text{Mean: } \mu = pn \qquad (6.1)$$

$$\text{standard deviation: } \sigma = \sqrt{npq} \qquad (6.2)$$

Within this normal distribution, each value of X has a corresponding z-score,

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}} \qquad (6.3)$$

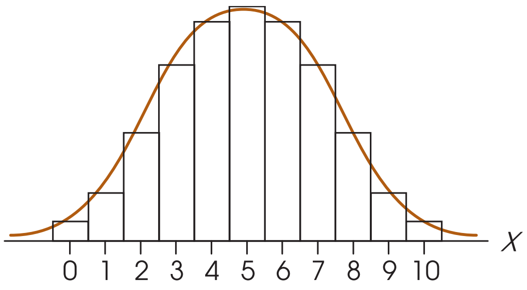
The fact that the binomial distribution tends to be normal in shape means that we can compute probability values directly from z-scores and the unit normal table.

Coin tosses produce discrete events. In a series of coin tosses, you may observe 1 head, 2 heads, 3 heads, and so on, but no values between them are possible (p. 21).

It is important to remember that the normal distribution is only an approximation of a true binomial distribution. Binomial values, such as the number of heads in a series of coin tosses, are *discrete*. The normal distribution is *continuous*. However, the *normal approximation* provides an extremely accurate model for computing binomial probabilities in many situations. Figure 6.18 shows the difference between the true binomial distribution, the discrete histogram, and the normal curve that approximates the binomial distribution. Although the two distributions are slightly different, the area under the distributions is nearly equivalent. *Remember, it is the area under the distribution that is used to find probabilities.*

FIGURE 6.18

The relationship between the binomial distribution and the normal distribution. The binomial distribution is always a discrete histogram, and the normal distribution is a continuous, smooth curve. Each X value is represented by a bar in the histogram or a section of the normal distribution.



To gain maximum accuracy when using the normal approximation, you must remember that each X value in the binomial distribution actually corresponds to a bar in the histogram. In the histogram in Figure 6.18, for example, the score $X = 6$ is represented by a bar that is bounded by real limits of 5.5 and 6.5. The actual probability of $X = 6$ is determined by the area contained in this bar. To approximate this probability using the normal distribution, you should find the area that is contained between the two real limits. Similarly, if you are using the normal approximation to find the probability of obtaining a score greater than $X = 6$, you should use the area beyond the real limit boundary of 6.5. The following example demonstrates how the normal approximation to the binomial distribution is used to compute probability values.

EXAMPLE 6.11

Suppose that you plan to test for ESP (extra-sensory perception) by asking people to predict the suit of a card that is randomly selected from a complete deck. Before you begin your test, however, you need to know what kind of performance is expected from people who do not have ESP and are simply guessing. For these people, there are two possible outcomes, correct or incorrect, on each trial. Because there are four different suits, the probability of a correct prediction (assuming that there is no ESP) is $p = \frac{1}{4}$ and the probability of an incorrect prediction is $q = \frac{3}{4}$. With a series of $n = 48$ trials, this situation meets the criteria for the normal approximation to the binomial distribution:

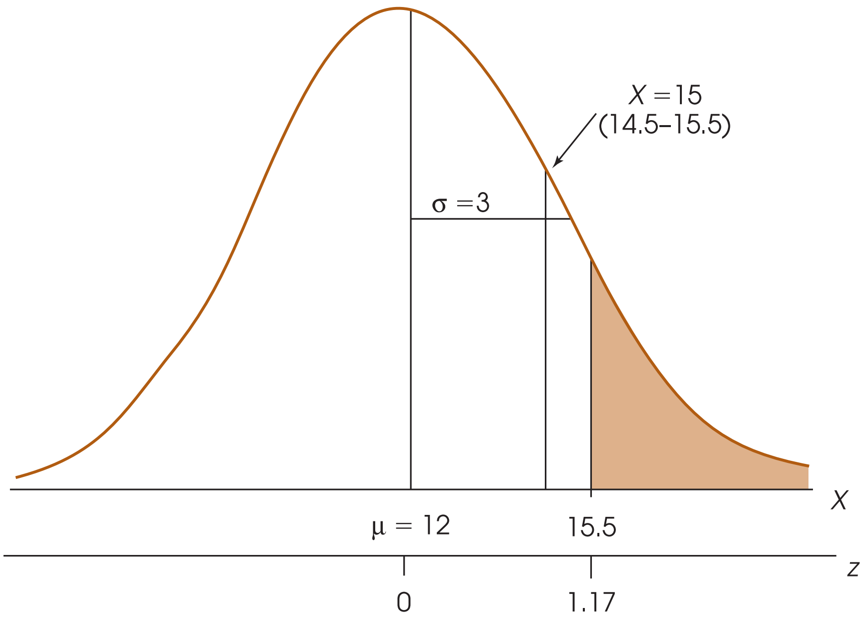
$$pn = \tfrac{1}{4}(48) = 12 \quad qn = \tfrac{3}{4}(48) = 36 \text{ Both are greater than 10.}$$

Thus, the distribution of correct predictions forms a normal-shaped distribution with a mean of $\mu = pn = 12$ and a standard deviation of $\sigma = \sqrt{npq} = \sqrt{9} = 3$. We can use this distribution to determine probabilities for different levels of performance. For example, we can calculate the probability that a person without ESP would guess correctly more than 15 times in a series of 48 trials.

Figure 6.19 shows the binomial distribution that we are considering. Because we want the probability of obtaining *more than* 15 correct predictions, we must find the

FIGURE 6.19

The normal approximation of the binomial distribution discussed in Example 6.11.



shaded area in the tail of the distribution beyond $X = 15.5$. (Remember that a score of 15 corresponds to an interval from 14.5 to 15.5. We want scores beyond this interval.) The first step is to find the z-score corresponding to $X = 15.5$.

Caution: If the question had asked for the probability of *15 or more* correct predictions, we would find the area beyond $X = 14.5$. Read the question carefully.

$$z = \frac{X - \mu}{\sigma} = \frac{15.5 - 12}{3} = 1.17$$

Next, look up the probability in the unit normal table. In this case, we want the proportion in the tail beyond $z = 1.17$. The value from the table is $p = 0.1210$. This is the answer we want. The probability of correctly predicting the suit of a card more than 15 times in a series of 48 trials is only $p = 0.1210$ or 12.10%. Thus, it is very unlikely for an individual without ESP to guess correctly more than 15 out of 48 trials.

LEARNING CHECK

1. Under what circumstances is the normal distribution an accurate approximation of the binomial distribution?
2. In the game Rock-Paper-Scissors, the probability that both players will select the same response and tie is $p = \frac{1}{3}$, and the probability that they will pick different responses is $p = \frac{2}{3}$. If two people play 72 rounds of the game and choose their responses randomly, what is the probability that they will choose the same response (tie) more than 28 times?
3. If you toss a balanced coin 36 times, you would expect, on the average, to get 18 heads and 18 tails. What is the probability of obtaining exactly 18 heads in 36 tosses?

ANSWERS

1. When pn and qn are both greater than 10
2. With $p = \frac{1}{3}$ and $q = \frac{2}{3}$, the binomial distribution is normal with $\mu = 24$ and $\sigma = 4$; $p(X > 28.5) = p(z > 1.13) = 0.1292$.
3. $X = 18$ is an interval with real limits of 17.5 and 18.5. The real limits correspond to $z = \pm 0.17$, and a probability of $p = 0.1350$.

6.5 LOOKING AHEAD TO INFERENTIAL STATISTICS

Probability forms a direct link between samples and the populations from which they come. As we noted at the beginning of this chapter, this link is the foundation for the inferential statistics in future chapters. The following example provides a brief preview of how probability is used in the context of inferential statistics.

We ended Chapter 5 with a demonstration of how inferential statistics are used to help interpret the results of a research study. A general research situation was shown in Figure 5.9 and is repeated here in Figure 6.20. The research begins with a population that forms a normal distribution with a mean of $\mu = 400$ and a standard deviation of $\sigma = 20$. A sample is selected from the population and a treatment is administered to the sample. The goal for the study is to evaluate the effect of the treatment.

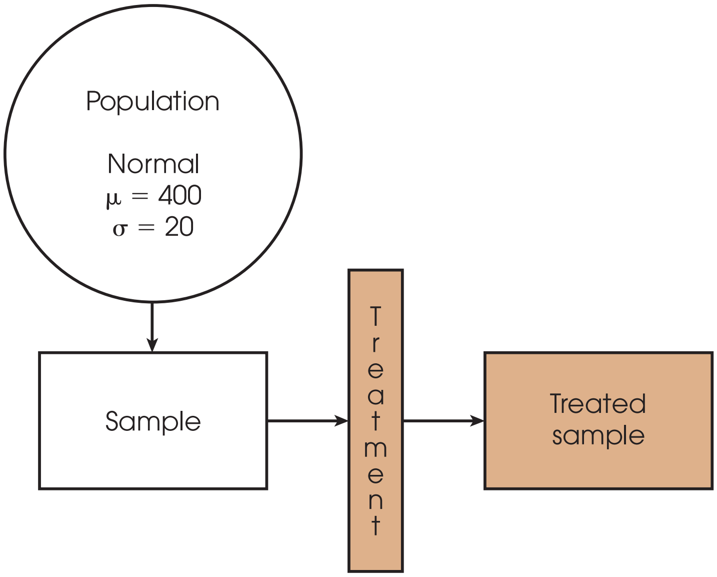


FIGURE 6.20

A diagram of a research study. A sample is selected from the population and receives a treatment. The goal is to determine whether the treatment has an effect.

To determine whether the treatment has an effect, the researcher simply compares the treated sample with the original population. If the individuals in the sample have scores around 400 (the original population mean), then we must conclude that the treatment appears to have no effect. On the other hand, if the treated individuals have scores that are noticeably different from 400, then the researcher has evidence that the treatment does have an effect. Notice that the study is using a sample to help answer a question about a population; this is the essence of inferential statistics.

The problem for the researcher is determining exactly what is meant by "noticeably different" from 400. If a treated individual has a score of $X = 415$, is that enough to say that the treatment has an effect? What about $X = 420$ or $X = 450$? In Chapter 5, we suggested that *z*-scores provide one method for solving this problem. Specifically, we suggested that a *z*-score value beyond $z = 2.00$ (or –2.00) was an extreme value and, therefore, noticeably different. However, the choice of $z = \pm 2.00$ was purely arbitrary. Now we have another tool, *probability,* to help us decide exactly where to set the boundaries.

Figure 6.21 shows the original population from our hypothetical research study. Note that most of the scores are located close to $\mu = 400$. Also note that we have added boundaries separating the middle 95% of the distribution from the extreme 5%, or 0.0500, in the two tails. Dividing the 0.0500 in half produces proportions of 0.0250 in the right-hand tail and 0.0250 in the left-hand tail. Using column C of the unit normal table, the *z*-score boundaries for the right and left tails are $z = +1.96$ and $z = –1.96$, respectively.

The boundaries set at $z = \pm 1.96$ provide objective criteria for deciding whether our sample provides evidence that the treatment has an effect. Specifically, we use the sample data to help decide between the following two alternatives:

1. The treatment has no effect. After treatment, the scores still average $\mu = 400$.
2. The treatment does have an effect. The treatment changes the scores so that, after treatment, they no longer average $\mu = 400$.

As a starting point, we assume that the first alternative is true and the treatment has no effect. In this case, treated individuals should be no different from the individuals in

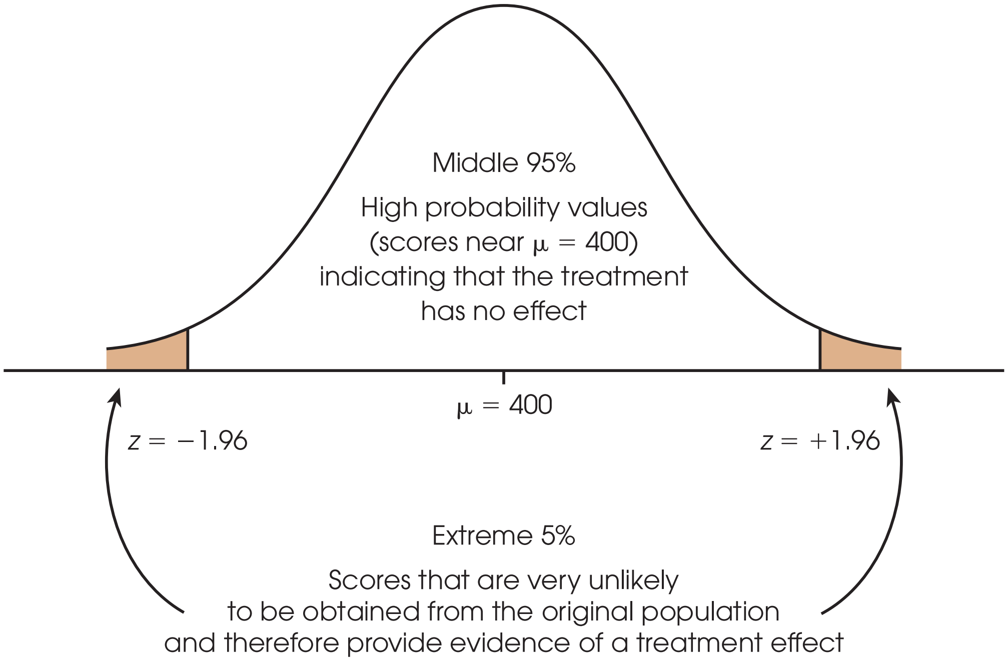

Middle 95%
High probability values
(scores near μ = 400)
indicating that the treatment
has no effect
μ = 400
z = −1.96
z = +1.96
Extreme 5%
Scores that are very unlikely
to be obtained from the original population
and therefore provide evidence of a treatment effect

FIGURE 6.21

Using probability to evaluate a treatment effect. Values that are extremely unlikely to be obtained from the original population are viewed as evidence of a treatment effect.

the original population, which is shown in Figure 6.21. Notice that, if our assumption is correct, it is extremely unlikely (probability less than 5%) for a treated individual to be outside the ±1.96 boundaries. Therefore, if we obtain a treated individual who is outside the boundaries, we must conclude that the assumption is probably not correct. In this case, we are left with the second alternative (the treatment does have an effect) as the more likely explanation.

Notice that we are comparing the treated sample with the original population to see if the sample is noticeably different. If it is different, we can conclude that the treatment seems to have an effect. Now we are defining “noticeably different” as meaning “very unlikely.” Specifically, if the sample is very unlikely to have come from a population of untreated individuals, then we must conclude that the treatment has an effect and has caused the sample to be different.

We are using the sample data and the ±1.96 boundaries, which were determined by probabilities, to make a general decision about the treatment. If the sample falls outside the boundaries we make the following logical conclusion:

a. This kind of sample is very unlikely to occur if the treatment has no effect.

b. Therefore, the treatment must have an effect that changed the sample.

On the other hand, if the sample falls between the ±1.96 boundaries, we conclude:

a. This is the kind of sample that is likely to occur if the treatment has no effect.

b. Therefore, the treatment does not appear to have had an effect.

SUMMARY

1. The probability of a particular event A is defined as a fraction or proportion:

$$p(A) = \frac{\text{number of outcomes classified as } A}{\text{total number of possible outcomes}}$$

2. Our definition of probability is accurate only for random samples. There are two requirements that must be satisfied for a random sample:
   - **a.** Every individual in the population has an equal chance of being selected.
   - **b.** When more than one individual is being selected, the probabilities must stay constant. This means that there must be sampling with replacement.

3. All probability problems can be restated as proportion problems. The "probability of selecting a king from a deck of cards" is equivalent to the "proportion of the deck that consists of kings." For frequency distributions, probability questions can be answered by determining proportions of area. The "probability of selecting an individual with an IQ greater than 108" is equivalent to the "proportion of the whole population that consists of IQs greater than 108."

4. For normal distributions, probabilities (proportions) can be found in the unit normal table. The table provides a listing of the proportions of a normal distribution that correspond to each z-score value. With the table, it is possible to move between X values and probabilities using a two-step procedure:
   - **a.** The z-score formula (Chapter 5) allows you to transform X to z or to change z back to X.
   - **b.** The unit normal table allows you to look up the probability (proportion) corresponding to each z-score or the z-score corresponding to each probability.

5. Percentiles and percentile ranks measure the relative standing of a score within a distribution (see Box 6.1). Percentile rank is the percentage of individuals with scores at or below a particular X value. A percentile is an X value that is identified by its rank. The percentile rank always corresponds to the proportion to the left of the score in question.

6. The binomial distribution is used whenever the measurement procedure classifies individuals into exactly two categories. The two categories are identified as A and B, with probabilities of

$$p(A) = p \qquad \text{and} \qquad p(B) = q$$

7. The binomial distribution gives the probability for each value of X, where X equals the number of occurrences of category A in a series of n events. For example, X equals the number of heads in $n = 10$ tosses of a coin.

   When pn and qn are both at least 10, the binomial distribution is closely approximated by a normal distribution with

$$\mu = pn \quad \sigma = \sqrt{npq}$$

8. In the normal approximation to the binomial distribution, each value of X has a corresponding z-score:

$$z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}}$$

   With the z-score and the unit normal table, you can find probability values associated with any value of X. For maximum accuracy, you should use the appropriate real limits for X when computing z-scores and probabilities.

KEY TERMS

probability (165)
random sample (167)
independent random sample (167)
sampling with replacement (168)
unit normal table (172)
percentile rank (179)
percentile (179)
binomial distribution (185)
normal approximation (binomial) (187)