CHAPTER

8

Introduction to Hypothesis Testing

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.

- z-Scores (Chapter 5)
- Distribution of sample means (Chapter 7)
  - Expected value
  - Standard error
  - Probability and sample means

Preview

Most of us spend more time looking down at our mobile devices than we do looking up at the clouds. But if you do watch the clouds and have a little imagination, you occasionally see them form into familiar shapes. Figure 8.1 is a photograph of a cloud formation seen over Kansas City around Christmas in 2008. Do you recognize a familiar image?

The cloud pattern shown in Figure 8.1 was formed simply by chance. Specifically, it was the random forces of wind and air currents that produced a portrait of Santa Claus. The clouds did not conspire to form the image, and it was not deliberately created by a team of professional skywriters. The point we would like to make is that what appear to be meaningful patterns can be produced by random chance.

The Problem Researchers often find meaningful patterns in the sample data obtained in research studies. The problem is deciding whether the patterns found in a sample reflect real patterns that exist in the population or are simply random, chance occurrences.

The Solution To differentiate between real, systematic patterns and random, chance occurrences, researchers rely on a statistical technique known as *hypothesis testing,* which is introduced in this chapter. As you will see, a hypothesis test first determines the probability that the pattern could have been produced by chance alone. If this probability is large enough, then we conclude that the pattern can reasonably be explained by chance. However, if the probability is extremely small, then we can rule out chance as a plausible explanation and conclude that some meaningful, systematic force has created the pattern. For example, it is reasonable, once in a lifetime, to see a cloud formation that resembles Santa Claus. However, it would be extremely unlikely if the clouds also included the words "Merry Christmas" spelled out beneath Santa's face. If this happened, we would conclude that the pattern was not produced by the random forces of chance, but rather was created by a deliberate, systematic act.

FIGURE 8.1

A cloud formation seen over Kansas City.

Mark Gravetter

8.1 THE LOGIC OF HYPOTHESIS TESTING

It usually is impossible or impractical for a researcher to observe every individual in a population. Therefore, researchers usually collect data from a sample and then use the sample data to help answer questions about the population. Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest.

Hypothesis testing is one of the most commonly used inferential procedures. In fact, most of the remainder of this book examines hypothesis testing in a variety of different situations and applications. Although the details of a hypothesis test change from one situation to another, the general process remains constant. In this chapter, we introduce the general procedure for a hypothesis test. You should notice that we use the statistical techniques that have been developed in the preceding three chapters—that is, we combine the concepts of *z*-scores, probability, and the distribution of sample means to create a new statistical procedure known as a *hypothesis test*.

DEFINITION

A hypothesis test is a statistical method that uses sample data to evaluate a hypothesis about a population.

In very simple terms, the logic underlying the hypothesis-testing procedure is as follows:

1. First, we state a hypothesis about a population. Usually the hypothesis concerns the value of a population parameter. For example, we might hypothesize that American adults gain an average of $\mu = 7$ pounds between Thanksgiving and New Year's Day each year.
2. Before we select a sample, we use the hypothesis to predict the characteristics that the sample should have. For example, if we predict that the average weight gain for the population is $\mu = 7$ pounds, then we would predict that our sample should have a mean *around* 7 pounds. Remember: The sample should be similar to the population, but you always expect a certain amount of error.
3. Next, we obtain a random sample from the population. For example, we might select a sample of $n = 200$ American adults and measure the average weight change for the sample between Thanksgiving and New Year's Day.
4. Finally, we compare the obtained sample data with the prediction that was made from the hypothesis. If the sample mean is consistent with the prediction, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the prediction, then we decide that the hypothesis is wrong.

A hypothesis test is typically used in the context of a research study. That is, a researcher completes a research study and then uses a hypothesis test to evaluate the results. Depending on the type of research and the type of data, the details of the hypothesis test change from one research situation to another. In later chapters, we examine different versions of hypothesis testing that are used for different kinds of research. For now, however, we focus on the basic elements that are common to all hypothesis tests. To accomplish this general goal, we examine a hypothesis test as it applies to the simplest possible situation—using a sample mean to test a hypothesis about a population mean.

In the six chapters that follow, we consider hypothesis testing in more complex research situations involving sample means and mean differences. In Chapters 15 and 16,

we look at correlational research and examine how the relationships obtained for sample data are used to evaluate hypotheses about relationships in the population. In Chapters 17 and 18, we examine how the proportions that exist in a sample are used to test hypotheses about the corresponding proportions in the population. Chapter 19 reviews the complete set of hypothesis tests and presents a guide to help you find the appropriate test for a specific set of data.

Once again, we introduce hypothesis testing with a situation in which a researcher is using one sample mean to evaluate a hypothesis about one unknown population mean.

The unknown population Figure 8.2 shows the general research situation that we use to introduce the process of hypothesis testing. Notice that the researcher begins with a known population. This is the set of individuals as they exist *before treatment*. For this example, we are assuming that the original set of scores forms a normal distribution with $\mu = 80$ and $\sigma = 20$. The purpose of the research is to determine the effect of a treatment on the individuals in the population. That is, the goal is to determine what happens to the population *after the treatment is administered*.

To simplify the hypothesis-testing situation, one basic assumption is made about the effect of the treatment: If the treatment has any effect, it is simply to add a constant amount to (or subtract a constant amount from) each individual's score. You should recall from Chapters 3 and 4 that adding (or subtracting) a constant changes the mean but does not change the shape of the population, nor does it change the standard deviation. Thus, we assume that the population after treatment has the same shape as the original population and the same standard deviation as the original population. This assumption is incorporated into the situation shown in Figure 8.2.

Note that the unknown population, after treatment, is the focus of the research question. Specifically, the purpose of the research is to determine what would happen if the treatment were administered to every individual in the population.

The sample in the research study The goal of the hypothesis test is to determine whether the treatment has any effect on the individuals in the population (see Figure 8.2). Usually, however, we cannot administer the treatment to the entire population, so the actual research study is conducted using a sample. Figure 8.3 shows the structure of the research study from the point of view of the hypothesis test. The original population, before treatment, is shown on the left-hand side. The unknown population, after treatment, is shown on the right-hand side. Note that the unknown population is actually *hypothetical* (the treatment is never administered to the entire population). Instead, we are asking *what would happen if* the treatment were administered to the entire population. The research

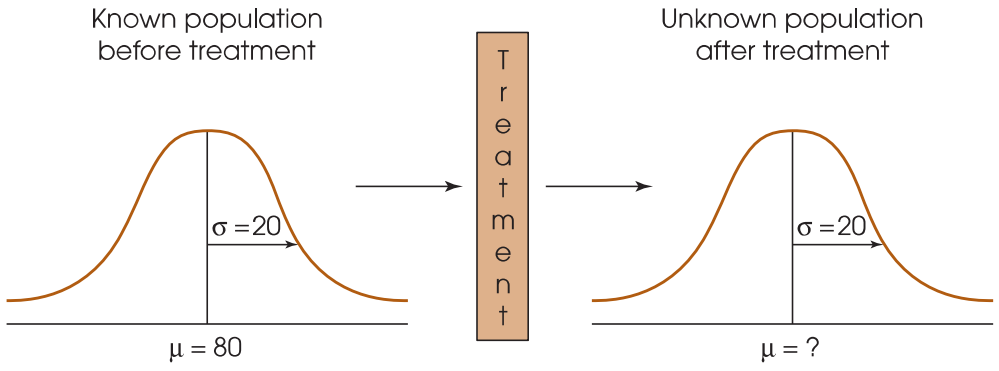
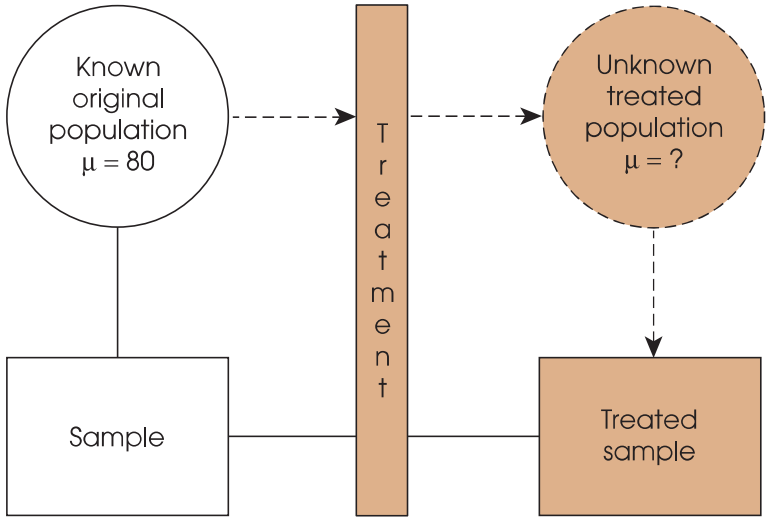


FIGURE 8.2
The basic experimental situation for hypothesis testing. It is assumed that the parameter μ is known for the population before treatment. The purpose of the experiment is to determine whether the treatment has an effect on the population mean.

FIGURE 8.3

From the point of view of the hypothesis test, the entire population receives the treatment and then a sample is selected from the treated population. In the actual research study, a sample is selected from the original population and the treatment is administered to the sample. From either perspective, the result is a treated sample that represents the treated population.



study involves selecting a sample from the original population, administering the treatment to the sample, and then recording scores for the individuals in the treated sample. Notice that the research study produces a treated sample. Although this sample was obtained indirectly, it is equivalent to a sample that is obtained directly from the unknown treated population. The hypothesis test uses the treated sample on the right-hand side of Figure 8.3 to evaluate a hypothesis about the unknown treated population on the right side of the figure.

A hypothesis test is a formalized procedure that follows a standard series of operations. In this way, researchers have a standardized method for evaluating the results of their research studies. Other researchers recognize and understand exactly how the data were evaluated and how conclusions were reached. To emphasize the formal structure of a hypothesis test, we present hypothesis testing as a four-step process that is used throughout the rest of the book. The following example provides a concrete foundation for introducing the hypothesis-testing procedure.

EXAMPLE 8.1

Researchers have noted a decline in cognitive functioning as people age (Bartus, 1990). However, the results from other research suggest that the antioxidants in foods such as blueberries can reduce and even reverse these age-related declines, at least in laboratory rats (Joseph et al., 1999). Based on these results, one might theorize that the same antioxidants might also benefit elderly humans. Suppose a researcher is interested in testing this theory.

Standardized neuropsychological tests such as the Wisconsin Card Sorting Test can be used to measure conceptual thinking ability and mental flexibility (Heaton, Chelune, Talley, Kay, & Curtiss, 1993). Performance on this type of test declines gradually with age. Suppose that our researcher selects a test for which adults older than 65 have an average score of $\mu = 80$ with a standard deviation of $\sigma = 20$. The distribution of test scores is approximately normal. The researcher's plan is to obtain a sample of $n = 25$ adults who are older than 65, and give each participant a daily dose of a blueberry supplement that is very high in antioxidants. After taking the supplement for 6 months, the participants are given the neuropsychological test to measure their level of cognitive function. If the mean score for the sample is

noticeably different from the mean for the general population of elderly adults, then the researcher can conclude that the supplement does appear to have an effect on cognitive function. On the other hand, if the sample mean is around 80 (the same as the general population mean), the researcher must conclude that the supplement does not appear to have any effect.

THE FOUR STEPS OF A HYPOTHESIS TEST

Figure 8.3 depicts the research situation that was described in the preceding example. Notice that the population after treatment is unknown. Specifically, we do not know what will happen to the mean score if the entire population of elderly adults is given the blueberry supplement. However, we do have a sample of $n = 25$ participants who have received the supplement and we can use this sample to help draw inferences about the unknown population. The following four steps outline the hypothesis-testing procedure that allows us to use sample data to answer questions about an unknown population.

STEP 1: STATE THE HYPOTHESIS

As the name implies, the process of hypothesis testing begins by stating a hypothesis about the unknown population. Actually, we state two opposing hypotheses. Notice that both hypotheses are stated in terms of population parameters.

The first, and most important, of the two hypotheses is called the *null hypothesis*. The null hypothesis states that the treatment has no effect. In general, the null hypothesis states that there is no change, no effect, no difference—nothing happened, hence the name *null*. The null hypothesis is identified by the symbol H_0. (The H stands for *hypothesis*, and the zero subscript indicates that this is the *zero-effect* hypothesis.) For the study in Example 8.1, the null hypothesis states that the blueberry supplement has no effect on cognitive functioning for the population of adults who are more than 65 years old. In symbols, this hypothesis is

The goal of inferential statistics is to make general statements about the population by using sample data. Therefore, when testing hypotheses, we make our predictions about the population parameters.

H_0: $\mu_{\text{with supplement}} = 80$ (Even with the supplement, the mean test score is still 80.)

DEFINITION

The **null hypothesis** (H_0) states that in the general population there is no change, no difference, or no relationship. In the context of an experiment, H_0 predicts that the independent variable (treatment) *has no effect* on the dependent variable (scores) for the population.

The second hypothesis is simply the opposite of the null hypothesis, and it is called the *scientific,* or *alternative, hypothesis* (H_1). This hypothesis states that the treatment has an effect on the dependent variable.

DEFINITION

The **alternative hypothesis** (H_1) states that there is a change, a difference, or a relationship for the general population. In the context of an experiment, H_1 predicts that the independent variable (treatment) *does have an effect* on the dependent variable.

The null hypothesis and the alternative hypothesis are mutually exclusive and exhaustive. They cannot both be true, and one of them must be true. The data determine which one should be rejected.

For this example, the alternative hypothesis states that the supplement does have an effect on cognitive functioning for the population and will cause a change in the mean score. In symbols, the alternative hypothesis is represented as

H_1: $\mu_{\text{with supplement}} \neq 80$ (With the supplement, the mean test score is different from 80.)

Notice that the alternative hypothesis simply states that there will be some type of change. It does not specify whether the effect will be increased or decreased test scores. In some circumstances, it is appropriate for the alternative hypothesis to specify the direction of the effect. For example, the researcher might hypothesize that the supplement will increase neuropsychological test scores ($\mu > 80$). This type of hypothesis results in a directional hypothesis test, which is examined in detail later in this chapter. For now we concentrate on nondirectional tests, for which the hypotheses simply state that the treatment has no effect (H_0) or has some effect (H_1).

STEP 2: SET THE CRITERIA FOR A DECISION

Eventually the researcher uses the data from the sample to evaluate the credibility of the null hypothesis. The data either provide support for the null hypothesis or tend to refute the null hypothesis. In particular, if there is a big discrepancy between the data and the null hypothesis, then we conclude that the null hypothesis is wrong.

To formalize the decision process, we use the null hypothesis to predict the kind of sample mean that ought to be obtained. Specifically, we determine exactly which sample means are consistent with the null hypothesis and which sample means are at odds with the null hypothesis.

For our example, the null hypothesis states that the supplement has no effect and the population mean is still $\mu = 80$. If this is true, then the sample mean should have a value around 80. Therefore, a sample mean near 80 is consistent with the null hypothesis. On the other hand, a sample mean that is very different from 80 is not consistent with the null hypothesis. To determine exactly which values are "near" 80 and which values are "very different from" 80, we examine all of the possible sample means that could be obtained if the null hypothesis is true. For our example, this is the distribution of sample means for $n = 25$. According to the null hypothesis, this distribution is centered at $\mu = 80$. The distribution of sample means is then divided into two sections:

1. Sample means that are likely to be obtained if H_0 is true; that is, sample means that are close to the null hypothesis
2. Sample means that are very unlikely to be obtained if H_0 is true; that is, sample means that are very different from the null hypothesis

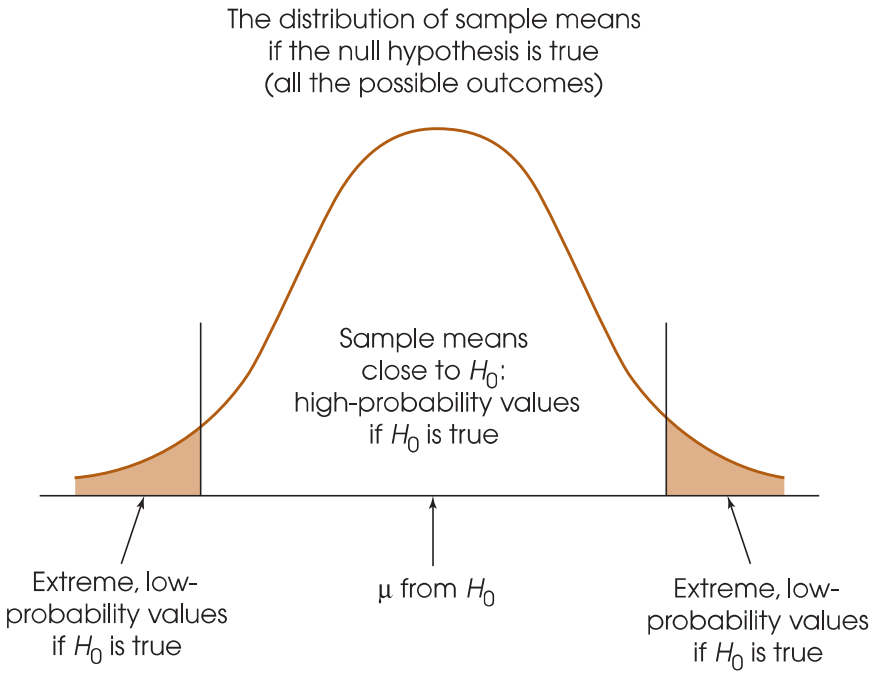
Figure 8.4 shows the distribution of sample means divided into these two sections. Notice that the high-probability samples are located in the center of the distribution and have sample means close to the value specified in the null hypothesis. On the other hand, the low-probability samples are located in the extreme tails of the distribution. After the distribution has been divided in this way, we can compare our sample data with the values in the distribution. Specifically, we can determine whether our sample mean is consistent with the null hypothesis (like the values in the center of the distribution) or whether our sample mean is very different from the null hypothesis (like the values in the extreme tails).

The alpha level To find the boundaries that separate the high-probability samples from the low-probability samples, we must define exactly what is meant by "low" probability and "high" probability. This is accomplished by selecting a specific probability value, which is known as the *level of significance,* or the *alpha level,* for the hypothesis test. The alpha (α) value is a small probability that is used to identify the low-probability samples. By convention, commonly used alpha levels are $\alpha = .05$ (5%), $\alpha = .01$ (1%), and $\alpha = .001$ (0.1%). For example, with $\alpha = .05$, we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values).

With rare exceptions, an alpha level is never larger than .05.

FIGURE 8.4

The set of potential samples is divided into those that are likely to be obtained and those that are very unlikely to be obtained if the null hypothesis is true.



The extremely unlikely values, as defined by the alpha level, make up what is called the *critical region*. These extreme values in the tails of the distribution define outcomes that are not consistent with the null hypothesis; that is, they are very unlikely to occur if the null hypothesis is true. Whenever the data from a research study produce a sample mean that is located in the critical region, we conclude that the data are not consistent with the null hypothesis, and we reject the null hypothesis.

DEFINITIONS

The **alpha level,** or the **level of significance,** is a probability value that is used to define the concept of "very unlikely" in a hypothesis test.

The **critical region** is composed of the extreme sample values that are very unlikely (as defined by the alpha level) to be obtained if the null hypothesis is true. The boundaries for the critical region are determined by the alpha level. If sample data fall in the critical region, the null hypothesis is rejected.

Technically, the critical region is defined by sample outcomes that are *very unlikely* to occur if the treatment has no effect (that is, if the null hypothesis is true). Reversing the point of view, we can also define the critical region as sample values that provide *convincing evidence* that the treatment really does have an effect. For our example, the regular population of elderly adults has a mean test score of $\mu = 80$. We selected a sample from this population and administered a treatment (the blueberry supplement) to the individuals in the sample. What kind of sample mean would convince you that the treatment has an effect? It should be obvious that the most convincing evidence would be a sample mean that is really different from $\mu = 80$. In a hypothesis test, the critical region is determined by sample values that are "really different" from the original population.

The boundaries for the critical region To determine the exact location for the boundaries that define the critical region, we use the alpha-level probability and the unit normal table. In most cases, the distribution of sample means is normal, and the unit normal table provides the precise z-score location for the critical region boundaries. With $\alpha = .05$, for example, the boundaries separate the extreme 5% from the middle 95%. Because the extreme 5% is split between two tails of the distribution, there is exactly 2.5% (or 0.0250) in each tail. In the unit normal table, you can look up a proportion of 0.0250 in column C (the tail) and find that the z-score boundary is $z = 1.96$. Thus, for any normal distribution, the extreme 5% is in the tails of the distribution beyond $z = +1.96$ and $z = -1.96$. These values define the boundaries of the critical region for a hypothesis test using $\alpha = .05$ (Figure 8.5).

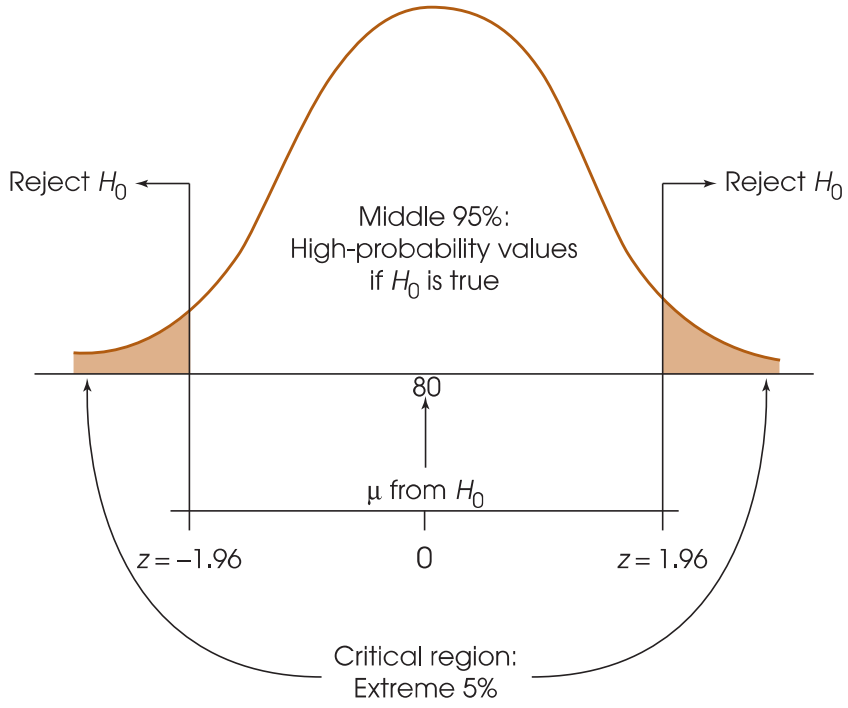
Similarly, an alpha level of $\alpha = .01$ means that 1%, or .0100, is split between the two tails. In this case, the proportion in each tail is .0050, and the corresponding z-score boundaries are $z = \pm 2.58$ (± 2.57 is equally good). For $\alpha = .001$, the boundaries are located at $z = \pm 3.30$. You should verify these values in the unit normal table and be sure that you understand exactly how they are obtained.

LEARNING CHECK

1. The city school district is considering increasing class size in the elementary schools. However, some members of the school board are concerned that larger classes may have a negative effect on student learning. In words, what would the null hypothesis say about the effect of class size on student learning?
2. If the alpha level is increased from $\alpha = .01$ to $\alpha = .05$, then the boundaries for the critical region move farther away from the center of the distribution. (True or false?)
3. If a researcher conducted a hypothesis test with an alpha level of $\alpha = .02$, what z-score values would form the boundaries for the critical region?

FIGURE 8.5

The critical region (very unlikely outcomes) for $\alpha = .05$.



ANSWERS

1. The null hypothesis would say that class size has no effect on student learning.
2. False. A larger alpha means that the boundaries for the critical region move closer to the center of the distribution.
3. The .02 would be split between the two tails, with .01 in each tail. The z-score boundaries would be $z = +2.33$ and $z = -2.33$.

STEP 3: COLLECT DATA AND COMPUTE SAMPLE STATISTICS

At this time, we select a sample of adults who are more than 65 years old and give each one a daily dose of the blueberry supplement. After 6 months, the neuropsychological test is used to measure cognitive function for the sample of participants. Notice that the data are collected *after* the researcher has stated the hypotheses and established the criteria for a decision. This sequence of events helps to ensure that a researcher makes an honest, objective evaluation of the data and does not tamper with the decision criteria after the experimental outcome is known.

Next, the raw data from the sample are summarized with the appropriate statistics: For this example, the researcher would compute the sample mean. Now it is possible for the researcher to compare the sample mean (the data) with the null hypothesis. This is the heart of the hypothesis test: comparing the data with the hypothesis.

The comparison is accomplished by computing a z-score that describes exactly where the sample mean is located relative to the hypothesized population mean from H_0. In step 2, we constructed the distribution of sample means that would be expected if the null hypothesis were true—that is, the entire set of sample means that could be obtained if the treatment has no effect (see Figure 8.5). Now we calculate a z-score that identifies where our sample mean is located in this hypothesized distribution. The z-score formula for a sample mean is

$$z = \frac{M - \mu}{\sigma_M}$$

In the formula, the value of the sample mean (M) is obtained from the sample data, and the value of μ is obtained from the null hypothesis. Thus, the z-score formula can be expressed in words as follows:

$$z = \frac{\text{sample mean} - \text{hypothesized population mean}}{\text{standard error between } M \text{ and } \mu}$$

Notice that the top of the z-score formula measures how much difference there is between the data and the hypothesis. The bottom of the formula measures the standard distance that ought to exist between a sample mean and the population mean.

STEP 4: MAKE A DECISION

In the final step, the researcher uses the z-score value obtained in step 3 to make a decision about the null hypothesis according to the criteria established in step 2. There are two possible outcomes:

1. The sample data are located in the critical region. By definition, a sample value in the critical region is very unlikely to occur if the null hypothesis is true. Therefore, we conclude that the sample is not consistent with H_0 and our decision is to *reject the null hypothesis*. Remember, the null hypothesis states that there is no treatment effect, so rejecting H_0 means that we are concluding that the treatment did have an effect.

For the example we have been considering, suppose that the sample produced a mean of $M = 92$ after taking the supplement for 6 months. The null hypothesis states that the population mean is $\mu = 80$ and, with $n = 25$ and $\sigma = 20$, the standard error for the sample mean is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

Thus, a sample mean of $M = 92$ produces a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{92 - 80}{4} = \frac{12}{4} = 3.00$$

With an alpha level of $\alpha = .05$, this z-score is far beyond the boundary of 1.96. Because the sample z-score is in the critical region, we reject the null hypothesis and conclude that the blueberry supplement did have an effect on cognitive functioning.

2. The second possibility is that the sample data are not in the critical region. In this case, the sample mean is reasonably close to the population mean specified in the null hypothesis (in the center of the distribution). Because the data do not provide strong evidence that the null hypothesis is wrong, our conclusion is to *fail to reject the null hypothesis*. This conclusion means that the treatment does not appear to have an effect.

   For the research study examining the blueberry supplement, suppose our sample produced a mean test score of $M = 84$. As before, the standard error for a sample of $n = 25$ is $\sigma_M = 4$, and the null hypothesis states that $\mu = 80$. These values produce a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{84 - 80}{4} = \frac{4}{4} = 1.00$$

   The z-score of 1.00 is not in the critical region. Therefore, we would fail to reject the null hypothesis and conclude that the blueberry supplement does not appear to have an effect on cognitive functioning.

In general, the final decision is made by comparing our treated sample with the distribution of sample means that would be obtained for untreated samples. If our treated sample looks much the same as samples that do not receive the blueberry treatment, we conclude that the treatment does not appear to have any effect. On the other hand, if the treated sample is noticeably different from the majority of untreated samples, we conclude that the treatment does have an effect.

An Analogy for Hypothesis Testing It may seem awkward to phrase both of the two possible decisions in terms of rejecting the null hypothesis; either we reject H_0 or we fail to reject H_0. These two decisions may be easier to understand if you think of a research study as an attempt to gather evidence to prove that a treatment works. From this perspective, the process of conducting a hypothesis test is similar to the process that takes place during a jury trial. For example,

1. The test begins with a null hypothesis stating that there is no treatment effect. The trial begins with a null hypothesis that the defendant did not commit a crime (innocent until proven guilty).

2. The research study gathers evidence to show that the treatment actually does have an effect, and the police gather evidence to show that the defendant really did commit a crime. Note that both are trying to refute the null hypothesis.
3. If there is enough evidence, the researcher rejects the null hypothesis and concludes that there really is a treatment effect. If there is enough evidence, the jury rejects the hypothesis and concludes that the defendant is guilty of a crime.
4. If there is not enough evidence, the researcher fails to reject the null hypothesis. Note that the researcher does not conclude that there is no treatment effect, simply that there is not enough evidence to conclude that there is an effect. Similarly, if there is not enough evidence, the jury fails to find the defendant guilty. Note that the jury does not conclude that the defendant is innocent, simply that there is not enough evidence for a guilty verdict.

A CLOSER LOOK AT THE *z*-SCORE STATISTIC

The *z*-score statistic that is used in the hypothesis test is the first specific example of what is called a *test statistic*. The term *test statistic* simply indicates that the sample data are converted into a single, specific statistic that is used to test the hypotheses. In the chapters that follow, we introduce several other test statistics that are used in a variety of different research situations. However, most of the new test statistics have the same basic structure and serve the same purpose as the *z*-score. We have already described the *z*-score equation as a formal method for comparing the sample data and the population hypothesis. In this section, we discuss the *z*-score from two other perspectives that may give you a better understanding of hypothesis testing and the role that *z*-scores play in this inferential technique. In each case, keep in mind that the *z*-score serves as a general model for other test statistics that come in future chapters.

The *z*-score formula as a recipe The *z*-score formula, like any formula, can be viewed as a recipe. If you follow instructions and use all of the right ingredients, the formula produces a *z*-score. In the hypothesis-testing situation, however, you do not have all of the necessary ingredients. Specifically, you do not know the value for the population mean (μ), which is one component, or ingredient, in the formula.

This situation is similar to trying to follow a cake recipe in which one of the ingredients is not clearly listed. For example, the recipe may call for flour but there is a grease stain that makes it impossible to read how much flour. Faced with this situation, you might try the following steps:

1. Make a hypothesis about the amount of flour. For example, hypothesize that the correct amount is 2 cups.
2. To test your hypothesis, add the rest of the ingredients along with the hypothesized amount of flour and bake the cake.
3. If the cake turns out to be good, you can reasonably conclude that your hypothesis was correct. But if the cake is terrible, you conclude that your hypothesis was wrong.

In a hypothesis test with *z*-scores, we do essentially the same thing. We have a formula (recipe) for *z*-scores but one ingredient is missing. Specifically, we do not know the value for the population mean, μ. Therefore, we try the following steps:

1. Make a hypothesis about the value of μ. This is the null hypothesis.
2. Plug the hypothesized value in the formula along with the other values (ingredients).

3. If the formula produces a z-score near zero (which is where z-scores are supposed to be), we conclude that the hypothesis was correct. On the other hand, if the formula produces an extreme value (a very unlikely result), we conclude that the hypothesis was wrong.

The z-score formula as a ratio In the context of a hypothesis test, the z-score formula has the following structure:

$$z = \frac{M - \mu}{\sigma_M} = \frac{\text{sample mean} - \text{hypothesized population mean}}{\text{standard error between } M \text{ and } \mu}$$

Notice that the numerator of the formula involves a direct comparison between the sample data and the null hypothesis. In particular, the numerator measures the obtained difference between the sample mean and the hypothesized population mean. The standard error in the denominator of the formula measures the standard amount of distance that exists naturally between a sample mean and the population mean without any treatment effect causing the sample to be different. Thus, the z-score formula (and most other test statistics) forms a ratio

$$z = \frac{\text{actual difference between the sample } (M) \text{ and the hypothesis } (\mu)}{\text{standard difference between } M \text{ and } \mu \text{ with no treatment effect}}$$

Thus, for example, a z-score of $z = 3.00$ means that the obtained difference between the sample and the hypothesis is 3 times bigger than would be expected if the treatment had no effect.

In general, a large value for a test statistic like the z-score indicates a large discrepancy between the sample data and the null hypothesis. Specifically, a large value indicates that the sample data are very unlikely to have occurred by chance alone. Therefore, when we obtain a large value (in the critical region), we conclude that it must have been caused by a treatment effect.

LEARNING CHECK

1. A researcher selects a sample of $n = 16$ individuals from a normal population with a mean of $\mu = 40$ and $\sigma = 8$. A treatment is administered to the sample and, after treatment, the sample mean is $M = 43$. If the researcher uses a hypothesis test to evaluate the treatment effect, what z-score would be obtained for this sample?
2. A small value (near zero) for the z-score statistic is evidence that the sample data are consistent with the null hypothesis. (True or false?)
3. A z-score value in the critical region means that you should reject the null hypothesis. (True or false?)

ANSWERS

1. The standard error is 2 points and $z = 3/2 = 1.50$.
2. True. A z-score near zero indicates that the data support the null hypothesis.
3. True. A z-score value in the critical region means that the sample is not consistent with the null hypothesis.

8.2 UNCERTAINTY AND ERRORS IN HYPOTHESIS TESTING

Hypothesis testing is an *inferential process,* which means that it uses limited information as the basis for reaching a general conclusion. Specifically, a sample provides only limited or incomplete information about the whole population, and yet a hypothesis test uses a sample to draw a conclusion about the population. In this situation, there is always the possibility that an incorrect conclusion will be made. Although sample data are usually representative of the population, there is always a chance that the sample is misleading and will cause a researcher to make the wrong decision about the research results. In a hypothesis test, there are two different kinds of errors that can be made.

TYPE I ERRORS

It is possible that the data will lead you to reject the null hypothesis when in fact the treatment has no effect. Remember: Samples are not expected to be identical to their populations, and some extreme samples can be very different from the populations that they are supposed to represent. If a researcher selects one of these extreme samples by chance, then the data from the sample may give the appearance of a strong treatment effect, even though there is no real effect. In the previous section, for example, we discussed a research study examining how a food supplement that is high in antioxidants affects the cognitive functioning of elderly adults. Suppose that the researcher selects a sample of $n = 25$ people who already have cognitive functioning that is well above average. Even if the blueberry supplement (the treatment) has no effect at all, these people will still score higher than average on the neuropsychological test when they are tested after 6 months of taking the supplement. In this case, the researcher is likely to conclude that the treatment does have an effect, when in fact it really does not. This is an example of what is called a *Type I error.*

DEFINITION

A **Type I error** occurs when a researcher rejects a null hypothesis that is actually true. In a typical research situation, a Type I error means that the researcher concludes that a treatment does have an effect when, in fact, it has no effect.

You should realize that a Type I error is not a stupid mistake in the sense that a researcher is overlooking something that should be perfectly obvious. On the contrary, the researcher is looking at sample data that appear to show a clear treatment effect. The researcher then makes a careful decision based on the available information. The problem is that the information from the sample is misleading.

In most research situations, the consequences of a Type I error can be very serious. Because the researcher has rejected the null hypothesis and believes that the treatment has a real effect, it is likely that the researcher will report or even publish the research results. A Type I error, however, means that this is a false report. Thus, Type I errors lead to false reports in the scientific literature. Other researchers may try to build theories or develop other experiments based on the false results. A lot of precious time and resources may be wasted.

The Probability of a Type I Error A Type I error occurs when a researcher unknowingly obtains an extreme, nonrepresentative sample. Fortunately, the hypothesis test is structured to minimize the risk that this will occur. Figure 8.5 shows the distribution of sample means and the critical region for the research study we have been discussing. This distribution contains all of the possible sample means for samples of $n = 25$ if the null hypothesis is true. Notice that most of the sample means are near the

hypothesized population mean, $\mu = 80$, and that means in the critical region are very unlikely to occur.

With an alpha level of $\alpha = .05$, only 5% of the samples have means in the critical region. Therefore, there is only a 5% probability ($p = .05$) that one of these samples will be obtained. Thus, the alpha level determines the probability of obtaining a sample mean in the critical region when the null hypothesis is true. In other words, the alpha level determines the probability of a Type I error.

DEFINITION

The **alpha level** for a hypothesis test is the probability that the test will lead to a Type I error. That is, the alpha level determines the probability of obtaining sample data in the critical region even though the null hypothesis is true.

In summary, whenever the sample data are in the critical region, the appropriate decision for a hypothesis test is to reject the null hypothesis. Normally this is the correct decision because the treatment has caused the sample to be different from the original population; that is, the treatment effect has pushed the sample mean into the critical region. In this case, the hypothesis test has correctly identified a real treatment effect. Occasionally, however, sample data are in the critical region just by chance, without any treatment effect. When this occurs, the researcher makes a Type I error; that is, the researcher concludes that a treatment effect exists when in fact it does not. Fortunately, the risk of a Type I error is small and is under the control of the researcher. Specifically, the probability of a Type I error is equal to the alpha level.

TYPE II ERRORS

Whenever a researcher rejects the null hypothesis, there is a risk of a Type I error. Similarly, whenever a researcher fails to reject the null hypothesis, there is a risk of a *Type II error*. By definition, a Type II error is the failure to reject a false null hypothesis. In more straightforward English, a Type II error means that a treatment effect really exists, but the hypothesis test fails to detect it.

DEFINITION

A **Type II error** occurs when a researcher fails to reject a null hypothesis that is really false. In a typical research situation, a Type II error means that the hypothesis test has failed to detect a real treatment effect.

A Type II error occurs when the sample mean is not in the critical region even though the treatment has had an effect on the sample. Often this happens when the effect of the treatment is relatively small. In this case, the treatment does influence the sample, but the magnitude of the effect is not big enough to move the sample mean into the critical region. Because the sample is not substantially different from the original population (it is not in the critical region), the statistical decision is to fail to reject the null hypothesis and to conclude that there is not enough evidence to say that there is a treatment effect.

The consequences of a Type II error are usually not as serious as those of a Type I error. In general terms, a Type II error means that the research data do not show the results that the researcher had hoped to obtain. The researcher can accept this outcome and conclude that the treatment either has no effect or has only a small effect that is not worth pursuing, or the researcher can repeat the experiment (usually with some improvement, such as a larger sample) and try to demonstrate that the treatment really does work.

Unlike a Type I error, it is impossible to determine a single, exact probability for a Type II error. Instead, the probability of a Type II error depends on a variety of factors and therefore is a function, rather than a specific number. Nonetheless, the probability of a Type II error is represented by the symbol β, the Greek letter *beta.*

In summary, a hypothesis test always leads to one of two decisions:

1. The sample data provide sufficient evidence to reject the null hypothesis and conclude that the treatment has an effect.
2. The sample data do not provide enough evidence to reject the null hypothesis. In this case, you fail to reject H_0 and conclude that the treatment does not appear to have an effect.

In either case, there is a chance that the data are misleading and the decision is wrong. The complete set of decisions and outcomes is shown in Table 8.1. The risk of an error is especially important in the case of a Type I error, which can lead to a false report. Fortunately, the probability of a Type I error is determined by the alpha level, which is completely under the control of the researcher. At the beginning of a hypothesis test, the researcher states the hypotheses and selects the alpha level, which immediately determines the risk that a Type I error will be made.

SELECTING AN ALPHA LEVEL

As you have seen, the alpha level for a hypothesis test serves two very important functions. First, the alpha level helps to determine the boundaries for the critical region by defining the concept of "very unlikely" outcomes. At the same time, the alpha level determines the probability of a Type I error. When you select a value for alpha at the beginning of a hypothesis test, your decision influences both of these functions.

The primary concern when selecting an alpha level is to minimize the risk of a Type I error. Thus, alpha levels tend to be very small probability values. By convention, the largest permissible value is $\alpha = .05$. When there is no treatment effect, an alpha level of .05 means that there is still a 5% risk, or a 1-in-20 probability, of rejecting the null hypothesis and committing a Type I error. Because the consequences of a Type I error can be relatively serious, many individual researchers and many scientific publications prefer to use a more conservative alpha level such as .01 or .001 to reduce the risk that a false report is published and becomes part of the scientific literature. (For more information on the origins of the .05 level of significance, see the excellent short article by Cowles and Davis, 1982.)

TABLE 8.1
Possible outcomes of a statistical decision

		Actual Situation	
		No Effect, H_0 True	Effect Exists, H_0 False
Experimenter's Decision	Reject H_0	Type I error	Decision correct
	Retain H_0	Decision correct	Type II error

At this point, it may appear that the best strategy for selecting an alpha level is to choose the smallest possible value to minimize the risk of a Type I error. However, there is a different kind of risk that develops as the alpha level is lowered. Specifically, a lower alpha level means less risk of a Type I error, but it also means that the hypothesis test demands more evidence from the research results.

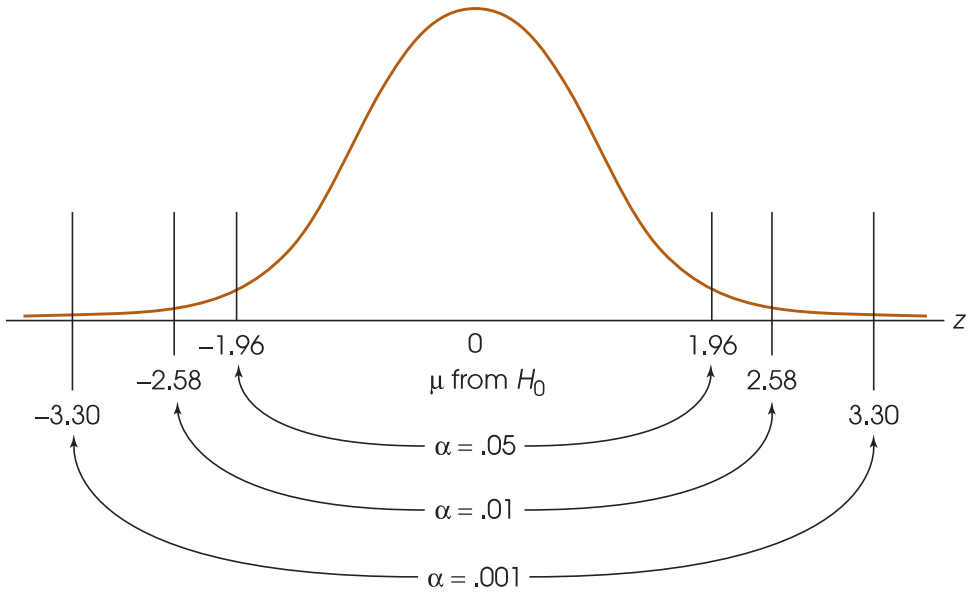
The trade-off between the risk of a Type I error and the demands of the test is controlled by the boundaries of the critical region. For the hypothesis test to conclude that the treatment does have an effect, the sample data must be in the critical region. If the treatment really has an effect, it should cause the sample to be different from the original population; essentially, the treatment should push the sample into the critical region. However, as the alpha level is lowered, the boundaries for the critical region move farther out and become more difficult to reach. Figure 8.6 shows how the boundaries for the critical region move farther into the tails as the alpha level decreases. Notice that $z = 0$, in the center of the distribution, corresponds to the value of μ specified in the null hypothesis. The boundaries for the critical region determine how much distance between the sample mean and μ is needed to reject the null hypothesis. As the alpha level gets smaller, this distance gets larger.

Thus, an extremely small alpha level, such as .000001 (one in a million), would mean almost no risk of a Type I error but would push the critical region so far out that it would become essentially impossible to ever reject the null hypothesis; that is, it would require an enormous treatment effect before the sample data would reach the critical boundaries.

In general, researchers try to maintain a balance between the risk of a Type I error and the demands of the hypothesis test. Alpha levels of .05, .01, and .001 are considered reasonably good values because they provide a low risk of error without placing excessive demands on the research results.

FIGURE 8.6

The locations of the critical region boundaries for three different levels of significance: $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.



LEARNING CHECK

1. Define a Type I error.
2. Define a Type II error.
3. Under what circumstances is a Type II error likely to occur?
4. If a sample mean is in the critical region with $\alpha = .05$, it would still (always) be in the critical region if alpha were changed to $\alpha = .01$. (True or false?)
5. If a sample mean is in the critical region with $\alpha = .01$, it would still (always) be in the critical region if alpha were changed to $\alpha = .05$. (True or false?)

ANSWERS

1. A Type I error is rejecting a true null hypothesis—that is, saying that the treatment has an effect when, in fact, it does not.
2. A Type II error is the failure to reject a false null hypothesis. In terms of a research study, a Type II error occurs when a study fails to detect a treatment effect that really exists.
3. A Type II error is likely to occur when the treatment effect is very small. In this case, a research study is more likely to fail to detect the effect.
4. False. With $\alpha = .01$, the boundaries for the critical region move farther out into the tails of the distribution. It is possible that a sample mean could be beyond the .05 boundary but not beyond the .01 boundary.
5. True. With $\alpha = .01$, the boundaries for the critical region are farther out into the tails of the distribution than for $\alpha = .05$. If a sample mean is beyond the .01 boundary it is definitely beyond the .05 boundary.

8.3 AN EXAMPLE OF A HYPOTHESIS TEST

At this time, we have introduced all the elements of a hypothesis test. In this section, we present a complete example of the hypothesis-testing process and discuss how the results from a hypothesis test are presented in a research report. For purposes of demonstration, the following scenario is used to provide a concrete background for the hypothesis-testing process.

EXAMPLE 8.2

Alcohol appears to be involved in a variety of birth defects, including low birth weight and retarded growth. A researcher would like to investigate the effect of prenatal alcohol exposure on birth weight. A random sample of $n = 16$ pregnant rats is obtained. The mother rats are given daily doses of alcohol. At birth, one pup is selected from each litter to produce a sample of $n = 16$ newborn rats. The average weight for the sample is $M = 15$ grams. The researcher would like to compare the sample with the general population of rats. It is known that regular newborn rats (not exposed to alcohol) have an average weight of $\mu = 18$ grams. The distribution of weights is normal with $\sigma = 4$. Figure 8.7 shows the overall research situation. Notice that the researcher's question concerns the unknown population that is exposed to alcohol. Also notice that we have a sample representing the unknown population, and we have a hypothesis about the unknown population mean. Specifically, the null hypothesis says that the alcohol has no effect and the unknown mean is still $\mu = 18$. The goal of the hypothesis test is to determine whether the sample data are compatible with the hypothesis.

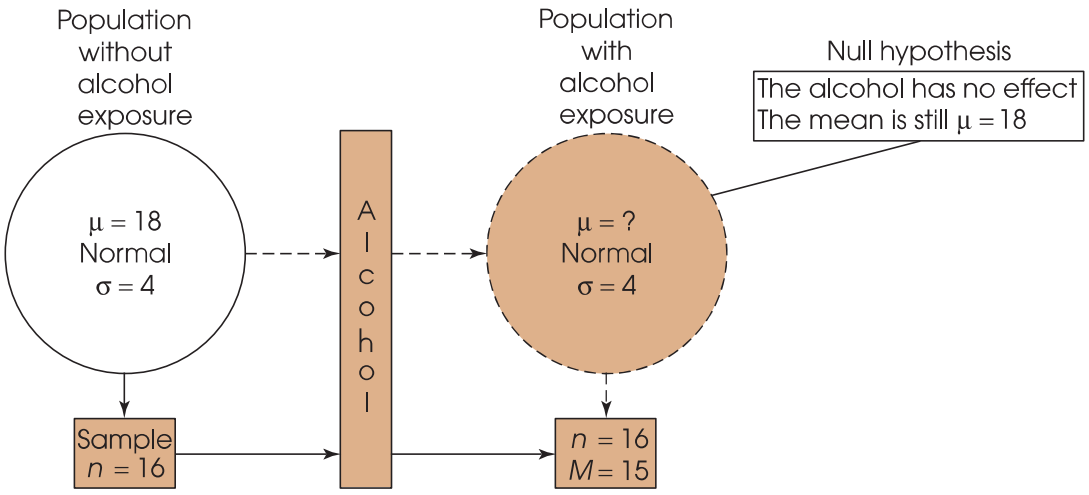


FIGURE 8.7

The structure of a research study to determine whether prenatal alcohol affects birth weight. A sample is selected from the original population and is exposed to alcohol. The question is what would happen if the entire population were exposed to alcohol. The treated sample provides information about the unknown treated population.

The following steps outline the hypothesis test that evaluates the effect of alcohol exposure on birth weight.

STEP 1 *State the hypotheses, and select the alpha level.* Both hypotheses concern the unknown population that is exposed to alcohol (the population on the right-hand side of Figure 8.7). The null hypothesis states that exposure to alcohol has no effect on birth weight. Thus, the population of rats with alcohol exposure should have the same mean birth weight as the regular, unexposed rats. In symbols,

H_0: $\mu_{\text{alcohol exposure}} = 18$ (Even with alcohol exposure, the rats still average 18 grams at birth.)

The alternative hypothesis states that alcohol exposure does affect birth weight, so the exposed population should be different from the regular rats. In symbols,

H_1: $\mu_{\text{alcohol exposure}} \neq 18$ (Alcohol exposure will change birth weight.)

Notice that both hypotheses concern the unknown population. For this test, we will use an alpha level of $\alpha = .05$. That is, we are taking a 5% risk of committing a Type I error.

STEP 2 *Set the decision criteria by locating the critical region.* By definition, the critical region consists of outcomes that are very unlikely if the null hypothesis is true. To locate the critical region we go through a three-stage process that is portrayed in Figure 8.8. We begin with the null hypothesis, which states that the alcohol has no effect on newborn rats. If H_0 is true, the population treated with alcohol is the same as the original population: that is, a normal distribution with $\mu = 18$ and $\sigma = 4$. Next, we consider all

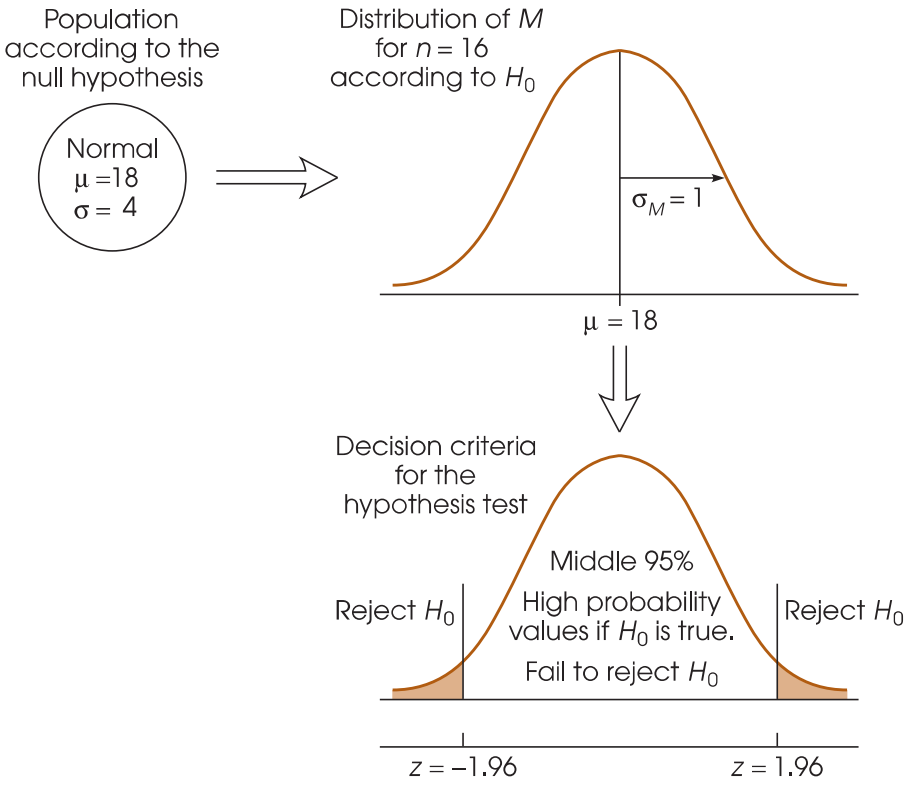


FIGURE 8.8

Locating the critical region as a three-step process. You begin with the population of scores that is predicted by the null hypothesis. Then, you construct the distribution of sample means for the sample size that is being used. The distribution of sample means corresponds to all the possible outcomes that could be obtained if H_0 is true. Finally, you use z-scores to separate the extreme outcomes (as defined by the alpha level) from the high-probability outcomes. The extreme values determine the critical region.

the possible outcomes for a sample of $n = 16$ newborn rats. This is the distribution of sample means for $n = 16$. For this example, the distribution of sample means is normal, is centered at $\mu = 18$ (according to H_0), and has a standard error of $\sigma_M = \frac{4}{\sqrt{16}} = 1$.

Finally, we use the distribution of sample means to identify the critical region, which consists of those outcomes that are very unlikely if the null hypothesis is true. With $\alpha = .05$, the critical region consists of the extreme 5% of the distribution. As we saw earlier, for any normal distribution, z-scores of $z = \pm 1.96$ separate the middle 95% from the extreme 5% (a proportion of 0.0250 in each tail). Thus, we have identified the sample means that, according to the null hypothesis, are very unlikely to occur. It is the unlikely sample means, those with z-score values beyond ± 1.96, that form the critical region for the test. If we obtain a sample mean that is in the critical region, we conclude that the sample is not compatible with the null hypothesis and we reject H_0.

STEP 3 *Collect the data, and compute the test statistic.* At this point, we would select one newborn pup from each of the $n = 16$ mothers that received alcohol during pregnancy. The birth weight is recorded for each pup and the sample mean is computed. For this example, we obtained a sample mean of $M = 15$ grams. The sample mean is then converted to a z-score, which is our test statistic.

$$z = \frac{M - \mu}{\sigma_M} = \frac{15 - 18}{1} = \frac{-3}{1} = -3.00$$

STEP 4 *Make a decision*. The z-score computed in step 3 has a value of -3.00, which is beyond the boundary of -1.96. Therefore, the sample mean is located in the critical region. This is a very unlikely outcome if the null hypothesis is true, so our decision is to reject the null hypothesis. In addition to this statistical decision concerning the null hypothesis, it is customary to state a conclusion about the results of the research study. For this example, we conclude that prenatal exposure to alcohol does have a significant effect on birth weight.

IN THE LITERATURE

REPORTING THE RESULTS OF THE STATISTICAL TEST

A special jargon and notational system are used in published reports of hypothesis tests. When you are reading a scientific journal, for example, you typically are not told explicitly that the researcher evaluated the data using a z-score as a test statistic with an alpha level of .05. Nor are you told that "the null hypothesis is rejected." Instead, you see a statement such as:

> The treatment with alcohol had a significant effect on the birth weight of newborn rats, $z = 3.00$, $p < .05$.

Let us examine this statement, piece by piece. First, what is meant by the word *significant*? In statistical tests, a *significant* result means that the null hypothesis has been rejected, which means that the result is very unlikely to have occurred merely by chance. For this example, the null hypothesis stated that the alcohol has no effect, however the data clearly indicate that the alcohol did have an effect. Specifically, it is very unlikely that the data would have been obtained if the alcohol did not have an effect.

DEFINITION

A result is said to be **significant,** or **statistically significant,** if it is very unlikely to occur when the null hypothesis is true. That is, the result is sufficient to reject the null hypothesis. Thus, a treatment has a significant effect if the decision from the hypothesis test is to reject H_0.

Next, what is the meaning of $z = 3.00$? The z indicates that a z-score was used as the test statistic to evaluate the sample data and that its value is 3.00. Finally, what is meant by $p < .05$? This part of the statement is a conventional way of specifying the alpha level that was used for the hypothesis test. It also acknowledges the possibility (and the probability) of a Type I error. Specifically, the researcher is reporting that the treatment had an effect but admits that this could be a false report. That is, it is possible that the sample mean was in the critical region even though the alcohol had no effect. However, the probability (p) of obtaining a sample mean in the critical region is extremely small (less than .05) if there is no treatment effect.

In circumstances in which the statistical decision is to *fail to reject* H_0, the report might state that.

> There was no evidence that the alcohol had an effect on birth weight, $z = 1.30$, $p > .05$.

The APA style does not use a leading zero in a probability value that refers to a level of significance.

In that case, we would be saying that the obtained result, $z = 1.30$, is not unusual (not in the critical region) and that it has a relatively high probability of occurring (greater than .05) even if the null hypothesis is true and there is no treatment effect.

Sometimes students become confused trying to differentiate between $p < .05$ and $p > .05$. Remember that you reject the null hypothesis with extreme, low-probability values, located in the critical region in the tails of the distribution. Thus, a significant result that rejects the null hypothesis corresponds to $p < .05$ (Figure 8.9).

When a hypothesis test is conducted using a computer program, the printout often includes not only a z-score value but also an exact value for p, the probability that the result occurred without any treatment effect. In this case, researchers are encouraged to report the exact p value instead of using the less-than or greater-than notation. For example, a research report might state that the treatment effect was significant, with $z = 2.45$, $p = .0142$. When using exact values for p, however, you must still satisfy the traditional criterion for significance; specifically, the p value must be smaller than .05 to be considered statistically significant. Remember: The p value is the probability that the result would occur if H_0 were true (without any treatment effect), which is also the probability of a Type I error. It is essential that this probability be very small.

FACTORS THAT INFLUENCE A HYPOTHESIS TEST

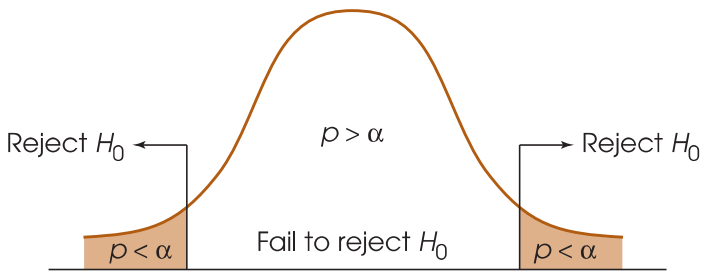
The final decision in a hypothesis test is determined by the value obtained for the z-score statistic. If the z-score is large enough to be in the critical region, then we reject the null hypothesis and conclude that there is a significant treatment effect. Otherwise, we fail to reject H_0 and conclude that the treatment does not have a significant effect. The most obvious factor influencing the size of the z-score is the difference between the sample mean and the hypothesized population mean from H_0. A big mean difference indicates that the treated sample is noticeably different from the untreated population and usually supports a conclusion that the treatment effect is significant. In addition to the mean difference, however, there are other factors that help determine whether the z-score is large enough to reject H_0. In this section we examine two factors that can influence the outcome of a hypothesis test.

1. The variability of the scores, which is measured by either the standard deviation or the variance. The variability influences the size of the standard error in the denominator of the z-score.
2. The number of scores in the sample. This value also influences the size of the standard error in the denominator.

We use the research study from Example 8.2, shown in Figure 8.7, to examine each of these factors. The study used a sample of $n = 16$ newborn rats and concluded that alcohol has a significant effect on birth weight, $z = -3.00$, $p < .05$.

FIGURE 8.9

Sample means that fall in the critical region (shaded areas) have a probability *less than* alpha ($p < \alpha$). In this case, H_0 should be rejected. Sample means that do not fall in the critical region have a probability *greater than* alpha ($p > \alpha$).



The variability of the scores In Chapter 4 (p. 124) we noted that high variability can make it very difficult to see any clear patterns in the results from a research study. In a hypothesis test, higher variability can reduce the chances of finding a significant treatment effect. For the study in Figure 8.7, the standard deviation is $\sigma = 4$. With a sample of $n = 16$, this produced a standard error of $\sigma_M = 1$ point and a significant z-score of $z = -300$. Now consider what happens if the standard deviation is increased to $\sigma = 12$. With the increased variability, the standard error becomes $\sigma_M = 12/\sqrt{16} = 3$ points. Using the same 3-points mean difference from the original example the new z-score becomes

$$z = \frac{M - \mu}{\sigma_M} = \frac{15 - 18}{3} = \frac{-3}{3} = -1.00$$

The z-score is no longer beyond the critical boundary of 1.96, so the statistical decision is to fail to reject the null hypothesis. The increased variability means that the sample data are no longer sufficient to conclude that the treatment has a significant effect. In general, increasing the variability of the scores produces a larger standard error and a smaller value (closer to zero) for the z-score. If other factors are held constant, then the larger the variability, the lower the likelihood of finding a significant treatment effect.

The number of scores in the sample The second factor that influences the outcome of a hypothesis test is the number of scores in the sample. The study in Figure 8.7 used a sample of $n = 16$ rats obtained a standard error of $\sigma_M = 4/\sqrt{16} = 1$ point and a significant z-score of $z = -3.00$. Now consider what happens if we increase the sample size to $n = 64$ rats. With $n = 64$, the standard error becomes $\sigma_M = 4/\sqrt{64} = 0.5$ points, and the z-score becomes

$$z = \frac{M - \mu}{\sigma_M} = \frac{15 - 18}{0.5} = \frac{-3}{0.5} = -6.00$$

Increasing the sample size from $n = 16$ to $n = 64$ has doubled the size of the z-score. In general, increasing the number of scores in the sample produces a smaller standard error and a larger value for the z-score. If all other factors are held constant, the larger the sample size, the greater the likelihood of finding a significant treatment effect. In simple terms, finding a 3-point treatment effect with large sample is more convincing than finding a 3-point effect with a small sample.

ASSUMPTIONS FOR HYPOTHESIS TESTS WITH *z*-SCORES

The mathematics used for a hypothesis test are based on a set of assumptions. When these assumptions are satisfied, you can be confident that the test produces a justified conclusion. However, if the assumptions are not satisfied, then the hypothesis test may be compromised. In practice, researchers are not overly concerned with the assumptions underlying a hypothesis test because the tests usually work well even when the assumptions are violated. However, you should be aware of the fundamental conditions that are associated with each type of statistical test to ensure that the test is being used appropriately. The assumptions for hypothesis tests with z-scores are summarized as follows.

Random sampling It is assumed that the participants used in the study were selected randomly. Remember, we wish to generalize our findings from the sample to the population. Therefore, the sample must be representative of the population from which it has been drawn. Random sampling helps to ensure that it is representative.

Independent observations The values in the sample must consist of *independent* observations. In everyday terms, two observations are independent if there is no consistent, predictable relationship between the first observation and the second. More precisely, two events (or observations) are independent if the occurrence of the first event has no effect on the probability of the second event. Specific examples of independence and non-independence are examined in Box 8.1. Usually, this assumption is satisfied by using a *random* sample, which also helps to ensure that the sample is representative of the population and that the results can be generalized to the population.

The value of σ is unchanged by the treatment A critical part of the *z*-score formula in a hypothesis test is the standard error, σ_M. To compute the value for the standard error, we must know the sample size (n) and the population standard deviation (σ). In a hypothesis test, however, the sample comes from an *unknown* population (see Figures 8.3 and 8.7). If the population is really unknown, it would suggest that we do not know the standard deviation and, therefore, we cannot calculate the standard error. To solve this dilemma, we have made an assumption. Specifically, we assume that the standard deviation for the unknown population (after treatment) is the same as it was for the population before treatment.

BOX 8.1 INDEPENDENT OBSERVATIONS

Independent observations are a basic requirement for nearly all hypothesis tests. The critical concern is that each observation or measurement is not influenced by any other observation or measurement. An example of independent observations is the set of outcomes obtained in a series of coin tosses. Assuming that the coin is balanced, each toss has a 50–50 chance of coming up either heads or tails. More important, each toss is *independent* of the tosses that came before. On the fifth toss, for example, there is a 50% chance of heads no matter what happened on the previous four tosses; the coin does not remember what happened earlier and is not influenced by the past. (*Note:* Many people fail to believe in the independence of events. For example, after a series of four tails in a row, it is tempting to think that the probability of heads must increase because the coin is overdue to come up heads. This is a mistake, called the "gambler's fallacy." Remember that the coin does not know what happened on the preceding tosses and cannot be influenced by previous outcomes.)

In most research situations, the requirement for independent observations is satisfied by using a random sample of separate, unrelated individuals. Thus, the measurement obtained for each individual is not influenced by other participants in the study. The following two situations demonstrate circumstances in which the observations are *not* independent.

1. A researcher is interested in examining television preferences for children. To obtain a sample of $n = 20$ children, the researcher selects 4 children from family A, 3 children from family B, 5 children from family C, 2 children from family D, and 6 children from family E.

   It should be obvious that the researcher does *not* have 20 independent observations. Within each family, the children probably share television preference (at least, they watch the same shows). Thus, the response, for each child is likely to be related to the responses of his or her siblings.
2. The principle of independent observations is violated if the sample is obtained using *sampling without replacement.* For example, if you are selecting from a group of 20 potential participants, each individual has a 1 in 20 chance of being selected first. After the first person is selected, however, there are only 19 people remaining and the probability of being selected changes to 1 in 19. Because the probability of the second selection depends on the first, the two selections are not independent.

Actually, this assumption is the consequence of a more general assumption that is part of many statistical procedures. This general assumption states that the effect of the treatment is to add a constant amount to (or subtract a constant amount from) every score in the population. You should recall that adding (or subtracting) a constant changes the mean but has no effect on the standard deviation. You also should note that this assumption is a theoretical ideal. In actual experiments, a treatment generally does not show a perfect and consistent additive effect.

Normal sampling distribution To evaluate hypotheses with z-scores, we have used the unit normal table to identify the critical region. This table can be used only if the distribution of sample means is normal.

LEARNING CHECK

1. After years of teaching driver's education, an instructor knows that students hit an average of $\mu = 10.5$ orange cones while driving the obstacle course in their final exam. The distribution of run-over cones is approximately normal with a standard deviation of $\sigma = 4.8$. To test a theory about text messaging and driving, the instructor recruits a sample of $n = 16$ student drivers to attempt the obstacle course while sending a text message. The individuals in this sample hit an average of $M = 15.9$ cones.
   a. Do the data indicate that texting has a significant effect on driving? Test with $\alpha = .01$.
   b. Write a sentence describing the outcome of the hypothesis test as it would appear in a research report.
2. In a research report, the term *significant* is used when the null hypothesis is rejected. (True or false?)
3. In a research report, the results of a hypothesis test include the phrase "$z = 3.15$, $p < .01$." This means that the test failed to reject the null hypothesis. (True or false?)
4. If other factors are held constant, increasing the size of the sample increases the likelihood of rejecting the null hypothesis. (True or false?)
5. If other factors are held constant, are you more likely to reject the null hypothesis with a standard deviation of $\sigma = 2$ or with $\sigma = 10$?

ANSWERS

1. a. With $\alpha = .01$, the critical region consists of z-scores in the tails beyond $z = \pm 2.58$. For these data, the standard error is 1.2 and $z = 4.50$. Reject the null hypothesis and conclude that texting has a significant effect on driving.
   b. Texting while driving had a significant effect on the number of cones hit by the participants, $z = 4.50$, $p < .01$.
2. True.
3. False. The probability is *less than* .01, which means it is very unlikely that the result occurred without any treatment effect. In this case, the data are in the critical region, and H_0 is rejected.
4. True. A larger sample produces a smaller standard error, which leads to a larger z-score.
5. $\sigma = 2$. A smaller standard deviation produces a smaller standard error, which leads to larger z-score.

8.4 DIRECTIONAL (ONE-TAILED) HYPOTHESIS TESTS

The hypothesis-testing procedure presented in Section 8.3 is the standard, or *two-tailed,* test format. The term *two-tailed* comes from the fact that the critical region is divided between the two tails of the distribution. This format is by far the most widely accepted procedure for hypothesis testing. Nonetheless, there is an alternative that is discussed in this section.

Usually a researcher begins an experiment with a specific prediction about the direction of the treatment effect. For example, a special training program is expected to *increase* student performance, or alcohol consumption is expected to *slow* reaction times. In these situations, it is possible to state the statistical hypotheses in a manner that incorporates the directional prediction into the statement of H_0 and H_1. The result is a directional test, or what commonly is called a *one-tailed test.*

DEFINITION

In a **directional hypothesis test,** or a **one-tailed test,** the statistical hypotheses (H_0 and H_1) specify either an increase or a decrease in the population mean. That is, they make a statement about the direction of the effect.

The following example demonstrates the elements of a one-tailed hypothesis test.

EXAMPLE 8.3

Earlier, in Example 8.1, we discussed a research study that examined the effect of antioxidants (such as those found in blueberries) on the cognitive skills of elderly adults. In the study, each participant in a sample of $n = 25$ received a blueberry supplement every day for 6 months and then was given a standardized test to measure cognitive skill. For the general population of elderly adults (without any supplement), the test scores form a normal distribution with a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$. For this example, the expected effect is that the blueberry supplement will improve cognitive performance. If the researcher obtains a sample mean of $M = 87$ for the $n = 25$ participants, is the result sufficient to conclude that the supplement really works?

THE HYPOTHESIS FOR A DIRECTIONAL TEST

Because a specific direction is expected for the treatment effect, it is possible for the researcher to perform a directional test. The first step (and the most critical step) is to state the statistical hypotheses. Remember that the null hypothesis states that there is no treatment effect and the alternative hypothesis says that there is an effect. For this example, the predicted effect is that the blueberry supplement will increase test scores. Thus, the two hypotheses would state:

H_0: Test scores are not increased. (The treatment does not work.)

H_1: Test scores are increased. (The treatment works as predicted.)

To express directional hypotheses in symbols, it usually is easier to begin with the alternative hypothesis (H_1). Again, we know that the general population has an average test score of $\mu = 80$, and H_1 states that test scores will be increased by the blueberry supplement. Therefore, expressed in symbols, H_1 states,

H_1: $\mu > 80$ (With the supplement, the average score is greater than 80.)

The null hypothesis states that the supplement does not increase scores. In symbols,

H_0: $\mu \leq 80$ (With the supplement, the average score is not greater than 80.)

Note again that the two hypotheses are mutually exclusive and cover all of the possibilities.

THE CRITICAL REGION FOR DIRECTIONAL TESTS

The critical region is defined by sample outcomes that are very unlikely to occur if the null hypothesis is true (that is, if the treatment has no effect). Earlier (p. 238), we noted that the critical region can also be defined in terms of sample values that provide *convincing evidence* that the treatment really does have an effect. For a directional test, the concept of "convincing evidence" is the simplest way to determine the location of the critical region. We begin with all of the possible sample means that could be obtained if the null hypothesis is true. This is the distribution of sample means and it is normal (because the population of test scores is normal), has an expected value of $\mu = 80$ (from H_0), and, for a sample of $n = 25$, has a standard error of $\sigma_M = 20/\sqrt{25} = 4$. The distribution is shown in Figure 8.10.

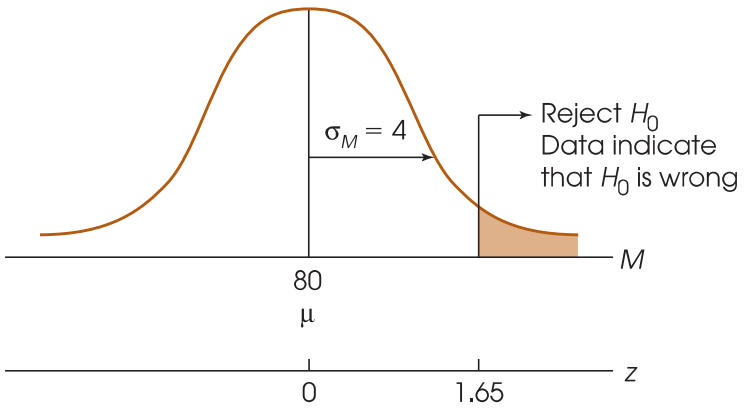
If the prediction is that the treatment will produce a *decrease* in scores, then the critical region is located entirely in the left-hand tail of the distribution.

For this example, the treatment is expected to increase test scores. If untreated adults average $\mu = 80$ on the test, then a sample mean that is substantially more than 80 would provide convincing evidence that the treatment worked. Thus, the critical region is located entirely in the right-hand tail of the distribution corresponding to sample means much greater than $\mu = 80$ (see Figure 8.10). Because the critical region is contained in one tail of the distribution, a directional test is commonly called a *one-tailed* test. Also note that the proportion specified by the alpha level is not divided between two tails, but rather is contained entirely in one tail. Using $\alpha = .05$ for example, the whole 5% is located in one tail. In this case, the z-score boundary for the critical region is $z = 1.65$, which is obtained by looking up a proportion of .05 in column C (the tail) of the unit normal table.

Notice that a directional (one-tailed) test requires two changes in the step-by-step hypothesis-testing procedure.

1. In the first step of the hypothesis test, the directional prediction is incorporated into the statement of the hypotheses.
2. In the second step of the process, the critical region is located entirely in one tail of the distribution.

FIGURE 8.10
Critical region for Example 8.3.



After these two changes, the remainder of a one-tailed test proceeds exactly the same as a regular two-tailed test. Specifically, you calculate the z-score statistic and then make a decision about H_0 depending on whether the z-score is in the critical region.

For this example, the researcher obtained a mean of $M = 87$ for the 25 participants who received the blueberry supplement. This sample mean corresponds to a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{87 - 80}{4} = \frac{7}{4} = 1.75$$

A z-score of $z = 1.75$ is in the critical region for a one-tailed test (see Figure 8.10). This is a very unlikely outcome if H_0 is true. Therefore, we reject the null hypothesis and conclude that the blueberry supplement produces a significant increase in cognitive performance scores. In the literature, this result would be reported as follows:

> The supplement produced a significant increase in scores, $z = 1.75$, $p < .05$, one tailed.

Note that the report clearly acknowledges that a one-tailed test was used.

COMPARISON OF ONE-TAILED VERSUS TWO-TAILED TESTS

The general goal of hypothesis testing is to determine whether a particular treatment has any effect on a population. The test is performed by selecting a sample, administering the treatment to the sample, and then comparing the result with the original population. If the treated sample is noticeably different from the original population, then we conclude that the treatment has an effect, and we reject H_0. On the other hand, if the treated sample is still similar to the original population, then we conclude that there is no convincing evidence for a treatment effect, and we fail to reject H_0. The critical factor in this decision is the *size of the difference* between the treated sample and the original population. A large difference is evidence that the treatment worked; a small difference is not sufficient to say that the treatment had any effect.

The major distinction between one-tailed and two-tailed tests is the criteria that they use for rejecting H_0. A one-tailed test allows you to reject the null hypothesis when the difference between the sample and the population is relatively small, provided that the difference is in the specified direction. A two-tailed test, on the other hand, requires a relatively large difference independent of direction. This point is illustrated in the following example.

EXAMPLE 8.4

Consider again the one-tailed test evaluating the effect of an antioxidant supplement. If we had used a standard two-tailed test, the hypotheses would be

H_0: $\mu = 80$ (The supplement has no effect on test scores.)

H_1: $\mu \neq 80$ (The supplement does have an effect on test scores.)

For a two-tailed test with $\alpha = .05$, the critical region consists of z-scores beyond ± 1.96. The data from Example 8.3 produced a sample mean of $M = 87$ and $z = 1.75$. For the two-tailed test, this z-score is not in the critical region, and we conclude that the supplement does not have a significant effect.

With the two-tailed test in Example 8.4, the 7-point difference between the sample mean and the hypothesized population mean ($M = 87$ and $\mu = 80$) is not big enough to reject the null hypothesis. However, with the one-tailed test introduced in

Example 8.3, the same 7-point difference is large enough to reject H_0 and conclude that the treatment had a significant effect.

All researchers agree that one-tailed tests are different from two-tailed tests. However, there are several ways to interpret the difference. One group of researchers contends that a two-tailed test is more rigorous and, therefore, more convincing than a one-tailed test. Remember that the two-tailed test demands more evidence to reject H_0 and thus provides a stronger demonstration that a treatment effect has occurred.

Other researchers feel that one-tailed tests are preferable because they are more sensitive. That is, a relatively small treatment effect may be significant with a one-tailed test but fail to reach significance with a two-tailed test. Also, there is the argument that one-tailed tests are more precise because they test hypotheses about a specific directional effect instead of an indefinite hypothesis about a general effect.

In general, two-tailed tests should be used in research situations when there is no strong directional expectation or when there are two competing predictions. For example, a two-tailed test would be appropriate for a study in which one theory predicts an increase in scores but another theory predicts a decrease. One-tailed tests should be used only in situations in which the directional prediction is made before the research is conducted and there is a strong justification for making the directional prediction. In particular, if a two-tailed test fails to reach significance, you should never follow up with a one-tailed test as a second attempt to salvage a significant result for the same data.

LEARNING CHECK

1. If a researcher predicts that a treatment will increase scores, then the critical region for a one-tailed test would be located in the right-hand tail of the distribution. (True or false?)
2. If the sample data are sufficient to reject the null hypothesis for a one-tailed test, then the same data would also reject H_0 for a two-tailed test. (True or false?)
3. A researcher obtains $z = 2.43$ for a hypothesis test. Using $\alpha = .01$, the researcher should reject the null hypothesis for a one-tailed test but fail to reject for a two-tailed test. (True or false?)

ANSWERS

1. True. A large sample mean, in the right-hand tail, would indicate that the treatment worked as predicted.
2. False. Because a two-tailed test requires a larger mean difference, it is possible for a sample to be significant for a one-tailed test but not for a two-tailed test.
3. True. The one-tailed critical value is $z = 2.33$ and the two-tailed value is $z = 2.58$.

8.5 CONCERNS ABOUT HYPOTHESIS TESTING: MEASURING EFFECT SIZE

Although hypothesis testing is the most commonly used technique for evaluating and interpreting research data, a number of scientists have expressed a variety of concerns about the hypothesis testing procedure (for example, see Loftus, 1996; Hunter, 1997; and Killeen, 2005).

There are two serious limitations with using a hypothesis test to establish the significance of a treatment effect. The first concern is that the focus of a hypothesis test is on the data rather than the hypothesis. Specifically, when the null hypothesis is rejected,

we are actually making a strong probability statement about the sample data, not about the null hypothesis. A significant result permits the following conclusion: "This specific sample mean is very unlikely ($p < .05$) if the null hypothesis is true." Note that the conclusion does not make any definite statement about the probability of the null hypothesis being true or false. The fact that the data are very unlikely *suggests* that the null hypothesis is also very unlikely, but we do not have any solid grounds for making a probability statement about the null hypothesis. Specifically, you cannot conclude that the probability of the null hypothesis being true is less than 5% simply because you rejected the null hypothesis with $\alpha = .05$ (see Box 8.2).

A second concern is that demonstrating a *significant* treatment effect does not necessarily indicate a *substantial* treatment effect. In particular, statistical significance

BOX 8.2 A FLAW IN THE LOGIC OF HYPOTHESIS TESTING

Suppose that you do a hypothesis test and reject the null hypothesis with $\alpha = .05$. Can you conclude that there is a 5% probability that you are making a Type I error? Can you also conclude that there is a 95% probability that your decision is correct and the treatment does have an effect? For both questions, the answer is no.

The problem is that the probabilities for a hypothesis test are well defined only when the null hypothesis is true. Specifically, a hypothesis test using $\alpha = .05$ is structured so that the error rate is $p < .05$ and the accuracy rate is $p \geq .95$ *if the null hypothesis is true*. If H_0 is false, however, these probabilities start to fall apart. When there is a treatment effect (H_0 is false), the probability that a hypothesis test will detect it and reject H_0 depends on a variety of factors. For example, if the treatment effect is very small, then a hypothesis test is unlikely to detect it. With a large treatment effect, the hypothesis test is more likely to detect it and the probability of rejecting H_0 increases. Thus, whenever there is a treatment effect (H_0 is false), it becomes impossible to define precisely the probability of rejecting the null hypothesis.

Most researchers begin research studies believing that there is a good likelihood that the null hypothesis is false and there really is a treatment effect. They are hoping that the study will provide evidence of the effect so they can convince their colleagues. Thus, most research begins with some probability that the null hypothesis is false. For the sake of argument, let's assume that there is an 80% probability that the null hypothesis is true.

p(there is no treatment effect—H_0 is true) = 0.80 and
p(there is a treatment effect—H_0 is false) = 0.20

In this situation, suppose that 125 researchers are all doing hypothesis tests with $\alpha = .05$. Of these researchers, 80% ($n = 100$) are testing a true H_0. For these researchers, the probability of rejecting the null hypothesis (and making a Type I error) is $\alpha = .05$. Therefore, the 100 hypothesis tests for this group should produce, at most, 5 tests that reject H_0.

Meanwhile, the other 20% of the researchers ($n = 25$) are testing a false null hypothesis. For this group, the probability of rejecting the null hypothesis is unknown. For the sake of argument, however, let's assume that the probability of detecting the treatment effect and correctly rejecting H_0 is 60%. This means that the 25 hypothesis tests should result in 15 tests (60%) that reject H_0 and 10 that fail to reject H_0.

Notice that there could be as many as 20 hypothesis tests that reject the null hypothesis (5 from the first group and 15 from the second group). Thus, a total of 20 researchers will find a statistically significant effect. Of these 20 "significant" results, however, the 5 from the first group are making a Type I error. In this case, the probability of a Type I error is 5 out of 20, or $p = 5/20 = .25$, which is five times greater than the alpha level of .05.

Based on this kind of argument, many scientists suspect that a large number of the results and conclusions published in research journals are simply wrong. Specifically, the Type I error rate in published research is almost certainly higher than the alpha levels used in the hypothesis tests that support the results (Siegfried, 2010).

does not provide any real information about the absolute size of a treatment effect. Instead, the hypothesis test has simply established that the results obtained in the research study are very unlikely to have occurred if there is no treatment effect. The hypothesis test reaches this conclusion by (1) calculating the standard error, which measures how much difference is reasonable to expect between M and μ, and (2) demonstrating that the obtained mean difference is substantially bigger than the standard error.

Notice that the test is making a *relative* comparison: the size of the treatment effect is being evaluated relative to the standard error. If the standard error is very small, then the treatment effect can also be very small and still be large enough to be significant. Thus, a significant effect does not necessarily mean a big effect.

The idea that a hypothesis test evaluates the relative size of a treatment effect, rather than the absolute size, is illustrated in the following example.

EXAMPLE 8.5

We begin with a population of scores that forms a normal distribution with $\mu = 50$ and $\sigma = 10$. A sample is selected from the population and a treatment is administered to the sample. After treatment, the sample mean is found to be $M = 51$. Does this sample provide evidence of a statistically significant treatment effect?

Although there is only a 1-point difference between the sample mean and the original population mean, the difference may be enough to be significant. In particular, the outcome of the hypothesis test depends on the sample size.

For example, with a sample of $n = 25$ the standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2.00$$

and the z-score for $M = 51$ is

$$z = \frac{M - \mu}{\sigma_M} = \frac{51 - 50}{2} = \frac{1}{2} = 0.50$$

This z-score fails to reach the critical boundary of $z = 1.96$, so we fail to reject the null hypothesis. In this case, the 1-point difference between M and μ is not significant because it is being evaluated relative to a standard error of 2 points.

Now consider the outcome with a sample of $n = 400$. With a larger sample, the standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{400}} = \frac{10}{20} = 0.50$$

and the z-score for $M = 51$ is

$$z = \frac{M - \mu}{\sigma_M} = \frac{51 - 50}{0.5} = \frac{1}{0.5} = 2.00$$

Now the z-score is beyond the 1.96 boundary, so we reject the null hypothesis and conclude that there is a significant effect. In this case, the 1-point difference between M and μ is considered statistically significant because it is being evaluated relative to a standard error of only 0.5 points.

The point of Example 8.5 is that a small treatment effect can still be statistically significant. If the sample size is large enough, any treatment effect, no matter how small, can be enough for us to reject the null hypothesis.

MEASURING EFFECT SIZE

As noted in the previous section, one concern with hypothesis testing is that a hypothesis test does not really evaluate the absolute size of a treatment effect. To correct this problem, it is recommended that whenever researchers report a statistically significant effect, they also provide a report of the effect size (see the guidelines presented by L. Wilkinson and the APA Task Force on Statistical Inference, 1999). Therefore, as we present different hypothesis tests we also present different options for measuring and reporting *effect size.*

DEFINITION

A measure of **effect size** is intended to provide a measurement of the absolute magnitude of a treatment effect, independent of the size of the sample(s) being used.

One of the simplest and most direct methods for measuring effect size is *Cohen's d.* Cohen (1988) recommended that effect size can be standardized by measuring the mean difference in terms of the standard deviation. The resulting measure of effect size is computed as

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{\mu_{\text{treatment}} - \mu_{\text{no treatment}}}{\sigma} \tag{8.1}$$

For the *z*-score hypothesis test, the mean difference is determined by the difference between the population mean before treatment and the population mean after treatment. However, the population mean after treatment is unknown. Therefore, we must use the mean for the treated sample in its place. Remember, the sample mean is expected to be representative of the population mean and provides the best measure of the treatment effect. Thus, the actual calculations are really estimating the value of Cohen's *d* as follows:

Cohen's *d* measures the distance between two means and is typically reported as a positive number even when the formula produces a negative value.

$$\text{estimated Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{M_{\text{treatment}} - \mu_{\text{no treatment}}}{\sigma} \tag{8.2}$$

The standard deviation is included in the calculation to standardize the size of the mean difference in much the same way that *z*-scores standardize locations in a distribution. For example, a 15-point mean difference can be a relatively large treatment effect or a relatively small effect depending on the size of the standard deviation. This phenomenon is demonstrated in Figure 8.11. The top portion of the figure (part a) shows the results of a treatment that produces a 15-point mean difference in SAT scores; before treatment, the average SAT score is $\mu = 500$, and after treatment the average is 515. Notice that the standard deviation for SAT scores is $\sigma = 100$, so the 15-point difference appears to be small. For this example, Cohen's *d* is

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{15}{100} = 0.15$$

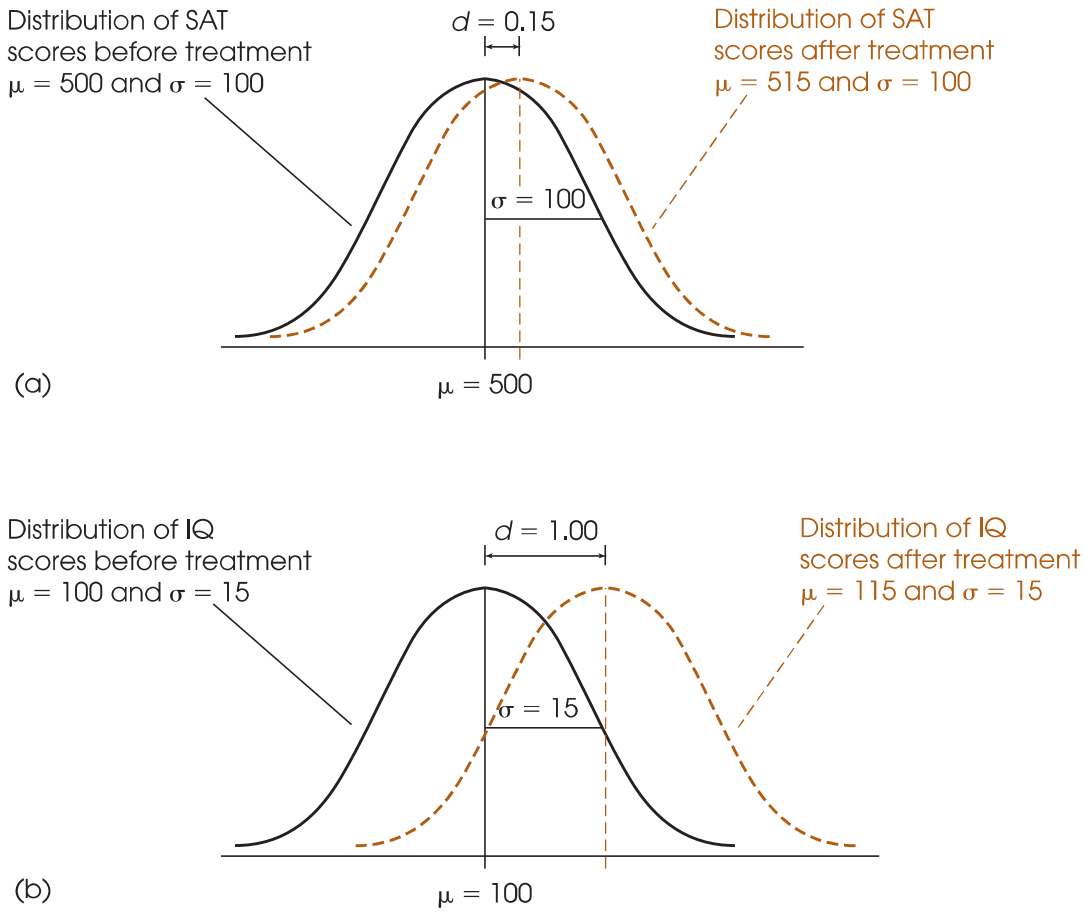


FIGURE 8.11

The appearance of a 15-point treatment effect in two different situations. In part (a), the standard deviation is $\sigma = 100$ and the 15-point effect is relatively small. In part (b), the standard deviation is $\sigma = 15$ and the 15-point effect is relatively large. Cohen's *d* uses the standard deviation to help measure effect size.

Now consider the treatment effect shown in Figure 8.11(b). This time, the treatment produces a 15-point mean difference in IQ scores; before treatment the average IQ is 100, and after treatment the average is 115. Because IQ scores have a standard deviation of $\sigma = 15$, the 15-point mean difference now appears to be large. For this example, Cohen's *d* is

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{15}{15} = 1.00$$

Notice that Cohen's *d* measures the size of the treatment effect in terms of the standard deviation. For example, a value of $d = 0.50$ indicates that the treatment changed the mean by half of a standard deviation; similarly, a value of $d = 1.00$ indicates that the size of the treatment effect is equal to one whole standard deviation. (See Box 8.3.)

Cohen (1988) also suggested criteria for evaluating the size of a treatment effect as shown in Table 8.2.

BOX 8.3 OVERLAPPING DISTRIBUTIONS

Figure 8.11(b) shows the results of a treatment with a Cohen's *d* of 1.00; that is, the effect of the treatment is to increase the mean by one full standard deviation. According to the guidelines in Table 8.2, a value of $d = 1.00$ is considered a large treatment effect. However, looking at the figure, you may get the impression that there really isn't that much difference between the distribution before treatment and the distribution after treatment. In particular, there is substantial overlap between the two distributions, so that many of the individuals who receive the treatment are not any different from the individuals who do not receive the treatment.

The overlap between distributions is a basic fact of life in most research situations; it is extremely rare for the scores after treatment to be *completely different* (no overlap) from the scores before treatment. Consider, for example, children's heights at different ages. Everyone knows that 8-year-old children are taller than 6-year-old children; on average, the difference is 3 or 4 inches. However, this does not mean that all 8-year-old children are taller than all 6-year-old children. In fact, there is considerable overlap between the two distributions, so that the tallest among the 6-year-old children are actually taller than most 8-year-old children. In fact, the height distributions for the two age groups would look a lot like the two distributions in Figure 8.10(b). Although there is a clear *mean difference* between the two distributions, there still can be substantial overlap.

Cohen's *d* measures the degree of separation between two distributions, and a separation of one standard deviation ($d = 1.00$) represents a large difference. Eight-year-old children really are bigger than 6-year-old children.

TABLE 8.2
Evaluating effect size with Cohen's *d*.

Magnitude of *d*	Evaluation of Effect Size
$d = 0.2$	Small effect (mean difference around 0.2 standard deviation)
$d = 0.5$	Medium effect (mean difference around 0.5 standard deviation)
$d = 0.8$	Large effect (mean difference around 0.8 standard deviation)

As one final demonstration of Cohen's *d*, consider the two hypothesis tests in Example 8.5. For each test, the original population had a mean of $\mu = 50$ with a standard deviation of $\sigma = 10$. For each test, the mean for the treated sample was $M = 51$. Although one test used a sample of $n = 25$ and the other test used a sample of $n = 400$, the sample size is not considered when computing Cohen's *d*. Therefore, both of the hypothesis tests would produce the same value:

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{15}{15} = 1.00$$

Notice that Cohen's *d* simply describes the size of the treatment effect and is not influenced by the number of scores in the sample. For both hypothesis tests, the original population mean was $\mu = 50$ and, after treatment, the sample mean was $M = 51$. Thus, treatment appears to have increased the scores by 1 point, which is equal to one-tenth of a standard deviation (Cohen's $d = 0.1$).

LEARNING CHECK

1. **a.** How does increasing sample size influence the outcome of a hypothesis test?
   b. How does increasing sample size influence the value of Cohen's *d*?
2. A researcher selects a sample from a population with $\mu = 45$ and $\sigma = 8$. A treatment is administered to the sample and, after treatment, the sample mean is found to be $M = 47$. Compute Cohen's *d* to measure the size of the treatment effect.

ANSWERS

1. **a.** Increasing sample size increases the likelihood of rejecting the null hypothesis.
   b. Cohen's *d* is not influenced at all by the sample size.
2. $d = 2/8 = 0.25$

8.6 STATISTICAL POWER

Instead of measuring effect size directly, an alternative approach to determining the size or strength of a treatment effect is to measure the power of the statistical test. The *power* of a test is defined as the probability that the test will reject the null hypothesis if the treatment really has an effect.

DEFINITION

The **power** of a statistical test is the probability that the test will correctly reject a false null hypothesis. That is, power is the probability that the test will identify a treatment effect if one really exists.

Whenever a treatment has an effect, there are only two possible outcomes for a hypothesis test: either fail to reject H_0 or reject H_0. Because there are only two possible outcomes, the probability for the first and the probability for the second must add up to 1.00. The first outcome, failing to reject H_0 when there is a real effect, was defined earlier (p. 245) as a Type II error with a probability identified as $p = \beta$. Therefore, the second outcome must have a probability of $1 - \beta$. However, the second outcome, rejecting H_0 when there is a real effect, is the power of the test. Thus, the power of a hypothesis test is equal to $1 - \beta$. In the examples that follow, we demonstrate the calculation of power for a hypothesis test; that is, the probability that the test will correctly reject the null hypothesis. At the same time, however, we are computing the probability that the test will result in a Type II error. For example, if the power of the test is 70% ($1 - \beta$) then the probability of a Type II error must be 30% (β).

Researchers typically calculate power as a means of determining whether a research study is likely to be successful. Thus, researchers usually calculate the power of a hypothesis test *before* they actually conduct the research study. In this way, they can determine the probability that the results will be significant (reject H_0) before investing time and effort in the actual research. To calculate power, however, it is first necessary to make assumptions about a variety of factors that influence the outcome of a hypothesis test. Factors such as the sample size, the size of the treatment effect, and the value chosen for the alpha level can all influence a hypothesis test. The following example demonstrates the calculation of power for a specific research situation.

EXAMPLE 8.6

We start with a normal-shaped population with a mean of $\mu = 80$ and a standard deviation of $\sigma = 10$. A researcher plans to select a sample of $n = 25$ individuals from

this population and administer a treatment to each individual. It is expected that the treatment will have an 8-point effect; that is, the treatment will add 8 points to each individual's score.

Figure 8.12 shows the original population distribution and two possible outcomes:

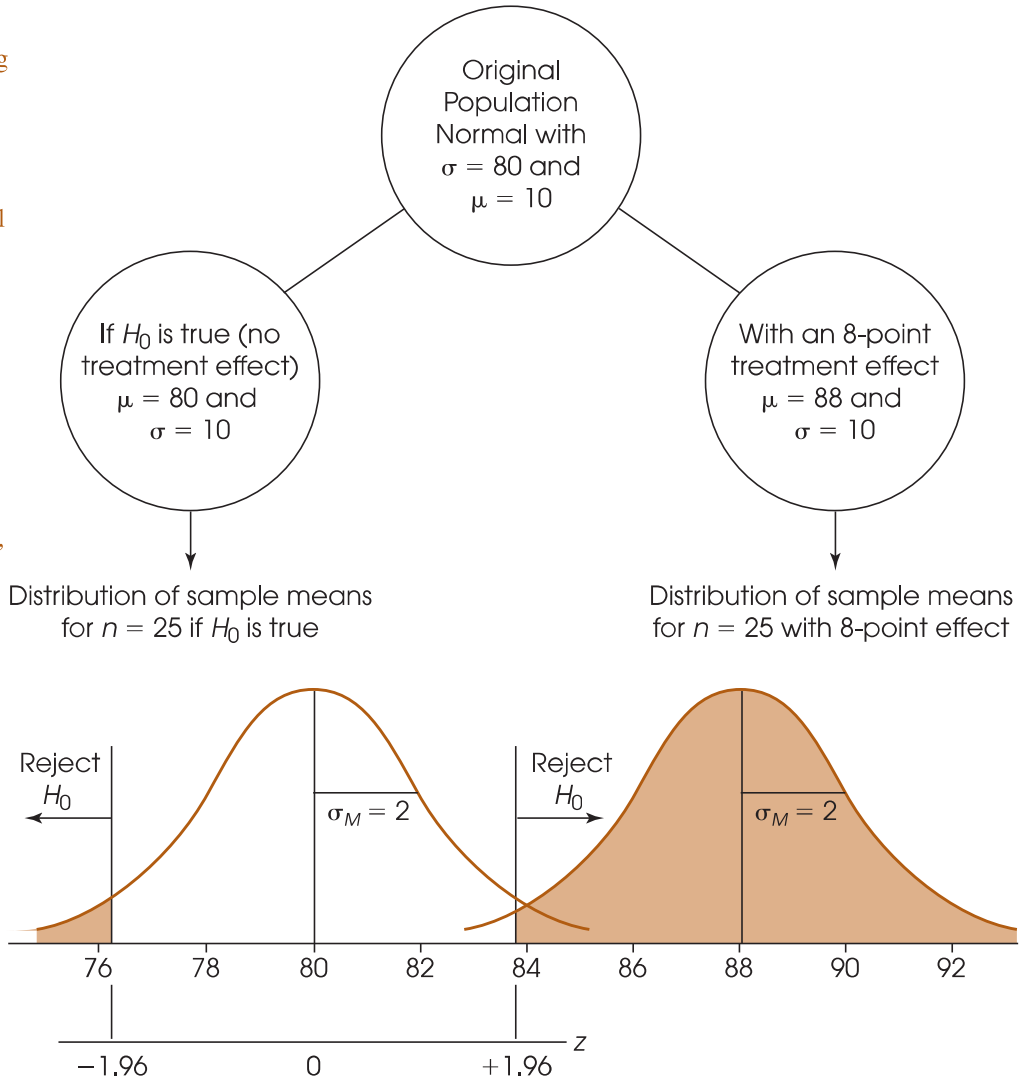
1. If the null hypothesis is true and there is no treatment effect.
2. If the researcher's expectation is correct and there is an 8-point effect.

The left-hand side of the figure shows what should happen according to the null hypothesis. In this case, the treatment has no effect and the population mean is still $\mu = 80$. On the right-hand side of the figure we show what would happen if the treatment has an 8-point effect. If the treatment adds 8 points to each person's score, then the population mean after treatment increases to $\mu = 88$.

Beneath each of the two populations, Figure 8.12 shows the distribution of sample means for $n = 25$. According to the null hypothesis, the sample means are centered

FIGURE 8.12

A demonstration of measuring power for a hypothesis test. The left-hand side shows the distribution of sample means that would occur if the null hypothesis is true. The critical region is defined for this distribution. The right-hand side shows the distribution of sample means that would be obtained if there were an 8-point treatment effect. Notice that if there is an 8-point effect, essentially all of the sample means would be in the critical region. Thus, the probability of rejecting H_0 (the power of the test) would be nearly 100% for an 8-point treatment effect.



around $\mu = 80$. With an 8-point treatment effect, the sample means are centered around $\mu = 88$. Both distributions have a standard error of

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

Notice that the distribution on the left shows all of the possible sample means if the null hypothesis is true. This is the distribution we use to locate the critical region for the hypothesis test. Using $\alpha = .05$, the critical region consists of extreme values in this distribution, specifically sample means beyond $z = 1.96$ or $z = -1.96$. These values are shown in Figure 8.12, and we have shaded all of the sample means located in the critical region.

Now turn your attention to the distribution on the right, which shows all of the possible sample means if there is an 8-point treatment effect. Notice that most of these sample means are located beyond the $z = 1.96$ boundary. This means that, if there is an 8-point treatment effect, you are almost guaranteed to obtain a sample mean in the critical region and reject the null hypothesis. Thus, the power of the test (the probability of rejecting H_0) is close to 100% if there is an 8-point treatment effect.

To calculate the exact value for the power of the test we must determine what portion of the distribution on the right-hand side is shaded. Thus, we must locate the exact boundary for the critical region, then find the probability value in the unit normal table. For the distribution on the left-hand side, the critical boundary of $z = +1.96$ corresponds to a location that is above $\mu = 80$ by a distance equal to

$$1.96\mu_M = 1.96(2) = 3.92 \text{ points}$$

Thus, the critical boundary of $z = +1.96$ corresponds to a sample mean of $M = 80 + 3.92 = 83.92$. Any sample mean greater than $M = 83.92$ is in the critical region and would lead to rejecting the null hypothesis. Next, we determine what proportion of the treated samples are greater than $M = 83.92$. For the treated distribution (right-hand side), the population mean is $\mu = 88$ and a sample mean of $M = 83.92$ corresponds to a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{83.92 - 88}{2} = \frac{-4.08}{2} = -2.04$$

Finally, look up $z = -2.04$ in the unit normal table and determine that the shaded area ($z > -2.04$) corresponds to $p = 0.9793$ (or 97.93%). Thus, if the treatment has an 8-point effect, 97.93% of all the possible sample means will be in the critical region and we will reject the null hypothesis. In other words, the power of the test is 97.93%. In practical terms, this means that the research study is almost guaranteed to be successful. If the researcher selects a sample of $n = 25$ individuals, and if the treatment really does have an 8-point effect, then 97.93% of the time the hypothesis test will conclude that there is a significant effect.

POWER AND EFFECT SIZE

Logically, it should be clear that power and effect size are related. Figure 8.12 shows the calculation of power for an 8-point treatment effect. Now consider what would happen if the treatment effect were only 4 points. With a 4-point treatment effect, the distribution on the right-hand side would shift to the left so that it is centered at $\mu = 84$.

In this new position, only about 50% of the treated sample means would be beyond the $z = 1.96$ boundary. Thus, with a 4-point treatment effect, there is only a 50% probability of selecting a sample that leads to rejecting the null hypothesis. In other words, the power of the test is only about 50% for a 4-point effect compared to nearly 98% with an 8-point effect (Example 8.6). Again, it is possible to find the z-score corresponding to the exact location of the critical boundary and to look up the probability value for power in the unit normal table. In this case, you should obtain $z = -0.04$ and the exact power of the test is $p = 0.5160$, or 51.60%.

In general, as the effect size increases, the distribution of sample means on the right-hand side moves even farther to the right so that more and more of the samples are beyond the $z = 1.96$ boundary. Thus, as the effect size increases, the probability of rejecting H_0 also increases, which means that the power of the test increases. Thus, measures of effect size such as Cohen's d and measures of power both provide an indication of the strength or magnitude of a treatment effect.

OTHER FACTORS THAT AFFECT POWER

Although the power of a hypothesis test is directly influenced by the size of the treatment effect, power is not meant to be a pure measure of effect size. Instead, power is influenced by several factors, other than effect size, that are related to the hypothesis test. Some of these factors are considered in the following section.

Sample size One factor that has a huge influence on power is the size of the sample. In Example 8.6 we demonstrated power for an 8-point treatment effect using a sample of $n = 25$. If the researcher decided to conduct the study using a sample of $n = 4$, then the power would be dramatically different. With $n = 4$, the standard error for the sample means would be

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{4}} = \frac{10}{2} = 5$$

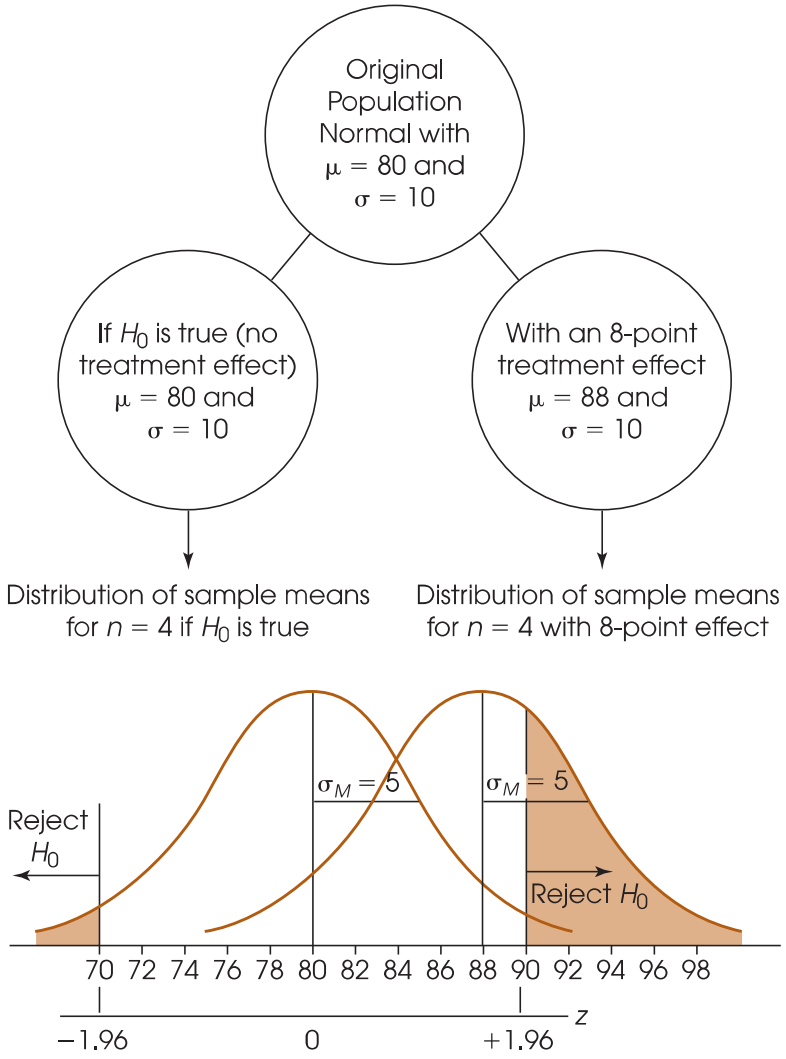
Figure 8.13 shows the two distributions of sample means with $n = 4$ and a standard error of $\sigma_M = 5$ points. Again, the distribution on the left is centered at $\mu = 80$ and shows all of the possible sample means if H_0 is true. As always, this distribution is used to locate the critical boundaries for the hypothesis test, $z = -1.96$ and $z = +1.96$. The distribution on the right is centered at $\mu = 88$ and shows all of the possible sample means if there is an 8-point treatment effect. Note that less than half of the treated sample means in the right-hand distribution are now located beyond the 1.96 boundary. Thus, with a sample of $n = 4$, there is less than a 50% probability that the hypothesis test would reject H_0, even though the treatment has an 8-point effect. Earlier, in Example 8.6, we found power equal to 97.93% for a sample of $n = 25$. However, when the sample size is reduced to $n = 4$, power decreases to less than 50%. In general, a larger sample produces greater power for a hypothesis test.

Because power is directly related to sample size, one of the primary reasons for computing power is to determine what sample size is necessary to achieve a reasonable probability for a successful research study. Before a study is conducted, researchers can compute power to determine the probability that their research will successfully reject the null hypothesis. If the probability (power) is too small, they always have the option of increasing sample size to increase power.

Alpha level Reducing the alpha level for a hypothesis test also reduces the power of the test. For example, lowering α from .05 to .01 lowers the power of the hypothesis

FIGURE 8.13

A demonstration of how sample size affects the power of a hypothesis test. As in Figure 8.12, the left-hand side shows the distribution of sample means if the null hypothesis were true. The critical region is defined for this distribution. The right-hand side shows the distribution of sample means that would be obtained if there were an 8-point treatment effect. Notice that reducing the sample size to $n = 4$ has reduced the power of the test to less than 50% compared to a power of nearly 100% with a sample of $n = 25$ in Figure 8.12.



test. The effect of reducing the alpha level can be seen by referring again to Figure 8.13. In this figure, the boundaries for the critical region are drawn using $\alpha = .05$. Specifically, the critical region on the right-hand side begins at $z = 1.96$. If α were changed to .01, the boundary would be moved farther to the right, out to $z = 2.58$. It should be clear that moving the critical boundary to the right means that a smaller portion of the treatment distribution (the distribution on the right-hand side) will be in the critical region. Thus, there would be a lower probability of rejecting the null hypothesis and a lower value for the power of the test.

One-tailed versus two-tailed tests Changing from a regular two-tailed test to a one-tailed test increases the power of the hypothesis test. Again, this effect can be seen by referring to Figure 8.13. The figure shows the boundaries for the critical region using a two-tailed test with $\alpha = .05$ so that the critical region on the right-hand side begins at $z = 1.96$. Changing to a one-tailed test would move the critical boundary to the left to a value of $z = 1.65$. Moving the boundary to the left would cause a larger proportion of the treatment distribution to be in the critical region and, therefore, would increase the power of the test.

LEARNING CHECK

1. For a particular hypothesis test, the power is .50 (50%) for a 5-point treatment effect. Will the power be greater or less than .50 for a 10-point treatment effect?
2. As the power of a test increases, what happens to the probability of a Type II error?
3. How does increasing sample size influence the power of a hypothesis test?
4. Find the exact value of the power for the hypothesis test shown in Figure 8.13.

ANSWERS

1. The hypothesis test is more likely to detect a 10-point effect, so power will be greater.
2. As power increases, the probability of a Type II error decreases.
3. Increasing sample size increases the power of a test.
4. With $n = 4$, the critical boundary of $z = 1.96$ corresponds to a sample mean of $M = 89.8$, and the exact value for power is $p = 0.3594$ or 35.945%.

SUMMARY

1. Hypothesis testing is an inferential procedure that uses the data from a sample to draw a general conclusion about a population. The procedure begins with a hypothesis about an unknown population. Then a sample is selected, and the sample data provide evidence that either supports or refutes the hypothesis.
2. In this chapter, we introduced hypothesis testing using the simple situation in which a sample mean is used to test a hypothesis about an unknown population mean; usually the mean for a population that has received a treatment. The question is to determine whether the treatment has an effect on the population mean (see Figure 8.2).
3. Hypothesis testing is structured as a four-step process that is used throughout the remainder of the book.
   - **a.** State the null hypothesis (H_0), and select an alpha level. The null hypothesis states that there is no effect or no change. In this case, H_0 states that the mean for the treated population is the same as the mean before treatment. The alpha level, usually $\alpha = .05$ or $\alpha = .01$, provides a definition of the term *very unlikely* and determines the risk of a Type I error. Also state an alternative hypothesis (H_1), which is the exact opposite of the null hypothesis.
   - **b.** Locate the critical region. The critical region is defined as extreme sample outcomes that would be very unlikely to occur if the null hypothesis is true. The alpha level defines "very unlikely."
   - **c.** Collect the data, and compute the test statistic. The sample mean is transformed into a z-score by the formula

$$z = \frac{M - \mu}{\sigma_M}$$

     The value of μ is obtained from the null hypothesis. The z-score test statistic identifies the location of the sample mean in the distribution of sample means.

   - **d.** Make a decision. If the obtained z-score is in the critical region, reject H_0 because it is very unlikely that these data would be obtained if H_0 were true. In this case, conclude that the treatment has changed the population mean. If the z-score is not in the critical region, fail to reject H_0 because the data are not significantly different from the null hypothesis. In this case, the data do not provide sufficient evidence to indicate that the treatment has had an effect.
4. Whatever decision is reached in a hypothesis test, there is always a risk of making the incorrect decision. There are two types of errors that can be committed.

   A Type I error is defined as rejecting a true H_0. This is a serious error because it results in falsely reporting a treatment effect. The risk of a Type I error is determined by the alpha level and, therefore, is under the experimenter's control.

   A Type II error is defined as the failure to reject a false H_0. In this case, the experiment fails to detect an effect

that actually occurred. The probability of a Type II error cannot be specified as a single value and depends in part on the size of the treatment effect. It is identified by the symbol β (beta).

5. When a researcher expects that a treatment will change scores in a particular direction (increase or decrease), it is possible to do a directional, or one-tailed, test. The first step in this procedure is to incorporate the directional prediction into the hypotheses. For example, if the prediction is that a treatment will increase scores, the null hypothesis says that there is no increase and the alternative hypothesis states that there is an increase. To locate the critical region, you must determine what kind of data would refute the null hypothesis by demonstrating that the treatment worked as predicted. These outcomes are located entirely in one tail of the distribution, so the entire critical region (5%, 1%, or 0.1% depending on α) will be in one tail.

6. A one-tailed test is used when there is prior justification for making a directional prediction. These *a priori* reasons may be previous reports and findings or theoretical considerations. In the absence of the *a priori* basis, a two-tailed test is appropriate. In this situation, you might be unsure of what to expect in the study, or you might be testing competing theories.

7. In addition to using a hypothesis test to evaluate the *significance* of a treatment effect, it is recommended that you also measure and report the *effect size*. One measure of effect size is Cohen's *d*, which is a standardized measure of the mean difference. Cohen's *d* is computed as

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}}$$

8. The power of a hypothesis test is defined as the probability that the test will correctly reject the null hypothesis.

9. To determine the power for a hypothesis test, you must first identify the treatment and null distributions. Also, you must specify the magnitude of the treatment effect. Next, you locate the critical region in the null distribution. The power of the hypothesis test is the portion of the treatment distribution that is located beyond the boundary (critical value) of the critical region.

10. As the size of the treatment effect increases, statistical power increases. Also, power is influenced by several factors that can be controlled by the experimenter:

    a. Increasing the alpha level increases power.
    b. A one-tailed test has greater power than a two-tailed test.
    c. A large sample results in more power than a small sample.

KEY TERMS

hypothesis test (233)
null hypothesis (236)
alternative hypothesis (236)
level of significance (237)
alpha level (237)
critical region (238)
test statistic (242)
Type I error (244)
Type II error (245)
beta (246)
significant (251)
directional test (256)
one-tailed test (256)
effect size (262)
Cohen's *d* (262)
power (265)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 8 on the book companion website. The website also provides access to a workshop titled *Hypothesis Testing,* which reviews the concept and logic of hypothesis testing.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics.**

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

The statistical computer package SPSS is not structured to conduct hypothesis tests using z-scores. In truth, the z-score test presented in this chapter is rarely used in actual research situations. The problem with the z-score test is that it requires that you know the value of the population standard deviation, and this information is usually not available. Researchers rarely have detailed information about the populations that they wish to study. Instead, they must obtain information entirely from samples. In the following chapters we introduce new hypothesis-testing techniques that are based entirely on sample data. These new techniques are included in SPSS.

FOCUS ON PROBLEM SOLVING

1. Hypothesis testing involves a set of logical procedures and rules that enable us to make general statements about a population when all we have are sample data. This logic is reflected in the four steps that have been used throughout this chapter. Hypothesis-testing problems are easier to tackle when you learn to follow the steps.

STEP 1 State the hypotheses and set the alpha level.

STEP 2 Locate the critical region.

STEP 3 Compute the test statistic (in this case, the z-score) for the sample.

STEP 4 Make a decision about H_0 based on the result of step 3.

2. Students often ask, "What alpha level should I use?" Or a student may ask, "Why is an alpha of .05 used?" as opposed to something else. There is no single correct answer to either of these questions. Keep in mind that the aim of setting an alpha level in the first place: *to reduce the risk of committing a Type I error*. Therefore,

the maximum acceptable value is $\alpha = .05$. However, some researchers prefer to take even less risk and use alpha levels of .01 or smaller.

Most statistical tests are now done with computer programs that provide an exact probability (p value) for a Type I error. Because an exact value is available, most researchers simply report the p value from the computer printout rather than setting an alpha level at the beginning of the test. However, the same criterion still applies: A result is not significant unless the p value is less than .05.

3. Take time to consider the implications of your decision about the null hypothesis. The null hypothesis states that there is no effect. Therefore, if your decision is to reject H_0, you should conclude that the sample data provide evidence for a treatment effect. However, it is an entirely different matter if your decision is to fail to reject H_0. Remember that when you fail to reject the null hypothesis, the results are inconclusive. It is impossible to *prove* that H_0 is correct; therefore, you cannot state with certainty that "there is no effect" when H_0 is not rejected. At best, all you can state is that "there is insufficient evidence for an effect."

4. It is very important that you understand the structure of the z-score formula (p. 242). It will help you understand many of the other hypothesis tests that are covered later.

5. When you are doing a directional hypothesis test, read the problem carefully, and watch for key words (such as increase or decrease, raise or lower, and more or less) that tell you which direction the researcher is predicting. The predicted direction determines the alternative hypothesis (H_1) and the critical region. For example, if a treatment is expected to *increase* scores, H_1 would contain a *greater than* symbol, and the critical region would be in the tail associated with high scores.

DEMONSTRATION 8.1

HYPOTHESIS TEST WITH *Z*

A researcher begins with a known population—in this case, scores on a standardized test that are normally distributed with $\mu = 65$ and $\sigma = 15$. The researcher suspects that special training in reading skills will produce a change in the scores for the individuals in the population. Because it is not feasible to administer the treatment (the special training) to everyone in the population, a sample of $n = 25$ individuals is selected, and the treatment is given to this sample. Following treatment, the average score for this sample is $M = 70$. Is there evidence that the training has an effect on test scores?

STEP 1 **State the hypothesis and select an alpha level.** The null hypothesis states that the special training has no effect. In symbols,

H_0: $\mu = 65$ (After special training, the mean is still 65.)

The alternative hypothesis states that the treatment does have an effect.

H_1: $\mu \neq 65$ (After training, the mean is different from 65.)

At this time you also select the alpha level. For this demonstration, we will use $\alpha = .05$. Thus, there is a 5% risk of committing a Type I error if we reject H_0.

STEP 2 **Locate the critical region.** With $\alpha = .05$, the critical region consists of sample means that correspond to z-scores beyond the critical boundaries of $z = \pm 1.96$.

STEP 3 **Obtain the sample data, and compute the test statistic.** For this example, the distribution of sample means, according to the null hypothesis, is normal with an expected value of $\mu = 65$ and a standard error of

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

In this distribution, our sample mean of $M = 70$ corresponds to a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{70 - 65}{3} = \frac{5}{3} = +1.67$$

STEP 4 **Make a decision about H_0, and state the conclusion.** The z-score we obtained is not in the critical region. This indicates that our sample mean of $M = 70$ is not an extreme or unusual value to be obtained from a population with $\mu = 65$. Therefore, our statistical decision is to *fail to reject* H_0. Our conclusion for the study is that the data do not provide sufficient evidence that the special training changes test scores.

DEMONSTRATION 8.2

EFFECT SIZE USING COHEN'S *D*

We will compute Cohen's d using the research situation and the data from Demonstration 8.1. Again, the original population mean was $\mu = 65$ and, after treatment (special training), the sample mean was $M = 70$. Thus, there is a 5-point mean difference. Using the population standard deviation, $\sigma = 15$, we obtain an effect size of

$$\text{Cohen's } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{5}{15} = 0.33$$

According to Cohen's evaluation standards (see Table 8.2), this is a medium treatment effect.

PROBLEMS

1. In the z-score formula as it is used in a hypothesis test,
   a. Explain what is measured by $M - \mu$ in the numerator.
   b. Explain what is measured by the standard error in the denominator.
2. The value of the z-score in a hypothesis test is influenced by a variety of factors. Assuming that all other variables are held constant, explain how the value of z is influenced by each of the following:
   a. Increasing the difference between the sample mean and the original population mean.
   b. Increasing the population standard deviation.
   c. Increasing the number of scores in the sample.
3. In words, define the alpha level and the critical region for a hypothesis test.
4. If the alpha level is changed from $\alpha = .05$ to $\alpha = .01$,
   a. What happens to the boundaries for the critical region?
   b. What happens to the probability of a Type I error?
5. Although there is a popular belief that herbal remedies such as ginkgo biloba and ginseng may improve

learning and memory in healthy adults, these effects are usually not supported by well-controlled research (Persson, Bringlov, Nilsson, & Nyberg, 2004). In a typical study, a researcher obtains a sample of $n = 36$ participants and has each person take the herbal supplements every day for 90 days. At the end of the 90 days, each person takes a standardized memory test. For the general population, scores from the test are normally distributed with a mean of $\mu = 80$ and a standard deviation of $\sigma = 18$. The sample of research participants had an average of $M = 84$.

a. Assuming a two-tailed test, state the null hypothesis in a sentence that includes the two variables being examined.
b. Using symbols, state the hypotheses (H_0 and H_1) for the two-tailed test.
c. Sketch the appropriate distribution, and locate the critical region for $\alpha = .05$.
d. Calculate the test statistic (z-score) for the sample.
e. What decision should be made about the null hypothesis, and what decision should be made about the effect of the herbal supplements?

6. Childhood participation in sports, cultural groups, and youth groups appears to be related to improved self-esteem for adolescents (McGee, Williams, Howden-Chapman, Martin, & Kawachi, 2006). In a representative study, a sample of $n = 100$ adolescents with a history of group participation is given a standardized self-esteem questionnaire. For the general population of adolescents, scores on this questionnaire form a normal distribution with a mean of $\mu = 40$ and a standard deviation of $\sigma = 12$. The sample of group-participation adolescents had an average of $M = 43.84$.

a. Does this sample provide enough evidence to conclude that self-esteem scores for these adolescents are significantly different from those of the general population? Use a two-tailed test with $\alpha = .01$.
b. Compute Cohen's d to measure the size of the difference.
c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

7. A local college requires an English composition course for all freshmen. This year they are evaluating a new online version of the course. A random sample of $n = 16$ freshmen is selected and the students are placed in the online course. At the end of the semester, all freshmen take the same English composition exam. The average score for the sample is $M = 76$. For the general population of freshmen who took the traditional lecture class, the exam scores form a normal distribution with a mean of $\mu = 80$.

a. If the final exam scores for the population have a standard deviation of $\sigma = 12$, does the sample provide enough evidence to conclude that the new online course is significantly different from the traditional class? Assume a two-tailed test with $\alpha = .05$.
b. If the population standard deviation is $\sigma = 6$, is the sample sufficient to demonstrate a significant difference? Again, assume a two-tailed test with $\alpha = .05$.
c. Comparing your answers for parts a and b, explain how the magnitude of the standard deviation influences the outcome of a hypothesis test.

8. A random sample is selected from a normal population with a mean of $\mu = 50$ and a standard deviation of $\sigma = 12$. After a treatment is administered to the individuals in the sample, the sample mean is found to be $M = 55$.

a. If the sample consists of $n = 16$ scores, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.
b. If the sample consists of $n = 36$ scores, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.
c. Comparing your answers for parts a and b, explain how the size of the sample influences the outcome of a hypothesis test.

9. A random sample of $n = 36$ scores is selected from a normal population with a mean of $\mu = 60$. After a treatment is administered to the individuals in the sample, the sample mean is found to be $M = 52$.

a. If the population standard deviation is $\sigma = 18$, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.
b. If the population standard deviation is $\sigma = 30$, is the sample mean sufficient to conclude that the treatment has a significant effect? Use a two-tailed test with $\alpha = .05$.
c. Comparing your answers for parts a and b, explain how the magnitude of the standard deviation influences the outcome of a hypothesis test.

10. Miller (2008) examined the energy drink consumption of college undergraduates and found that males use energy drinks significantly more often than females. To further investigate this phenomenon, suppose that a researcher selects a random sample of $n = 36$ male undergraduates and a sample of $n = 25$ females. On average, the males reported consuming $M = 2.45$ drinks per month and females had an average of $M = 1.28$. Assume that the overall level of consumption for college undergraduates averages $\mu = 1.85$ energy drinks per month, and that the distribution of monthly consumption scores is approximately normal with a standard deviation of $\sigma = 1.2$.

a. Did this sample of males consume significantly more energy drinks than the overall population average? Use a one-tailed test with $\alpha = .01$.

b. Did this sample of females consume significantly fewer energy drinks than the overall population average? Use a one-tailed test with $\alpha = .01$

11. A random sample is selected from a normal population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 10$. After a treatment is administered to the individuals in the sample, the sample mean is found to be $M = 42$.

a. How large a sample is necessary for this sample mean to be statistically significant? Assume a two-tailed test with $\alpha = .05$.

b. If the sample mean were $M = 41$, what sample size is needed to be significant for a two-tailed test with $\alpha = .05$?

12. There is some evidence that REM sleep, associated with dreaming, may also play a role in learning and memory processing. For example, Smith and Lapp (1991) found increased REM activity for college students during exam periods. Suppose that REM activity for a sample of $n = 16$ students during the final exam period produced an average score of $M = 143$. Regular REM activity for the college population averages $\mu = 110$ with a standard deviation of $\sigma = 50$. The population distribution is approximately normal.

a. Do the data from this sample provide evidence for a significant increase in REM activity during exams? Use a one-tailed test with $\alpha = .01$.

b. Compute Cohen's *d* to estimate the size of the effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

13. There is some evidence indicating that people with visible tattoos are viewed more negatively than people without visible tattoos (Resenhoeft, Villa, & Wiseman, 2008). In a similar study, a researcher first obtained overall ratings of attractiveness for a woman with no tattoos shown in a color photograph. On a 7-point scale, the woman received an average rating of $\mu = 4.9$, and the distribution of ratings was normal with a standard deviation of $\sigma = 0.84$. The researcher then modified the photo by adding a tattoo of a butterfly on the woman's left arm. The modified photo was then shown to a sample of $n = 16$ students at a local community college and the students used the same 7-point scale to rate the attractiveness of the woman. The average score for the photo with the tattoo was $M = 4.2$.

a. Do the data indicate a significant difference in rated attractiveness when the woman appeared to have a tattoo? Use a two-tailed test with $\alpha = .05$.

b. Compute Cohen's *d* to measure the size of the effect.

c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

14. A psychologist is investigating the hypothesis that children who grow up as the only child in the household develop different personality characteristics than those who grow up in larger families. A sample of $n = 30$ only children is obtained and each child is given a standardized personality test. For the general population, scores on the test from a normal distribution with a mean of $\mu = 50$ and a standard deviation of $\sigma = 15$. If the mean for the sample is $M = 58$, can the researcher conclude that there is a significant difference in personality between only children and the rest of the population? Use a two-tailed test with $\alpha = .05$.

15. A researcher is testing the hypothesis that consuming a sports drink during exercise improves endurance. A sample of $n = 50$ male college students is obtained and each student is given a series of three endurance tasks and asked to consume 4 ounces of the drink during each break between tasks. The overall endurance score for this sample is $M = 53$. For the general population of male college students, without any sports drink, the scores for this task average $\mu = 50$ with a standard deviation of $\sigma = 12$.

a. Can the researcher conclude that endurance scores with the sports drink are significantly higher than scores without the drink? Use a one-tailed test with $\alpha = .05$.

b. Can the researcher conclude that endurance scores with the sports drink are significantly different than scores without the drink? Use a two-tailed test with $\alpha = .05$.

c. You should find that the two tests lead to different conclusions. Explain why.

16. Montarello and Martins (2005) found that fifth-grade students completed more mathematics problems correctly when simple problems were mixed in with their regular math assignments. To further explore this phenomenon, suppose that a researcher selects a standardized mathematics achievement test that produces a normal distribution of scores with a mean of $\mu = 100$ and a standard deviation of $\sigma = 18$. The researcher modifies the test by inserting a set of very easy problems among the standardized questions, and gives the modified test to a sample of $n = 36$ students. If the average test score for the sample is $M = 104$, is this result sufficient to conclude that inserting the easy questions improves student performance? Use a one-tailed test with $\alpha = .01$.

17. Researchers have often noted increases in violent crimes when it is very hot. In fact, Reifman, Larrick, and Fein (1991) noted that this relationship even extends to baseball. That is, there is a much greater chance of a batter being hit by a pitch when the temperature increases. Consider the following hypothetical data. Suppose that over the past 30 years, during any given week of the major-league season, an average of $\mu = 12$ players are hit by wild pitches. Assume that the distribution is nearly normal with $\sigma = 3$. For a sample of $n = 4$ weeks in which the daily temperature was extremely hot, the weekly average of hit-by-pitch players was $M = 15.5$. Are players more likely to get hit by pitches during hot weeks? Set alpha to .05 for a one-tailed test.

18. A researcher plans to conduct an experiment testing the effect of caffeine on reaction time during a driving simulation task. A sample of $n = 9$ participants is selected and each person receives a standard dose of caffeine before being tested on the simulator. The caffeine is expected to lower reaction time by an average of 30 msec. Scores on the simulator task for the regular population (without caffeine) form a normal distribution with $\mu = 240$ msec. and $\sigma = 30$.

a. If the researcher uses a two-tailed test with $\alpha = .05$, what is the power of the hypothesis test?

b. Again assuming a two-tailed test with $\alpha = .05$, what is the power of the hypothesis test if the sample size is increased to $n = 25$?

19. A sample of $n = 40$ is selected from a normal population with $\mu = 75$ msec. and $\sigma = 12$, and a treatment is administered to the sample. The treatment is expected to increase scores by an average of 4 points.

a. If the treatment effect is evaluated with a two-tailed hypothesis test using $\alpha = .05$, what is the power of the test?

b. What is the power of the test if the researcher uses a one-tailed test with $\alpha = .05$?

20. Briefly explain how increasing sample size influences each of the following. Assume that all other factors are held constant.

a. The size of the z-score in a hypothesis test.

b. The size of Cohen's d.

c. The power of a hypothesis test.

21. Explain how the power of a hypothesis test is influenced by each of the following. Assume that all other factors are held constant.

a. Increasing the alpha level from .01 to .05.

b. Changing from a one-tailed test to a two-tailed test.

22. A researcher is investigating the effectiveness of a new medication for lowering blood pressure for individuals with systolic pressure greater than 140. For this population, systolic scores average $\mu = 160$ with a standard deviation of $\sigma = 20$, and the scores form a normal-shaped distribution. The researcher plans to select a sample of $n = 25$ individuals, and measure their systolic blood pressure after they take the medication for 60 days. If the researcher uses a two-tailed test with $\alpha = .05$,

a. What is the power of the test if the medication has a 5-point effect?

b. What is the power of the test if the medication has a 10-point effect?

23. A researcher is evaluating the influence of a treatment using a sample selected from a normally distributed population with a mean of $\mu = 80$ and a standard deviation of $\sigma = 20$. The researcher expects a 12-point treatment effect and plans to use a two-tailed hypothesis test with $\alpha = .05$.

a. Compute the power of the test if the researcher uses a sample of $n = 16$ individuals. (See Example 8.6.)

b. Compute the power of the test if the researcher uses a sample of $n = 25$ individuals.

This page intentionally left blank

PART II

REVIEW

After completing this part, you should understand the basic procedures that form the foundation of inferential statistics. These include:

1. The ability to transform scores into z-scores to describe locations within a distribution and to standardize entire distributions.
2. The ability to determine probabilities associated with individual scores selected from a distribution, especially for scores from normal distributions.
3. The ability to transform sample means into z-scores and to determine the probabilities associated with sample means.
4. The ability to use a sample mean to evaluate a hypothesis about an unknown population mean.

The general goal of inferential statistics is to use the limited information from a sample to answer general questions about an unknown population. In Chapter 8, we introduced hypothesis testing, one of the most commonly used inferential procedures. The hypothesis test presented in Chapter 8 integrates z-scores from Chapter 5, probability from Chapter 6, and the distribution of sample means from Chapter 7 into a single procedure that allows researchers to use a sample from an unknown population to evaluate a hypothesis about the population mean. The researcher first obtains a sample from the unknown population and computes the sample mean. The sample mean and a hypothesized value for the population mean are then used to compute a z-score. If the resulting z-score is a high-probability value, near the center of the distribution of sample means, then the researcher concludes that the sample data fit the hypothesis and the decision is to fail to reject the hypothesis. On the other hand, if the resulting z-score is a low-probability value, out in the tails of the distribution of sample means, then the researcher concludes that the sample data do not fit the hypothesis and the decision is to reject the hypothesis.

REVIEW EXERCISES

1. Find each of the requested values for a population with a mean of $\mu = 40$ and a standard deviation of $\sigma = 8$.
   - **a.** What is the z-score corresponding to $X = 52$?
   - **b.** What is the X value corresponding to $z = -0.50$?
   - **c.** If all of the scores in the population are transformed into z-scores, what will be the values for the mean and standard deviation for the complete set of z-scores?
   - **d.** What is the z-score corresponding to a sample mean of $M = 42$ for a sample of $n = 4$ scores?
   - **e.** What is the z-score corresponding to a sample mean of $M = 42$ for a sample of $n = 16$ scores?

2. A survey of female high school seniors shows that the average amount of time spent on clothes, hair, and makeup each morning before school is $\mu = 35$ minutes. Assume that the distribution of preparation times is approximately normal with a standard deviation of $\sigma = 14$ minutes, and find each of the requested values.
   - **a.** What proportion of female high school seniors spend more than 40 minutes preparing themselves for going to school each morning?
   - **b.** What is the probability of randomly selecting a female high school senior who spends less than 10 minutes on her clothes, hair, and makeup each morning?
   - **c.** What is the probability of obtaining a mean preparation time less than $M = 30$ minutes for a sample of $n = 49$ female high school students?

3. Brunt, Rhee, and Zhong (2008) surveyed 557 undergraduate college students to examine their weight status, health behaviors, and diet. Using body mass index (BMI), they classified the students into four categories: underweight, healthy weight, overweight, and obese. They also measured dietary variety by counting the number of different foods each student ate from several food groups. Note that the researchers are not measuring the amount of food eaten, but rather the number of different foods eaten (variety, not quantity). Nonetheless, it was somewhat surprising that the results showed no differences among the four weight categories that were related to eating fatty and/or sugary snacks.

   Suppose a researcher conducting a follow up study obtains a sample of $n = 25$ students classified as healthy weight and a sample of $n = 36$ students classified as overweight. Each student completes the food variety questionnaire, and the healthy-weight group produces a mean of $M = 4.01$ for the fatty, sugary snack category compared to a mean of $M = 4.48$ for the overweight group. The results from the Brunt, Rhee, and Zhong study showed an overall mean variety score of $\mu = 4.22$ for the discretionary sweets or fats food group. Assume that the distribution of scores is approximately normal with a standard deviation of $\sigma = 0.60$.
   - **a.** Does the sample of $n = 36$ indicate that number of fatty, sugary snacks eaten by overweight students is significantly different from the overall population mean? Use a two-tailed test with $\alpha = .05$.
   - **b.** Based on the sample of $n = 25$ healthy-weight students, can you conclude that healthy-weight students eat significantly fewer fatty, sugary snacks than the overall population? Use a one-tailed test with $\alpha = .05$.

This page intentionally left blank