CHAPTER 8

Introduction to Hypothesis Testing

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

1. Identify the four steps of hypothesis testing.
2. Define null hypothesis, alternative hypothesis, level of significance, test statistic, *p* value, and statistical significance.
3. Define Type I error and Type II error, and identify the type of error that researchers control.
4. Calculate the one-independent sample *z* test and interpret the results.
5. Distinguish between a one-tailed and two-tailed test, and explain why a Type III error is possible only with one-tailed tests.
6. Explain what effect size measures and compute a Cohen's *d* for the one-independent sample *z* test.
7. Define power and identify six factors that influence power.
8. Summarize the results of a one-independent sample *z* test in American Psychological Association (APA) format.

8.1 INFERENTIAL STATISTICS AND HYPOTHESIS TESTING

We use inferential statistics because it allows us to measure behavior in samples to learn more about the behavior in populations that are often too large or inaccessible. We use samples because we know how they are related to populations. For example, suppose the average score on a standardized exam in a given population is 1,000. In Chapter 7, we showed that the sample mean as an unbiased estimator of the population mean—if we selected a random sample from a population, then on average the value of the sample mean will equal the population mean. In our example, if we select a random sample from this population with a mean of 1,000, then on average, the value of a sample mean will equal 1,000. On the basis of the central limit theorem, we know that the probability of selecting any other sample mean value from this population is normally distributed.

In behavioral research, we select samples to learn more about populations of interest to us. In terms of the mean, we measure a sample mean to learn more about the mean in a population. Therefore, we will use the sample mean to describe the population mean. We begin by stating the value of a population mean, and then we select a sample and measure the mean in that sample. On average, the value of the sample mean will equal the population mean. The larger the difference or discrepancy between the sample mean and population mean, the less likely it is that we could have selected that sample mean, if the value of the population mean is correct. This type of experimental situation, using the example of standardized exam scores, is illustrated in Figure 8.1.

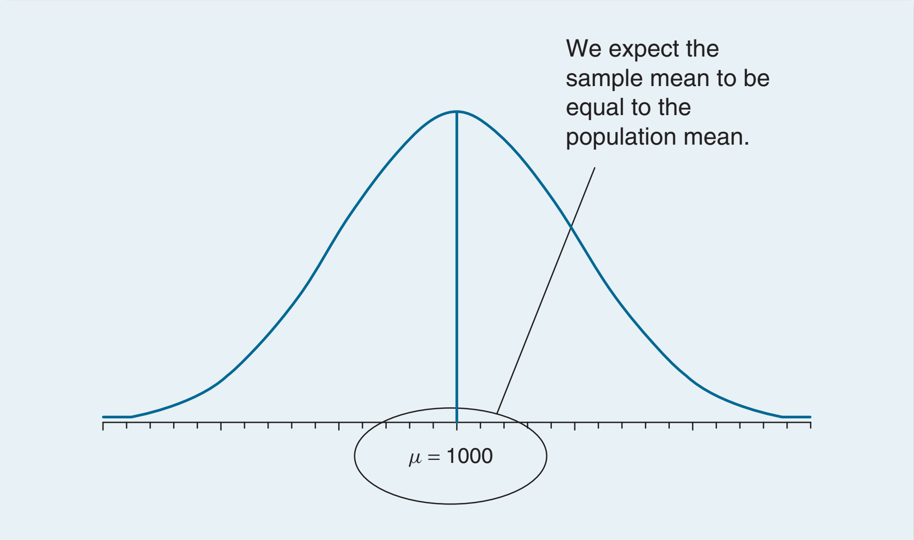


FIGURE 8.1

The sampling distribution for a population mean is equal to 1,000. If 1,000 is the correct population mean, then we know that, on average, the sample mean will equal 1,000 (the population mean). Using the empirical rule, we know that about 95% of all samples selected from this population will have a sample mean that falls within two standard deviations (*SD*) of the mean. It is therefore unlikely (less than a 5% probability) that we will measure a sample mean beyond 2 *SD* from the population mean, if the population mean is indeed correct.

The method in which we select samples to learn more about characteristics in a given population is called **hypothesis testing**. Hypothesis testing is really a systematic way to test claims or ideas about a group or population. To illustrate, suppose we read an article stating that children in the United States watch an average of 3 hours of TV per week. To test whether this claim is true, we record the time (in hours) that a group of 20 American children (the sample), among all children in the United States (the population), watch TV. The mean we measure for these 20 children is a sample mean. We can then compare the sample mean we select to the population mean stated in the article.

DEFINITION

Hypothesis testing or **significance testing** is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

The method of hypothesis testing can be summarized in four steps. We will describe each of these four steps in greater detail in Section 8.2.

1. To begin, we identify a hypothesis or claim that we feel should be tested. For example, we might want to test the claim that the mean number of hours that children in the United States watch TV is 3 hours.
2. We select a criterion upon which we decide that the claim being tested is true or not. For example, the claim is that children watch 3 hours of TV per week. Most samples we select should have a mean close to or equal to 3 hours if the claim we are testing is true. So at what point do we decide that the discrepancy between the sample mean and 3 is so big that the claim we are testing is likely not true? We answer this question in this step of hypothesis testing.
3. Select a random sample from the population and measure the sample mean. For example, we could select 20 children and measure the mean time (in hours) that they watch TV per week.
4. Compare what we observe in the sample to what we expect to observe if the claim we are testing is true. We expect the sample mean to be around 3 hours. If the discrepancy between the sample mean and population mean is small, then we will likely decide that the claim we are testing is indeed true. If the discrepancy is too large, then we will likely decide to reject the claim as being not true.

NOTE: Hypothesis testing is the method of testing whether claims or hypotheses regarding a population are likely to be true.

LEARNING CHECK 1

1. On average, what do we expect the sample mean to be equal to?
2. True or false: Researchers select a sample from a population to learn more about characteristics in that sample.

Answers: 1. The population mean; 2. False. Researchers select a sample from a population to learn more about characteristics in the population that the sample was selected from.

8.2 FOUR STEPS TO HYPOTHESIS TESTING

The goal of hypothesis testing is to determine the likelihood that a population parameter, such as the mean, is likely to be true. In this section, we describe the four steps of hypothesis testing that were briefly introduced in Section 8.1:

Step 1: State the hypotheses.

Step 2: Set the criteria for a decision.

Step 3: Compute the test statistic.

Step 4: Make a decision.

Step 1: State the hypotheses. We begin by stating the value of a population mean in a **null hypothesis,** which we presume is true. For the children watching TV example, we state the null hypothesis that children in the United States watch an average of 3 hours of TV per week. This is a starting point so that we can decide whether this is likely to be true, similar to the presumption of innocence in a courtroom. When a defendant is on trial, the jury starts by assuming that the defendant is innocent. The basis of the decision is to determine whether this assumption is true. Likewise, in hypothesis testing, we start by assuming that the hypothesis or claim we are testing is true. This is stated in the null hypothesis. The basis of the decision is to determine whether this assumption is likely to be true.

DEFINITION

The **null hypothesis (H_0)**, stated as the **null**, is a statement about a population parameter, such as the population mean, that is assumed to be true.

The null hypothesis is a starting point. We will test whether the value stated in the null hypothesis is likely to be true.

Keep in mind that the only reason we are testing the null hypothesis is because we think it is wrong. We state what we think is wrong about the null hypothesis in an **alternative hypothesis.** For the children watching TV example, we may have reason to believe that children watch more than (>) or less than (<) 3 hours of TV per week. When we are uncertain of the direction, we can state that the value in the null hypothesis is not equal to (≠) 3 hours.

NOTE: In hypothesis testing, we conduct a study to test whether the null hypothesis is likely to be true.

In a courtroom, since the defendant is assumed to be innocent (this is the null hypothesis so to speak), the burden is on a prosecutor to conduct a trial to show evidence that the defendant is not innocent. In a similar way, we assume the null hypothesis is true, placing the burden on the researcher to conduct a study to show evidence that the null hypothesis is unlikely to be true. Regardless, we always make a decision about the null hypothesis (that it is likely or unlikely to be true). The alternative hypothesis is needed for Step 2.

DEFINITION

An **alternative hypothesis (H_1)** is a statement that directly contradicts a null hypothesis by stating that that the actual value of a population parameter is less than, greater than, or not equal to the value stated in the null hypothesis.

The alternative hypothesis states what we think is wrong about the null hypothesis, which is needed for Step 2.

MAKING SENSE: Testing the Null Hypothesis

A decision made in hypothesis testing centers on the null hypothesis. This means two things in terms of making a decision:

1. Decisions are made about the null hypothesis. Using the courtroom analogy, a jury decides whether a defendant is guilty or not guilty. The jury does not make a decision of guilty or *innocent* because the defendant is assumed to be innocent. All evidence presented in a trial is to show that a defendant is guilty. The evidence either shows guilt (decision: guilty) or does not (decision: not guilty). In a similar way, the null hypothesis is assumed to be correct. A researcher conducts a study showing evidence that this assumption is unlikely (we reject the null hypothesis) or fails to do so (we retain the null hypothesis).
2. The bias is to do nothing. Using the courtroom analogy, for the same reason the courts would rather let the guilty go free than send the innocent to prison, researchers would rather do nothing (accept previous notions of truth stated by a null hypothesis) than make statements that are not correct. For this reason, we assume the null hypothesis is correct, thereby placing the burden on the researcher to demonstrate that the null hypothesis is not likely to be correct.

Step 2: Set the criteria for a decision. To set the criteria for a decision, we state the **level of significance** for a test. This is similar to the criterion that jurors use in a criminal trial. Jurors decide whether the evidence presented shows guilt *beyond a reasonable doubt* (this is the criterion). Likewise, in hypothesis testing, we collect data to show that the null hypothesis is not true, based on the likelihood of selecting a sample mean from a population (the likelihood is the criterion). The likelihood or level of significance is typically set at 5% in behavioral research studies. When the probability of obtaining a sample mean is less than 5% if the null hypothesis were true, then we conclude that the sample we selected is too unlikely and so we reject the null hypothesis.

DEFINITION

Level of significance, or significance level, refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value stated in the null hypothesis were true.

In behavioral science, the criterion or level of significance is typically set at 5%. When the probability of obtaining a sample mean is less than 5% if the null hypothesis were true, then we reject the value stated in the null hypothesis.

The alternative hypothesis establishes where to place the level of significance. Remember that we know that the sample mean will equal the population mean on average if the null hypothesis is true. All other possible values of the sample mean are normally distributed (central limit theorem). The empirical rule tells us that at least 95% of all sample means fall within about 2 standard deviations (*SD*) of the population mean, meaning that there is less than a 5% probability of obtaining a

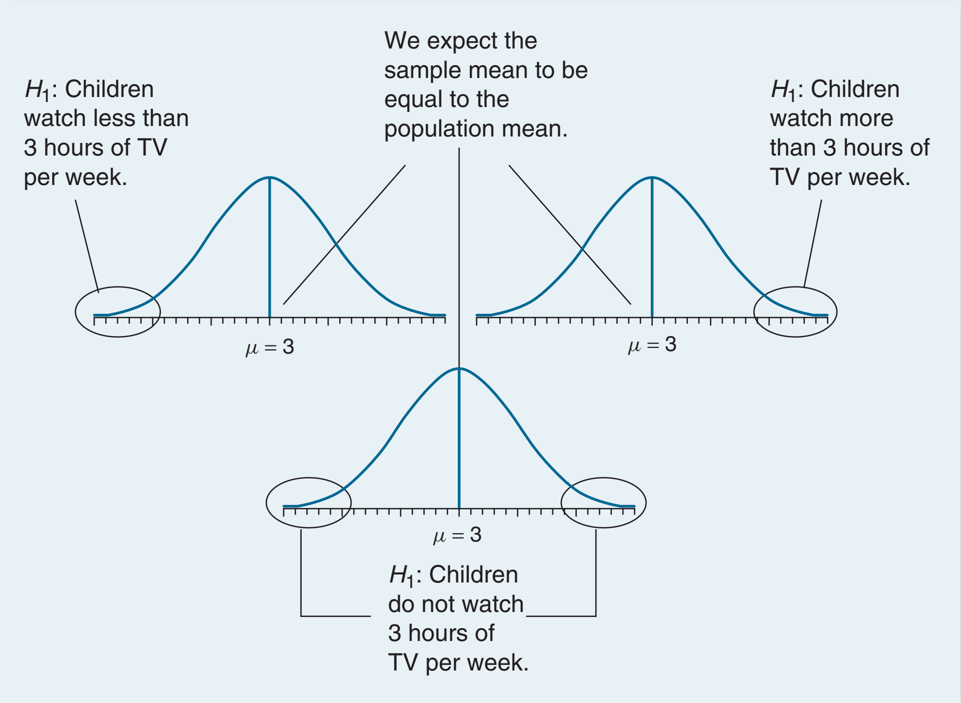

H_1: Children watch less than 3 hours of TV per week.
We expect the sample mean to be equal to the population mean.
H_1: Children watch more than 3 hours of TV per week.
$\mu = 3$
$\mu = 3$
$\mu = 3$
H_1: Children do not watch 3 hours of TV per week.

FIGURE 8.2

The alternative hypothesis determines whether to place the level of significance in one or both tails of a sampling distribution. Sample means that fall in the tails are unlikely to occur (less than a 5% probability) if the value stated for a population mean in the null hypothesis is true.

sample mean that is beyond 2 *SD* from the population mean. For the children watching TV example, we can look for the probability of obtaining a sample mean beyond 2 *SD* in the upper tail (greater than 3), the lower tail (less than 3), or both tails (not equal to 3). Figure 8.2 shows that the alternative hypothesis is used to determine which tail or tails to place the level of significance for a hypothesis test.

NOTE: The level of significance in hypothesis testing is the criterion we use to decide whether the value stated in the null hypothesis is likely to be true.

Step 3: Compute the test statistic. Suppose we measure a sample mean equal to 4 hours per week that children watch TV. To make a decision, we need to evaluate how likely this sample outcome is, if the population mean stated by the null hypothesis (3 hours per week) is true. We use a **test statistic** to determine this likelihood**.** Specifically, a test statistic tells us how far, or how many standard deviations, a sample mean is from the population mean. The larger the value of the test statistic, the further the distance, or number of standard deviations, a sample mean is from the population mean stated in the null hypothesis. The value of the test statistic is used to make a decision in Step 4.

DEFINITION

The **test statistic** is a mathematical formula that allows researchers to determine the likelihood of obtaining sample outcomes if the null hypothesis were true. The value of the test statistic is used to make a decision regarding the null hypothesis.

NOTE: We use the value of the test statistic to make a decision regarding the null hypothesis.

Step 4: Make a decision. We use the value of the test statistic to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true. If the probability of obtaining a sample mean is less than 5% when the null hypothesis is true, then the decision is to reject the null hypothesis. If the probability of obtaining a sample mean is greater than 5% when the null hypothesis is true, then the decision is to retain the null hypothesis. In sum, there are two decisions a researcher can make:

1. Reject the null hypothesis. The sample mean is associated with a low probability of occurrence when the null hypothesis is true.
2. Retain the null hypothesis. The sample mean is associated with a high probability of occurrence when the null hypothesis is true.

The probability of obtaining a sample mean, given that the value stated in the null hypothesis is true, is stated by the ***p* value.** The p value is a probability: It varies between 0 and 1 and can never be negative. In Step 2, we stated the criterion or probability of obtaining a sample mean at which point we will decide to reject the value stated in the null hypothesis, which is typically set at 5% in behavioral research. To make a decision, we compare the p value to the criterion we set in Step 2.

DEFINITION

A ***p* value** is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. The p value for obtaining a sample outcome is compared to the level of significance.

Significance, or **statistical significance**, describes a decision made concerning a value stated in the null hypothesis. When the null hypothesis is rejected, we reach significance. When the null hypothesis is retained, we fail to reach significance.

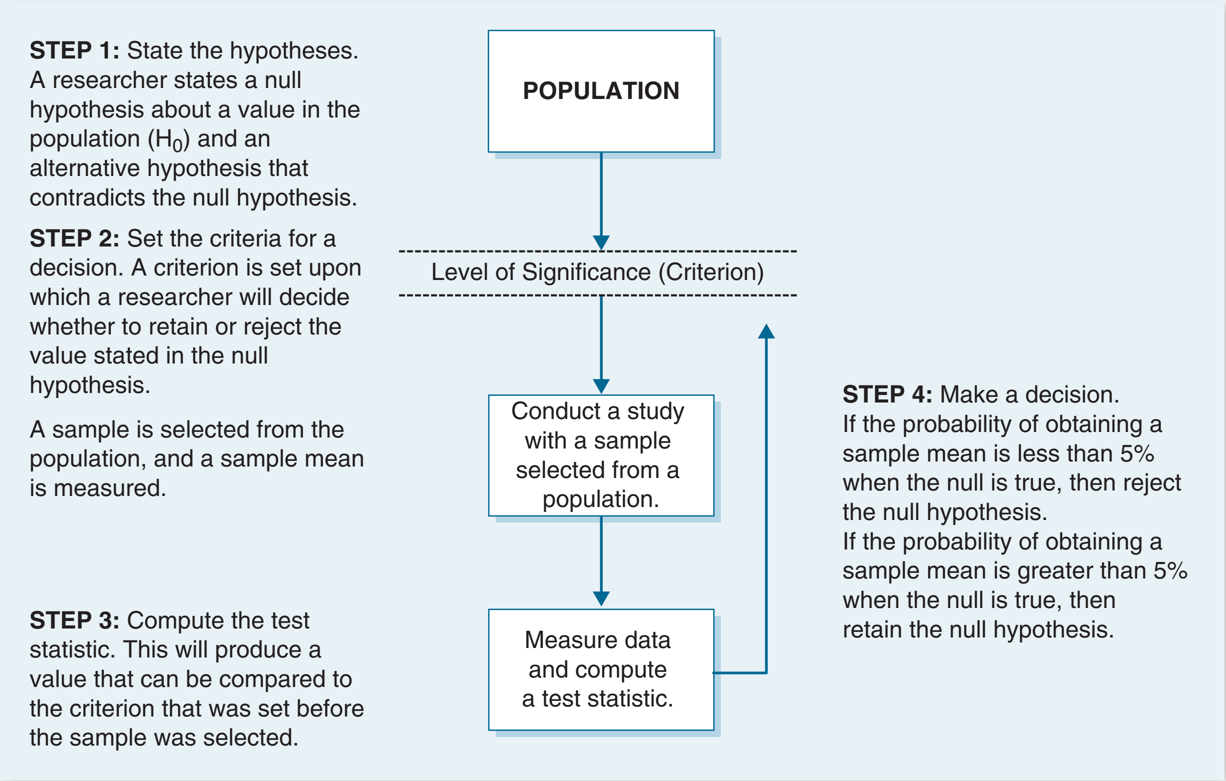
When the p value is less than 5% ($p < .05$), we reject the null hypothesis. We will refer to $p < .05$ as the criterion for deciding to reject the null hypothesis, although note that when $p = .05$, the decision is also to reject the null hypothesis. When the p value is greater than 5% ($p > .05$), we retain the null hypothesis. The decision to reject or retain the null hypothesis is called **significance.** When the p value is less than .05, we reach significance; the decision is to reject the null hypothesis. When the p value is greater than .05, we fail to reach significance; the decision is to retain the null hypothesis. Figure 8.3 shows the four steps of hypothesis testing.

NOTE: Researchers make decisions regarding the null hypothesis. The decision can be to retain the null ($p > .05$) or reject the null ($p < .05$).

LEARNING CHECK 2

1. State the four steps of hypothesis testing.
2. The decision in hypothesis testing is to retain or reject which hypothesis: the null or alternative hypothesis?
3. The criterion or level of significance in behavioral research is typically set at what probability value?
4. A test statistic is associated with a p value less than .05 or 5%. What is the decision for this hypothesis test?
5. If the null hypothesis is rejected, then did we reach significance?

Answers: 1. Step 1: State the null and alternative hypothesis. Step 2: Determine the level of significance. Step 3: Compute the test statistic. Step 4: Make a decision; 2. Null; 3. A .05 or 5% likelihood for obtaining a sample outcome; 4. Reject the null; 5. Yes.



STEP 1: State the hypotheses. A researcher states a null hypothesis about a value in the population (H_0) and an alternative hypothesis that contradicts the null hypothesis.

STEP 2: Set the criteria for a decision. A criterion is set upon which a researcher will decide whether to retain or reject the value stated in the null hypothesis.

A sample is selected from the population, and a sample mean is measured.

STEP 3: Compute the test statistic. This will produce a value that can be compared to the criterion that was set before the sample was selected.

STEP 4: Make a decision. If the probability of obtaining a sample mean is less than 5% when the null is true, then reject the null hypothesis. If the probability of obtaining a sample mean is greater than 5% when the null is true, then retain the null hypothesis.

POPULATION
Level of Significance (Criterion)
Conduct a study with a sample selected from a population.
Measure data and compute a test statistic.

FIGURE 8.3

A summary of hypothesis testing.

8.3 HYPOTHESIS TESTING AND SAMPLING DISTRIBUTIONS

The logic of hypothesis testing is rooted in an understanding of the sampling distribution of the mean. In Chapter 7, we showed three characteristics of the mean, two of which are particularly relevant in this section:

1. The sample mean is an unbiased estimator of the population mean. On average, a randomly selected sample will have a mean equal to that in the population. In hypothesis testing, we begin by stating the null hypothesis. We expect that, if the null hypothesis is true, then a random sample selected from a given population will have a sample mean equal to the value stated in the null hypothesis.
2. Regardless of the distribution in the population, the sampling distribution of the sample mean is normally distributed. Hence, the probabilities of all other possible sample means we could select are normally distributed. Using this distribution, we can therefore state an alternative hypothesis to locate the probability of obtaining sample means with less than a 5% chance of being selected if the value stated in the null hypothesis is true. Figure 8.2 shows that we can identify sample mean outcomes in one or both tails.

To locate the probability of obtaining a sample mean in a sampling distribution, we must know (1) the population mean and (2) the standard error of the mean (*SEM;* introduced in Chapter 7). Each value is entered in the test statistic formula computed in Step 3, thereby allowing us to make a decision in Step 4. To review, Table 8.1 displays the notations used to describe populations, samples, and sampling distributions. Table 8.2 summarizes the characteristics of each type of distribution.

TABLE 8.1 A review of the notation used for the mean, variance, and standard deviation in population, sample, and sampling distributions.

Characteristic	Population	Sample	Sampling Distribution
Mean	μ	M or $\bar{X}$	$\mu_M = \mu$
Variance	σ^2	s^2 or SD^2	$\sigma_M^2 = \frac{\sigma^2}{n}$
Standard deviation	σ	s or SD	$\sigma_M = \frac{\sigma}{\sqrt{n}}$

TABLE 8.2 A review of the key differences between population, sample, and sampling distributions.

	Population Distribution	Sample Distribution	Distribution of Sample Means
What is it?	Scores of all persons in a population	Scores of a select portion of persons from the population	All possible sample means that can be drawn, given a certain sample size
Is it accessible?	Typically, no	Yes	Yes
What is the shape?	Could be any shape	Could be any shape	Normally distributed

LEARNING CHECK 3

1. For the following statement, write increases or decreases as an answer. The likelihood that we reject the null hypothesis (increases or decreases):
   a. The closer the value of a sample mean is to the value stated by the null hypothesis?
   b. The further the value of a sample mean is from the value stated in the null hypothesis?
2. A researcher selects a sample of 49 students to test the null hypothesis that the average student exercises 90 minutes per week. What is the mean for the sampling distribution for this population of interest if the null hypothesis is true?

Answers: 1. (a) Decreases, (b) Increases; 2. 90 minutes.

8.4 MAKING A DECISION: TYPES OF ERROR

In Step 4, we decide whether to retain or reject the null hypothesis. Because we are observing a sample and not an entire population, it is possible that a conclusion may be wrong. Table 8.3 shows that there are four decision alternatives regarding the truth and falsity of the decision we make about a null hypothesis:

1. The decision to retain the null hypothesis could be correct.
2. The decision to retain the null hypothesis could be incorrect.
3. The decision to reject the null hypothesis could be correct.
4. The decision to reject the null hypothesis could be incorrect.

TABLE 8.3 Four outcomes for making a decision. The decision can be either correct (correctly reject or retain null) or wrong (incorrectly reject or retain null).

		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT $1-\alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1-\beta$ POWER

We investigate each decision alternative in this section. Since we will observe a sample, and not a population, it is impossible to know for sure the truth in the population. So for the sake of illustration, we will assume we know this. This assumption is labeled as truth in the population in Table 8.3. In this section, we will introduce each decision alternative.

DECISION: RETAIN THE NULL HYPOTHESIS

When we decide to retain the null hypothesis, we can be correct or incorrect. The correct decision is to retain a true null hypothesis. This decision is called a null result or null finding. This is usually an uninteresting decision because the decision is to retain what we already assumed: that the value stated in the null hypothesis is correct. For this reason, null results alone are rarely published in behavioral research.

The incorrect decision is to retain a false null hypothesis. This decision is an example of a **Type II error,** or **β error.** With each test we make, there is always some probability that the decision could be a Type II error. In this decision, we decide to retain previous notions of truth that are in fact false. While it's an error, we still did nothing; we retained the null hypothesis. We can always go back and conduct more studies.

DEFINITION

Type II error, or **beta (β) error**, is the probability of retaining a null hypothesis that is actually false.

DECISION: REJECT THE NULL HYPOTHESIS

NOTE: A Type II error, or beta (β) error, is the probability of incorrectly retaining the null hypothesis.

When we decide to reject the null hypothesis, we can be correct or incorrect. The incorrect decision is to reject a true null hypothesis. This decision is an example of a **Type I error.** With each test we make, there is always some probability that our decision is a Type I error. A researcher who makes this error decides to reject previous notions of truth that are in fact true. Making this type of error is analogous to finding an innocent person guilty. To minimize this error, we assume a defendant is innocent when beginning a trial. Similarly, to minimize making a Type I error, we assume the null hypothesis is true when beginning a hypothesis test.

DEFINITION

Type I error is the probability of rejecting a null hypothesis that is actually true. Researchers directly control for the probability of committing this type of error.

An **alpha (α) level** is the level of significance or criterion for a hypothesis test. It is the largest probability of committing a Type I error that we will allow and still decide to reject the null hypothesis.

Since we assume the null hypothesis is true, we control for Type I error by stating a level of significance. The level we set, called the **alpha level** (symbolized as α), is the largest probability of committing a Type I error that we will allow and still decide to reject the null hypothesis. This criterion is usually set at .05 ($\alpha = .05$), and we compare the alpha level to the p value. When the probability of a Type I error is less than 5% ($p < .05$), we decide to reject the null hypothesis; otherwise, we retain the null hypothesis.

NOTE: Researchers directly control for the probability of a Type I error by stating an alpha (α) level.

The correct decision is to reject a false null hypothesis. There is always some probability that we decide that the null hypothesis is false when it is indeed false. This decision is called the **power** of the decision-making process. It is called power because it is the decision we aim for. Remember that we are only testing the null hypothesis because we think it is wrong. Deciding to reject a false null hypothesis, then, is the power, inasmuch as we learn the most about populations when we accurately reject false notions of truth. This decision is the most published result in behavioral research.

NOTE: The power in hypothesis testing is the probability of correctly rejecting the value stated in the null hypothesis.

DEFINITION

The **power** in hypothesis testing is the probability of rejecting a false null hypothesis. Specifically, it is the probability that a randomly selected sample will show that the null hypothesis is false when the null hypothesis is indeed false.

LEARNING CHECK 4

1. What type of error do we directly control?
2. What type of error is associated with decisions to retain the null?
3. What type of error is associated with decisions to reject the null?
4. State the two correct decisions that a researcher can make.

Answers: 1. Type I error; 2. Type II error; 3. Type I error; 4. Retain a true null hypothesis and reject a false null hypothesis.

8.5 TESTING A RESEARCH HYPOTHESIS: EXAMPLES USING THE *Z* TEST

The test statistic in Step 3 converts the sampling distribution we observe into a standard normal distribution, thereby allowing us to make a decision in Step 4. The test statistic we use depends largely on what we know about the population. When we know the mean and standard deviation in a single population, we can use the **one–independent sample *z* test**, which we will use in this section to illustrate the four steps of hypothesis testing.

DEFINITION

The **one–independent sample *z* test** is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance.

NOTE: The z test is used to test hypotheses about a population mean when the population variance is known.

Recall that we can state one of three alternative hypotheses: A population mean is greater than (>), less than (<), or not equal (≠) to the value stated in a null hypothesis. The alternative hypothesis determines which tail of a sampling distribution to place the level of significance, as illustrated in Figure 8.2. In this section, we will use an example for each type of alternative hypothesis.

NONDIRECTIONAL, TWO-TAILED HYPOTHESIS TESTS (H1: ≠)

NOTE: Nondirectional tests are used to test hypotheses when we are interested in any alternative from the null hypothesis.

In Example 8.1, we will use the *z* test for a **nondirectional,** or **two-tailed test,** where the alternative hypothesis is stated as *not equal to* (≠) the null hypothesis. For this test, we will place the level of significance in both tails of the sampling distribution. We are therefore interested in any alternative from the null hypothesis. This is the most common alternative hypothesis tested in behavioral science.

DEFINITION

Nondirectional tests, or **two-tailed tests**, are hypothesis tests where the alternative hypothesis is stated as *not equal to* (≠). The researcher is interested in any alternative from the null hypothesis.

EXAMPLE 8.1

Templer and Tomeo (2002) reported that the population mean score on the quantitative portion of the Graduate Record Examination (GRE) General Test for students taking the exam between 1994 and 1997 was 558 ± 139 ($\mu \pm \sigma$). Suppose we select a sample of 100 participants ($n = 100$). We record a sample mean equal to 585 ($M = 585$). Compute the one–independent sample *z* test for whether or not we will retain the null hypothesis ($\mu = 558$) at a .05 level of significance ($\alpha = .05$).

Step 1: State the hypotheses. The population mean is 558, and we are testing whether the null hypothesis is (=) or is not (≠) correct:

H_0: $\mu = 558$ Mean test scores are equal to 558 in the population.

H_1: $\mu \neq 558$ Mean test scores are not equal to 558 in the population.

Step 2: Set the criteria for a decision. The level of significance is .05, which makes the alpha level $\alpha = .05$. To locate the probability of obtaining a sample mean from a given population, we use the standard normal distribution. We will locate the *z* scores in a standard normal distribution that are the cutoffs, or **critical values,** for sample mean values with less than a 5% probability of occurrence if the value stated in the null ($\mu = 558$) is true.

DEFINITION

> A **critical value** is a cutoff value that defines the boundaries beyond which less than 5% of sample means can be obtained if the null hypothesis is true. Sample means obtained beyond a critical value will result in a decision to reject the null hypothesis.

In a nondirectional two-tailed test, we divide the alpha value in half so that an equal proportion of area is placed in the upper and lower tail. Table 8.4 gives the critical values for one- and two-tailed tests at a .05, .01, and .001 level of significance. Figure 8.4 displays a graph with the critical values for Example 8.1 shown. In this example $\alpha = .05$, so we split this probability in half:

$$\text{Splitting } \alpha \text{ in half: } \frac{\alpha}{2} = \frac{.05}{2} = .0250 \text{ in each tail}$$

TABLE 8.4 Critical values for one- and two-tailed tests at three commonly used levels of significance.

Level of Significance (α)	Type of Test	
	One-Tailed	Two-Tailed
0.05	+1.645 or –1.645	±1.96
0.01	+2.33 or –2.33	±2.58
0.001	+3.09 or –3.09	±3.30

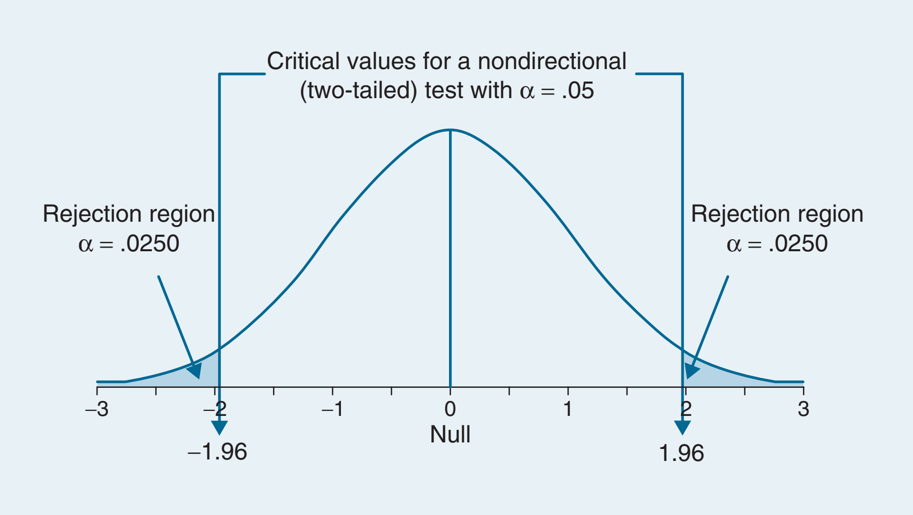


FIGURE 8.4

The critical values (±1.96) for a nondirectional (two-tailed) test with a .05 level of significance.

NOTE: For two-tailed tests, the alpha is split in half and placed in each tail of a standard normal distribution.

NOTE: A critical value marks the cutoff for the rejection region.

To locate the critical values, we use the unit normal table given in Table B1 in Appendix B and look up the proportion .0250 toward the tail in column C. This value, .0250, is listed for a *z*-score equal to $z = 1.96$. This is the critical value for the upper tail of the standard normal distribution. Since the normal distribution is symmetrical, the critical value in the bottom tail will be the same distance below the mean, or $z = –1.96$. The regions beyond the critical values, displayed in Figure 8.4, are called the **rejection regions.** If the value of the test statistic falls in these regions, then the decision is to reject the null hypothesis; otherwise, we retain the null hypothesis.

DEFINITION

The **rejection region** is the region beyond a critical value in a hypothesis test. When the value of a test statistic is in the rejection region, we decide to reject the null hypothesis; otherwise, we retain the null hypothesis.

Step 3: Compute the test statistic. Step 2 sets the stage for making a decision because the criterion is set. The probability is less than 5% that we will obtain a sample mean that is at least 1.96 standard deviations above or below the value of the population mean stated in the null hypothesis. In this step, we will compute a test statistic to determine whether the sample mean we selected is beyond or within the critical values we stated in Step 2.

The test statistic for a one–independent sample *z* test is called the ***z* statistic.** The *z* statistic converts any sampling distribution into a standard normal distribution. The *z* statistic is therefore a *z* transformation. The solution of the formula gives the number of standard deviations, or *z*-scores, that a sample mean falls above or below the population mean stated in the null hypothesis. We can then compare the value of the *z* statistic, called the **obtained value,** to the critical values we determined in Step 2. The *z* statistic formula is the sample mean minus the population mean stated in the null hypothesis, divided by the standard error of the mean:

$$\textbf{\textit{z} statistic: } z_{obt} = \frac{M - \mu}{\sigma_M}, \text{ where } \sigma_M = \frac{\sigma}{\sqrt{n}}.$$

DEFINITION

The ***z* statistic** is an inferential statistic used to determine the number of standard deviations in a standard normal distribution that a sample mean deviates from the population mean stated in the null hypothesis.

The **obtained value** is the value of a test statistic. This value is compared to the critical value(s) of a hypothesis test to make a decision. When the obtained value exceeds a critical value, we decide to reject the null hypothesis; otherwise, we retain the null hypothesis.

To calculate the *z* statistic, first compute the standard error (σ_M), which is the denominator for the *z* statistic:

NOTE: The z statistic measures the number of standard deviations, or z-scores, that a sample mean falls above or below the population mean stated in the null hypothesis.

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{139}{\sqrt{100}} = 13.9.$$

Then compute the *z* statistic by substituting the values of the sample mean, $M = 585$; the population mean stated by the null hypothesis, $\mu = 558$; and the standard error we just calculated, $\sigma_M = 13.9$:

$$z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{585 - 558}{13.9} = 1.94.$$

Step 4: Make a decision. To make a decision, we compare the obtained value to the critical values. We reject the null hypothesis if the obtained value exceeds a critical value. Figure 8.5 shows that the obtained value ($Z_{obt} = 1.94$) is less than the critical value; it does not fall in the rejection region. The decision is to retain the null hypothesis.

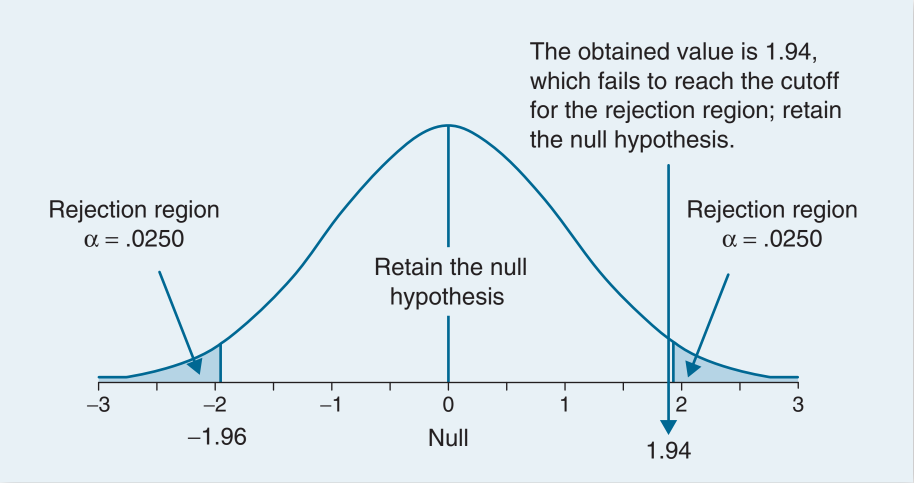


FIGURE 8.5

Since the obtained value fails to reach the rejection region (it is within the critical values of ±1.96), we decide to retain the null hypothesis.

The probability of obtaining $Z_{obt} = 1.94$ is stated by the p value. To locate the p value or probability of obtaining the z statistic, we refer to the unit normal table in Table B1 in Appendix B. Look for a z score equal to 1.94 in column A, then locate the probability toward the tail in column C. The value is .0262. Finally, multiply the value given in column C times the number of tails for alpha. Since this is a two-tailed test, we multiply .0262 times 2: $p = (.0262) \times 2$ tails = .0524. Table 8.5 summarizes how to determine the p value for one- and two-tailed tests. (We will compute one-tailed tests in Examples 8.2 and 8.3.)

TABLE 8.5 To find the p value for the z statistic, find its probability (toward the tail) in the unit normal table and multiply this probability times the number of tails for alpha.

	One-Tailed Test	Two-Tailed Test
Number of tails	1	2
Probability	p	p
p value calculation	$1p$	$2p$

We found in Example 8.1 that if the null hypothesis were true, then $p = .0524$ that we could have selected this sample mean from this population. The criteria we set in Step 2 was that the probability must be less than 5% that we obtain a sample mean, if the null hypothesis were true. Since p is greater than 5%, we decide to retain the null hypothesis. We conclude that the mean score on the GRE General Test in this population is 558 (the value stated in the null hypothesis).

DIRECTIONAL, UPPER-TAIL CRITICAL HYPOTHESIS TESTS (H1: >)

NOTE: An upper-tail critical test is conducted when it is not possible or highly unlikely that a sample mean will fall below the population mean stated in the null hypothesis.

In Example 8.2, we will use the *z* test for a **directional,** or **one-tailed test,** where the alternative hypothesis is stated as *greater than* (>) the null hypothesis. A directional test can also be stated as *less than* (<) the null hypothesis (an example for this alternative is given in Example 8.3). For an upper-tail critical test, or a greater than statement, we place the level of significance in the upper tail of the sampling distribution. So we are interested in any alternative greater than the value stated in the null hypothesis. This test is appropriate when it is not possible or highly unlikely that a sample mean will fall below the population mean stated in the null hypothesis.

DEFINITION

Directional tests, or **one-tailed tests**, are hypothesis tests where the alternative hypothesis is stated as greater than (>) or less than (<) a value stated in the null hypothesis. Hence, the researcher is interested in a specific alternative from the null hypothesis.

EXAMPLE 8.2

Using the same study from Example 8.1, Templer and Tomeo (2002) reported that the population mean on the quantitative portion of the GRE General Test for students taking the exam between 1994 and 1997 was 558 ± 139 ($\mu \pm \sigma$). Suppose we select a sample of 100 students enrolled in an elite private school ($n = 100$). We hypothesize that students at this elite school will score higher than the general population. We record a sample mean equal to 585 ($M = 585$), same as measured in Example 8.1. Compute the one–independent sample *z* test at a .05 level of significance.

Step 1: State the hypotheses. The population mean is 558, and we are testing whether the alternative is greater than (>) this value:

H_0: $\mu = 558$ Mean test scores are equal to 558 in the population of students at the elite school.

H_1: $\mu > 558$ Mean test scores are greater than 558 in the population of students at the elite school.

Step 2: Set the criteria for a decision. The level of significance is .05, which makes the alpha level $\alpha = .05$. To determine the critical value for an upper-tail critical test, we locate the probability .0500 toward the tail in column C in the unit normal table. The *z*-score associated with this probability is between $z = 1.64$ and $z = 1.65$. The average of these *z*-scores is $z = 1.645$. This is the critical value or cutoff for the rejection region. Figure 8.6 shows that for this test, we place all the value of alpha in the upper tail of the standard normal distribution.

NOTE: For one-tailed tests, the alpha level is placed in a single tail of a distribution. For upper-tail critical tests, the alpha level is placed above the mean in the upper tail.

Step 3: Compute the test statistic. Step 2 sets the stage for making a decision because the criterion is set. The probability is less than 5% that we will obtain a sample mean that is at least 1.645 standard deviations above the value of the population mean stated in the null hypothesis. In this step, we will compute a test statistic to determine whether or not the sample mean we selected is beyond the critical value we stated in Step 2.

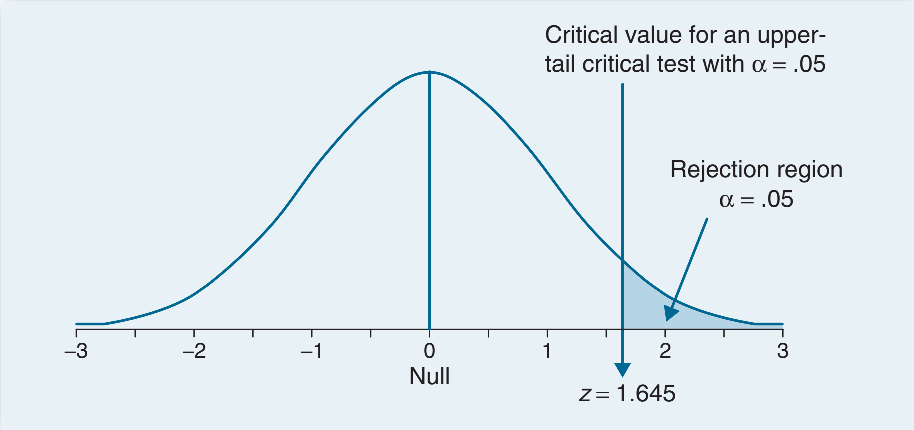


FIGURE 8.6

The critical value (1.645) for a directional (upper-tail critical) hypothesis test at a .05 level of significance. When the test statistic exceeds 1.645, we reject the null hypothesis; otherwise, we retain the null hypothesis.

The test statistic does not change from that in Example 8.1. We are testing the same population, and we measured the same value of the sample mean. We changed only the location of the rejection region in Step 2. The *z* statistic is the same computation as that shown in Example 8.1:

$$z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{585 - 558}{13.9} = 1.94.$$

Step 4: Make a decision. To make a decision, we compare the obtained value to the critical value. We reject the null hypothesis if the obtained value exceeds the critical value. Figure 8.7 shows that the obtained value (Z_{obt} = 1.94) is greater than the critical value; it falls in the rejection region. The decision is to reject the null hypothesis. The *p* value for this test is .0262 (*p* = .0262). We do not double the *p* value for one-tailed tests.

We found in Example 8.2 that if the null hypothesis were true, then *p* = .0262 that we could have selected this sample mean from this population. The criteria we set in Step 2 was that the probability must be less than 5% that we obtain a sample mean, if the null hypothesis were true. Since *p* is less than 5%, we decide to reject the null hypothesis. We decide that the mean score on the GRE General Test in this

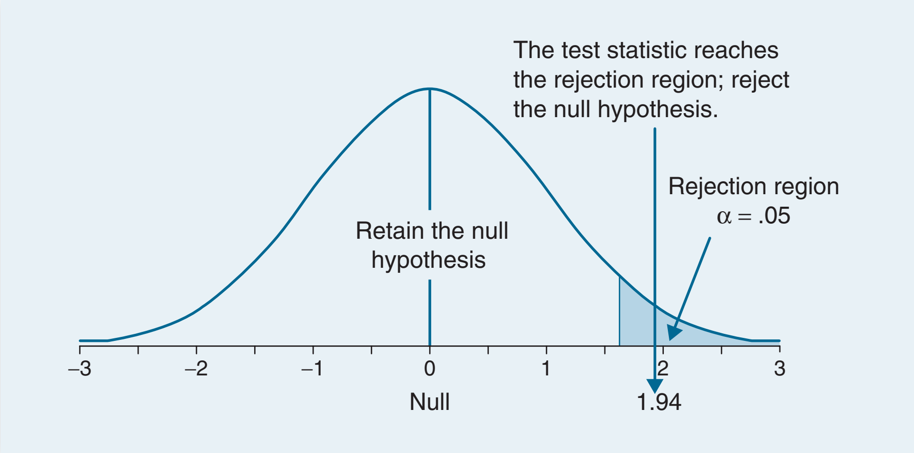


FIGURE 8.7

Since the obtained value reaches the rejection region, we decide to reject the null hypothesis.

population is not 558, which was the value stated in the null hypothesis. Also, notice that we made two different decisions using the same data in Examples 8.1 and 8.2. This outcome is explained further in Section 8.6.

DIRECTIONAL, LOWER-TAIL CRITICAL HYPOTHESIS TESTS (H1: <)

NOTE: A lower-tail critical test is conducted when it is not possible or highly unlikely that a sample mean will fall above the population mean stated in the null hypothesis.

In Example 8.3, we will use the z test for a directional, or one-tailed test, where the alternative hypothesis is stated as *less than* (<) the null hypothesis. For a lower-tail critical test, or a less than statement, we place the level of significance or critical value in the lower tail of the sampling distribution. So we are interested in any alternative less than the value stated in the null hypothesis. This test is appropriate when it is not possible or highly unlikely that a sample mean will fall above the population mean stated in the null hypothesis.

EXAMPLE 8.3

Using the same study from Example 8.1, Templer and Tomeo (2002) reported that the population mean on the quantitative portion of the GRE General Test for those taking the exam between 1994 and 1997 was 558 ± 139 ($\mu \pm \sigma$). Suppose we select a sample of 100 students enrolled in a school with low funding and resources ($n = 100$). We hypothesize that students at this school will score lower than the general population. We record a sample mean equal to 585 ($M = 585$), same as measured in Examples 8.1 and 8.2. Compute the one–independent sample z test at a .05 level of significance.

Step 1: State the hypotheses. The population mean is 558, and we are testing whether the alternative is less than (<) this value:

H_0: $\mu = 558$ Mean test scores are equal to 558 in the population at this school.

H_1: $\mu < 558$ Mean test scores are less than 558 in the population at this school.

NOTE: For one-tailed tests, the alpha level is placed in a single tail of the distribution. For lower-tail critical tests, the alpha is placed below the mean in the lower tail.

Step 2: Set the criteria for a decision. The level of significance is .05, which makes the alpha level $\alpha = .05$. To determine the critical value for a lower-tail critical test, we locate the probability .0500 toward the tail in column C in the unit normal table. The z-score associated with this probability is again $z = 1.645$. Since this test is a lower-tail critical test, we place the critical value the same distance below the mean: The critical value for this test is $z = -1.645$. All of the alpha level is placed in the lower tail of the distribution beyond the critical value. Figure 8.8 shows the standard normal distribution, with the rejection region beyond the critical value.

Step 3: Compute the test statistic. Step 2 sets the stage for making a decision because the criterion is set. The probability is less than 5% that we will obtain a sample mean that is at least 1.645 standard deviations below the value of the population mean stated in the null hypothesis. In this step, we will compute a test statistic to determine whether or not the sample mean we selected is beyond the critical value we stated in Step 2.

The test statistic does not change from that used in Example 8.1. We are testing the same population, and we measured the same value of the sample mean. We changed

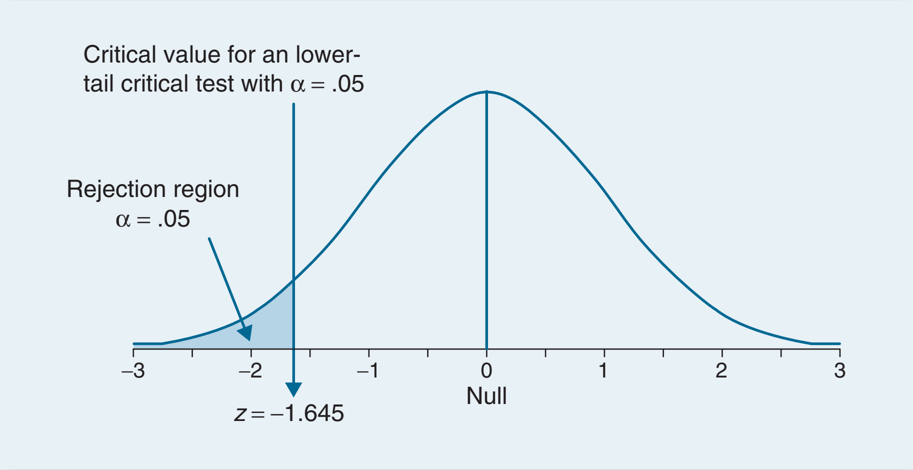


FIGURE 8.8

The critical value (−1.645) for a directional (lower-tail critical) test at a .05 level of significance. When the test statistic is less than −1.645, we reject the null hypothesis; otherwise, we retain the null hypothesis.

only the location of the rejection region in Step 2. The *z* statistic is the same computation as that shown in Example 8.1:

$$z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{585 - 558}{13.9} = 1.94.$$

Step 4: Make a decision. To make a decision, we compare the obtained value to the critical value. We reject the null hypothesis if the obtained value exceeds the critical value. Figure 8.9 shows that the obtained value (Z_{obt} = +1.94) does not exceed the critical value. Instead, the value we obtained is located in the opposite tail. The decision is to retain the null hypothesis.

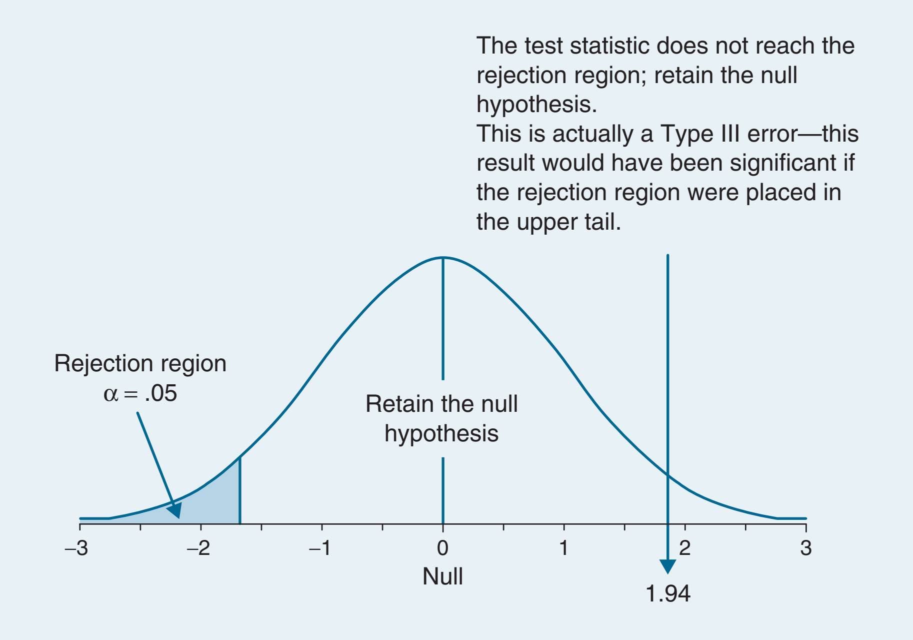


FIGURE 8.9

Since the obtained value does not reach the rejection region, we decide to retain the null hypothesis.

NOTE: A Type III error occurs when the rejection region is located in the wrong tail. This type of error is only possible for one-tailed tests.

The decision in Example 8.3 was to retain the null hypothesis, although if we placed the rejection region in the upper tail (as we did in Example 8.2), we would have decided to reject the null hypothesis. We anticipated that scores would be worse, and instead, they were better than the value stated in the null hypothesis. When we fail to reject the null hypothesis because we placed the rejection region in the wrong tail, we commit a **Type III error** (Kaiser, 1960).

DEFINITION

A **Type III error** occurs with one-tailed tests, where the researcher decides to retain the null hypothesis because the rejection region was located in the wrong tail.

The "wrong tail" refers to the opposite tail from where a difference was observed and would have otherwise been significant.

8.6 RESEARCH IN FOCUS: DIRECTIONAL VERSUS NONDIRECTIONAL TESTS

Kruger and Savitsky (2006) conducted a study in which they performed two tests on the same data. They completed an upper-tail critical test at $\alpha = .05$ and a two-tailed test at $\alpha = .10$. A shown in Figure 8.10, these are similar tests, except in the upper-tail test, all the alpha level is placed in the upper tail, and in the two-tailed test, the alpha level is split so that .05 is placed in each tail. When the researchers showed these results to a group of participants, they found that participants were more persuaded by a significant result when it was described as a one-tailed test, $p < .05$, than when it was described as a two-tailed test, $p < .10$. This was interesting because the two results were identical—both tests were associated with the same critical value in the upper tail.

Most editors of peer-reviewed journals in behavioral research will not publish the results of a study where the level of significance is greater than .05. Although the two-tailed test, $p < .10$, was significant, it is unlikely that the results would be published in a peer-reviewed scientific journal. Reporting the same results as a one-tailed test, $p < .05$, makes it more likely that the data will be published.

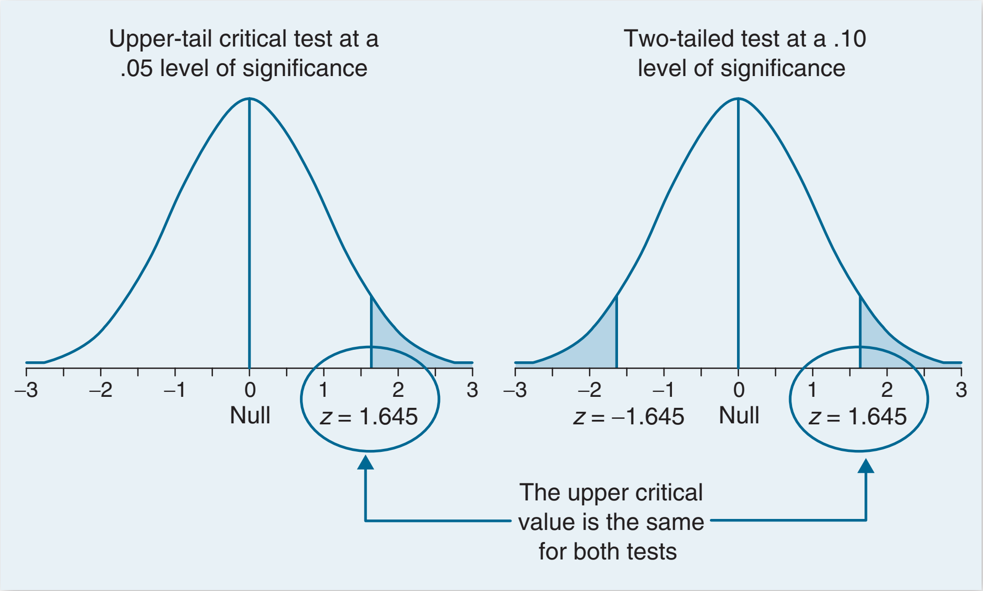


FIGURE 8.10

When $\alpha = .05$, all of that value is placed in the upper tail for an upper-tail critical test. The two-tailed equivalent would require a test with $\alpha = .10$, such that .05 is placed in each tail.

The two-tailed test is more conservative; it makes it more difficult to reject the null hypothesis. It also eliminates the possibility of committing a Type III error. The one-tailed test, though, is associated with greater power. If the value stated in the null hypothesis is false, then a one-tailed test will make it easier to detect this (i.e., lead to a decision to reject the null hypothesis). Because the one-tailed test makes it easier to reject the null hypothesis, it is important that we justify that an outcome can occur in only one direction. Justifying that an outcome can occur in only one direction is difficult for much of the data that behavioral researchers measure. For this reason, most studies in behavioral research are two-tailed tests.

NOTE: Two-tailed tests are more conservative and eliminate the possibility of committing a Type III error. One-tailed tests are associated with more power, assuming the value stated in the null hypothesis is wrong.

LEARNING CHECK 5

1. Is the following set of hypotheses appropriate for a directional or a nondirectional hypothesis test?
   H_0: $\mu = 35$
   H_1: $\mu \neq 35$

2. A researcher conducts a one–independent sample *z* test. The z statistic for the upper-tail critical test at a .05 level of significance was $Z_{obt} = 1.84$. What is the decision for this test?

3. A researcher conducts a hypothesis test and finds that the probability of selecting the sample mean is $p = .0689$ if the value stated in the null hypothesis is true. What is the decision for a hypothesis test at a .05 level of significance?

4. Which type of test, one-tailed or two-tailed, is associated with greater power to detect an effect when the null hypothesis is false?

Answers: 1. A nondirectional (two-tailed) hypothesis test; 2. Reject the null; 3. Retain the null; 4. One-tailed tests.

MEASURING THE SIZE OF AN EFFECT: COHEN'S *d* 8.7

A decision to reject the null hypothesis means that an effect is significant. For a one-sample test, an **effect** is the difference between a sample mean and the population mean stated in the null hypothesis. In Example 8.2, we found a significant effect, meaning that the sample mean, $M = 585$, was significantly larger than the value stated in the null hypothesis, $\mu = 558$. Hypothesis testing identifies whether an effect exists in a population. When a sample mean is likely to occur if the null hypothesis were true ($p > .05$), we decide that an effect doesn't exist in a population; the effect is insignificant. When a sample mean is unlikely to occur if the null hypothesis were true ($p < .05$), we decide that an effect does exist in a population; the effect is significant. Hypothesis testing does not, however, inform us of how big the effect is.

To determine the size of an effect, we compute **effect size.** There are two ways to calculate the size of an effect. We can determine:

1. How far scores shifted in the population

2. The percent of variance that can be explained by a given variable

DEFINITION

For a single sample, an **effect** is the difference between a sample mean and the population mean stated in the null hypothesis. In hypothesis testing, an effect is insignificant when we retain the null hypothesis; an effect is significant when we reject the null hypothesis.

Effect size is a statistical measure of the size of an effect in a population, which allows researchers to describe how far scores shifted in the population, or the percent of variance that can be explained by a given variable.

NOTE: Cohen's d is a measure of the number of standard deviations an effect is shifted above or below the population mean stated by the null hypothesis.

Effect size is most meaningfully reported with significant effects when the decision was to reject the null hypothesis. If an effect is not significant, as in instances when we retain the null hypothesis, then we are concluding that an effect does not exist in a population. It makes little sense to compute the size of an effect that we just concluded doesn't exist. In this section, we describe how far scores shifted in the population using a measure of effect size called Cohen's *d*.

Cohen's *d* measures the number of standard deviations an effect shifted above or below the population mean stated by the null hypothesis. The formula for Cohen's *d* replaces the standard error in the denominator of the test statistic with the population standard deviation (Cohen, 1988):

$$\text{Cohen's } d = \frac{M - \mu}{\sigma}.$$

The value of Cohen's *d* is zero when there is no difference between two means and increases as the differences get larger. To interpret values of *d,* we refer to **Cohen's effect size conventions** outlined in Table 8.6. The sign of *d* indicates the direction of the shift. When values of *d* are positive, an effect shifted above the population mean; when values of *d* are negative, an effect shifted below the population mean.

NOTE: Hypothesis testing determines whether an effect exists in a population. Effect size measures the size of an observed effect from small to large.

TABLE 8.6 Cohen's effect size conventions.

Description of Effect	Effect Size (*d*)
Small	$d < 0.2$
Medium	$0.2 < d < 0.8$
Large	$d < 0.8$

DEFINITION

Cohen's *d* is a measure of effect size in terms of the number of standard deviations that mean scores shifted above or below the population mean stated by the null hypothesis. The larger the value of *d,* the larger the effect in the population.

Cohen's effect size conventions are standard rules for identifying small, medium, and large effects based on typical findings in behavioral research.

In Example 8.4, we will compute effect size for the research study in Examples 8.1 to 8.3. Since we tested the same population and measured the same sample mean in each example, the effect size estimate will be the same for all examples.

EXAMPLE 8.4

In Examples 8.1 to 8.3, we used data given by Templer and Tomeo (2002). They reported that the population mean on the quantitative portion of the GRE General Test for those taking the exam between 1994 and 1997 was 558 ± 139 ($\mu \pm \sigma$). In each example, the mean test score in the sample was 585 ($M = 585$). What is the effect size for this test using Cohen's *d?*

The numerator for Cohen's *d* is the difference between the sample mean ($M = 585$) and the population mean ($\mu = 558$). The denominator is the population standard deviation ($\sigma = 139$):

$$d = \frac{M - \mu}{\sigma} = \frac{27}{139} = 0.19.$$

We conclude that the observed effect shifted 0.19 standard deviations above the mean in the population. This way of interpreting effect size is illustrated in Figure 8.11. We are stating that students in the elite school scored 0.19 standard deviations higher, on average, than students in the general population. This interpretation is most meaningfully reported with Example 8.2 since we decided to reject the null hypothesis using this example. Table 8.7 compares the basic characteristics of hypothesis testing and effect size.

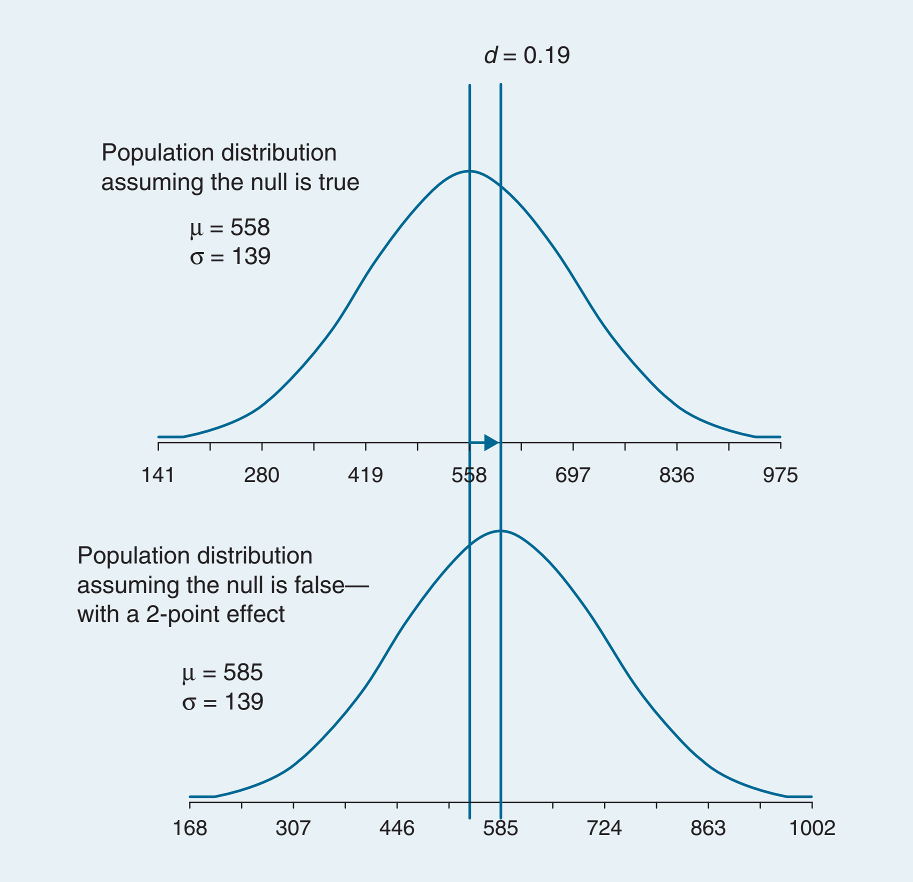


FIGURE 8.11

Effect size. Cohen's *d* estimates the size of an effect using the population standard deviation as an absolute comparison. A 27-point effect shifted the distribution of scores in the population by 0.19 standard deviations.

TABLE 8.7 Distinguishing characteristics for significance testing and effect size.

	Hypothesis (Significance) Testing	Effect Size (Cohen's *d*)
Value being measured?	*p* value	*d*
What type of distribution is the test based upon?	Sampling distribution	Population distribution
What does the test measure?	The probability of obtaining a measured sample mean	The size of a measured treatment effect in the population
What can be inferred from the test?	Whether the null hypothesis is true or false	Whether the size of a treatment effect is small to large
Can this test stand alone in research reports?	Yes, the test statistic can be reported without an effect size	No, effect size is almost always reported with a test statistic

LEARNING CHECK 6

1. ________ measures the size of an effect in a population, whereas ______________ measures whether an effect exists in a population.

2. The scores for a population are normally distributed with a mean equal to 25 and standard deviation equal to 6. A researcher selects a sample of 36 students and measures a sample mean equal to 23 (M = 23). For this example:
   a. What is the value of Cohen's d?
   b. Is this effect size small, medium, or large?

Answers: 1. Effect size, hypothesis or significance testing; 2. (a) $d = \frac{23-25}{6} = -0.33$, (b) Medium effect size.

8.8 EFFECT SIZE, POWER, AND SAMPLE SIZE

One advantage of knowing effect size, *d,* is that its value can be used to determine the power of detecting an effect in hypothesis testing. The likelihood of detecting an effect, called power, is critical in behavioral research because it lets the researcher know the probability that a randomly selected sample will lead to a decision to reject the null hypothesis, if the null hypothesis is false. In this section, we describe how effect size and sample size are related to power.

THE RELATIONSHIP BETWEEN EFFECT SIZE AND POWER

As effect size increases, power increases. To illustrate, we will use a random sample of quiz scores in two statistics classes shown in Table 8.8. Notice that only the standard deviation differs between these populations. Using the values given in Table 8.8, we already have enough information to compute effect size:

TABLE 8.8 Characteristics for two hypothetical populations of quiz scores.

Class 1	Class 2
$M_1 = 40$	$M_2 = 40$
$\mu_1 = 38$	$\mu_2 = 38$
$\sigma_1 = 10$	$\sigma_2 = 2$

$$\text{Effect size for Class 1}: d = \frac{M - \mu}{\sigma} = \frac{40 - 38}{10} = 0.20.$$

$$\text{Effect size for Class 2}: d = \frac{M - \mu}{\sigma} = \frac{40 - 38}{10} = 1.00.$$

The numerator for each effect size estimate is the same. The mean difference between the sample mean and the population mean is 2 points. Although there is a 2-point effect in both Class 1 and Class 2, Class 2 is associated with a much larger effect size in the population because the standard deviation is smaller. Since a larger effect size is associated with greater power, we should find that it is easier to detect the 2-point effect in Class 2. To determine whether this is true, suppose we select a sample of 30 students ($n = 30$) from each class and measure the same sample mean value that is listed in Table 8.8. Let's determine the power of each test when we conduct an upper-tail critical test at a .05 level of significance.

To determine the power, we will first construct the sampling distribution for each class, with a mean equal to the population mean and standard error equal to $\frac{\sigma}{\sqrt{n}}$:

$$\text{Sampling distribution for Class 1: Mean: } \mu_M = 38$$

$$\text{Standard error: } \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{30}} = 1.82$$

$$\text{Sampling distribution for Class 2: Mean: } \mu_M = 38$$

$$\text{Standard error: } \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.37$$

If the null hypothesis is true, then the sampling distribution of the mean for alpha (α), the type of error associated with a true null hypothesis, will have a mean equal to 38. We can now determine the smallest value of the sample mean that is the cutoff for the rejection region, where we decide to reject that the true population mean is 38. For an upper-tail critical test using a .05 level of significance, the critical value is 1.645. We can use this value to compute a *z* transformation to determine what sample mean value is 1.645 standard deviations above 38 in a sampling distribution for samples of size 30:

$$\text{Cutoff for } \alpha \text{ (Class 1): } 1.645 = \frac{M - 38}{1.82}$$

$$M = 40.99$$

$$\text{Cutoff for } \alpha \text{ (Class 2): } 1.645 = \frac{M - 38}{0.37}$$

$$M = 38.61$$

If we obtain a sample mean equal to 40.99 or higher in Class 1, then we will reject the null hypothesis. If we obtain a sample mean equal to 38.61 or higher in Class 2, then we will reject the null hypothesis. To determine the power for this test, we assume that the sample mean we selected ($M = 40$) is the true population mean—we are therefore assuming that the null hypothesis is false. We are asking the following question: If we are correct and there is a 2-point effect, then what is the probability that we will detect the effect? In other words, what is the probability that a sample randomly selected from this population will lead to a decision to reject the null hypothesis?

NOTE: As the size of an effect increases, the power to detect the effect also increases.

If the null hypothesis is false, then the sampling distribution of the mean for β, the type of error associated with a false null hypothesis, will have a mean equal to 40. This is what we believe is the true population mean, and this is the only change; we do not change the standard error. Figure 8.12 shows the sampling distribution for Class 1, and Figure 8.13 shows the sampling distribution for Class 2, assuming the null hypothesis is correct (top graph) and assuming the 2-point effect exists (bottom graph).

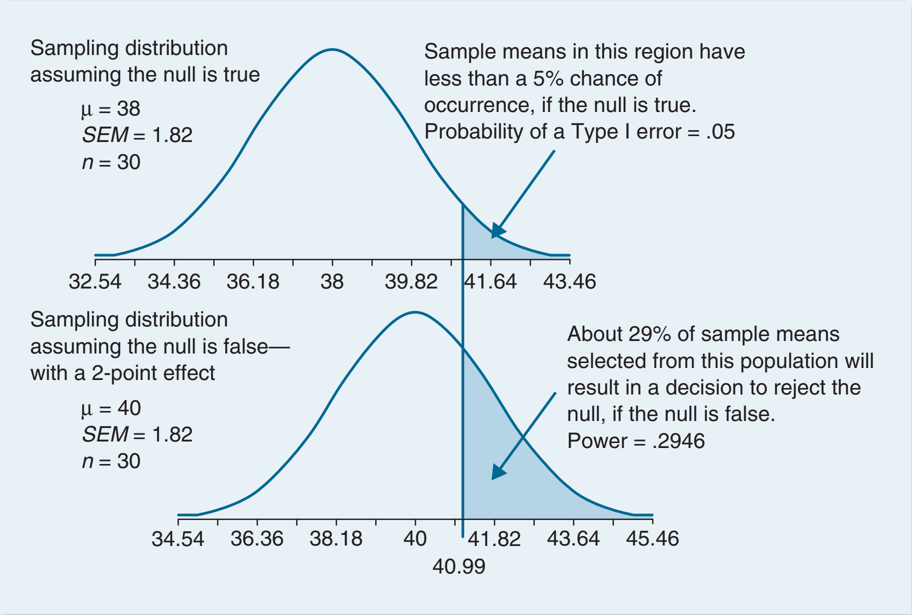


FIGURE 8.12

Small effect size and low power for Class 1. In this example, when alpha is .05, the critical value or cutoff for alpha is 40.99. When α = .05, notice that only about 29% of samples will detect this effect (the power). So even if the researcher is correct, and the null is false (with a 2-point effect), only about 29% of the samples he or she selects at random will result in a decision to reject the null hypothesis.

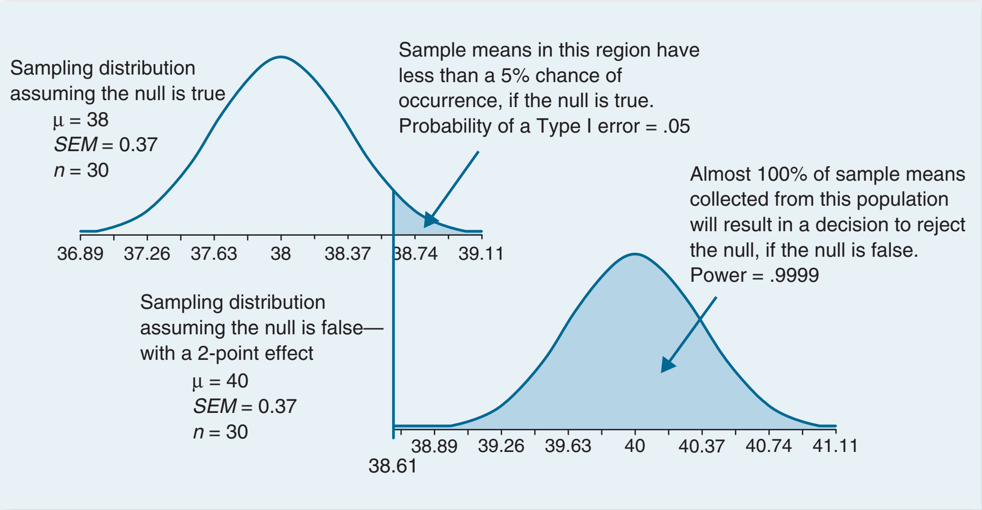


FIGURE 8.13

Large effect size and high power for Class 2. In this example, when alpha is .05, the critical value or cutoff for alpha is 38.61. When $\alpha = .05$, notice that practically any sample will detect this effect (the power). So if the researcher is correct, and the null is false (with a 2-point effect), nearly 100% of the samples he or she selects at random will result in a decision to reject the null hypothesis.

If we are correct, and the 2-point effect exists, then we are much more likely to detect the effect in Class 2 for $n = 30$. Class 1 has a small effect size ($d = .20$). Even if we are correct, and a 2-point effect does exist in this population, then of all the samples of size 30 we could select from this population, only about 29% (power = .2946) of those samples will show the effect (i.e., lead to a decision to reject the null). The probability of correctly rejecting the null hypothesis (power) is low.

Class 2 has a large effect size ($d = 1.00$). If we are correct, and a 2-point effect does exist in this population, then of all the samples of size 30 we could select from this population, nearly 100% (power = .9999) of those samples will show the effect (i.e., lead to a decision to reject the null hypothesis). Hence, we have more power to detect an effect in this population, and correctly reject the null hypothesis.

THE RELATIONSHIP BETWEEN SAMPLE SIZE AND POWER

To overcome low effect size, we can increase the sample size. Increasing sample size decreases standard error, thereby increasing power. To illustrate, let's compute the test statistic for the one-tailed significance test for Class 1, which had a small effect size. The data for Class 1 are given in Table 8.8 for a sample of 30 participants. The test statistic for Class 1 when $n = 30$ is:

$$z_{obt} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{40 - 38}{\frac{10}{\sqrt{30}}} = 1.10.$$

For a one-tailed test that is upper-tail critical, the critical value is 1.645. The value of the test statistic (+1.10) does not exceed the critical value (+1.645), so we retain the null hypothesis.

Increase the sample size to $n = 100$. The test statistic for Class 1 when $n = 100$ is:

$$z_{obt} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{40 - 38}{\frac{10}{\sqrt{100}}} = 2.00.$$

NOTE: Increasing the sample size increases power by reducing the standard error, thereby increasing the value of the test statistic in hypothesis testing.

The critical value is still 1.645. The value of the test statistic (+2.00) now exceeds the critical value (+1.645), so we reject the null hypothesis.

Notice that increasing the sample size alone led to a decision to reject the null hypothesis. Hence, increasing sample size increases power: It makes it more likely that we will detect an effect, assuming that an effect exists in some population.

LEARNING CHECK 7

1. As effect size increases, what happens to the power?
2. As effect size decreases, what happens to the power?
3. When a population is associated with a small effect size, what can a researcher do to increase the power of the study?
4. True or false: The effect size, power, and sample size associated with a study can affect the decisions we make in hypothesis testing.

Answers: 1. Power increases; 2. Power decreases; 3. Increase the sample size (*n*); 4. True.

8.9 ADDITIONAL FACTORS THAT INCREASE POWER

The power is the likelihood of detecting an effect. Behavioral research often requires a great deal of time and money to select, observe, measure, and analyze data. And the institutions that supply the funding for research studies want to know that they are spending their money wisely and that researchers conduct studies that will show results. Consequently, to receive a research grant, researchers are often required to state the likelihood that they will detect the effect they are studying, assuming they are correct. In other words, researchers must disclose the power of their study.

The typical standard for power is .80. Researchers try to make sure that at least 80% of the samples they select will show an effect when an effect exists in a population. In Section 8.8, we showed that increasing effect size and sample size increases power. In this section, we introduce four additional factors that influence power.

INCREASING POWER: INCREASE EFFECT SIZE, SAMPLE SIZE, AND ALPHA

NOTE: To increase power: increase effect size, sample size, and alpha; decrease beta, population standard deviation, and standard error.

Increasing effect size, sample size, and the alpha level will increase power. Section 8.8 showed that increasing effect size and sample size increases power; here we discuss increasing alpha. The alpha level is the probability of a Type I error; it is the rejection region for a hypothesis test. The larger the rejection region, the greater the likelihood of rejecting the null hypothesis, and the greater the power will be. This was illustrated by the difference in the decisions made for Examples 8.1 and 8.2. Increasing the size of the rejection region in the upper tail in Example 8.2 increased the power to detect the 27-point effect. This is why one-tailed tests are more powerful than two-tailed tests: They increase alpha in the direction that an effect is expected to occur, thereby increasing the power to detect an effect.

INCREASING POWER: DECREASE BETA, STANDARD DEVIATION (σ), AND STANDARD ERROR

Decreasing three factors can increase power. Decreasing beta error (β) increases power. In Table 8.3, β is given as the probability of a Type II error, and 1 – β is given as the power. So the lower β is, the greater the solution will be for 1 – β. For example, say β = .20. In this case, 1 – β = (1 – .20) = .80. If we decreased β, say, to β = .10, the power will increase: 1 – β = (1 – .10) = .90. Hence, decreasing beta error increases power.

Decreasing the population standard deviation (σ) and standard error (σ_M) will also increase power. The population standard deviation is the numerator for computing standard error. Decreasing the population standard deviation will decrease the standard error, thereby increasing the value of the test statistic. To illustrate, suppose that we select a sample from a population of students with quiz scores equal to 10 ± 8 (μ ± σ). We select a sample of 16 students from this population and measure a sample mean equal to 12. In this example, the standard error is:

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{16}} = 2.0.$$

To compute the *z* statistic, we subtract the sample mean from the population mean and divide by the standard error:

$$z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{12 - 10}{2} = 1.00.$$

An obtained value equal to 1.00 does not exceed the critical value for a one-tailed test (critical value = 1.645) or a two-tailed test (critical values = ±1.96). The decision is to retain the null hypothesis.

If the population standard deviation is smaller, the standard error will be smaller, thereby making the value of the test statistic larger. Suppose, for example, that we reduce the population standard deviation to 4. The standard error in this example is now:

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{16}} = 1.0.$$

To compute the *z* statistic, we subtract the sample mean from the population mean and divide by this smaller standard error:

$$z_{obt} = \frac{M - \mu}{\sigma_M} = \frac{12 - 10}{1} = 2.00.$$

An obtained value equal to 2.00 does exceed the critical value for a one-tailed test (critical value = 1.645) and a two-tailed test (critical values = ±1.96). Now the decision is to reject the null hypothesis. Assuming that an effect exists in the population, decreasing the population standard deviation decreases standard error and increases the power to detect an effect. Table 8.9 lists each factor that increases power.

TABLE 8.9 A summary of factors that increase power—the probability of rejecting a false null hypothesis.

To increase power:	
Increase	**Decrease**
d (Effect size)	β (Type II error)
n (Sample size)	σ (Standard deviation)
α (Type I error)	σ_M (Standard error)

8.10 SPSS IN FOCUS: A PREVIEW FOR CHAPTERS 9 TO 18

As discussed in Section 8.5, it is rare that we know the value of the population variance, so the *z* test is not a common hypothesis test. It is so uncommon that SPSS can't be used to compute this test statistic, although it can be used to compute all other test statistics described in this book. For each analysis, SPSS provides output analyses that indicate the significance of a hypothesis test and provide the information needed to compute effect size and even power. SPSS statistical software can be used to compute nearly any statistic or measure used in behavioral research. For this reason, most researchers use SPSS software to analyze their data.

8.11 APA IN FOCUS: REPORTING THE TEST STATISTIC AND EFFECT SIZE

To report the results of a *z* test, we report the test statistic, *p* value, and effect size of a hypothesis test. Here is how we could report the significant result for the *z* statistic in Example 8.2:

> Test scores for students in the elite school were significantly higher than the standard performance of test takers, $z = 1.94$, $p < .03$.

Notice that when we report a result, we do not state that we reject or retain the null hypothesis. Instead, we report whether a result is significant (the decision was to reject the null hypothesis) or not significant (the decision was to retain the null hypothesis). Also, you are not required to report the exact *p* value, although it is recommended. An alternative is to report it in terms of the closest value to the hundredths or thousandths place that its value is less than. In this example, we stated $p < .03$ for a *p* value actually equal to .0262.

Finally, it is often necessary to include a figure or table to illustrate a significant effect and the effect size associated with it. For example, we could describe the effect size in one additional sentence supported by the following figure:

> As shown in Figure 8.14, students in the elite school scored an average of 27 points higher on the exam compared to the general population ($d = .19$).

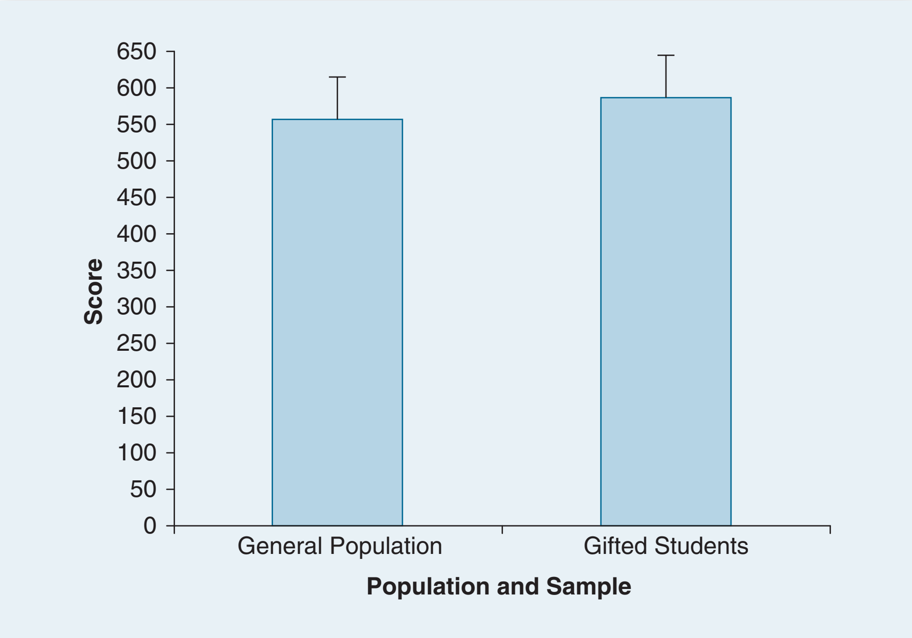


FIGURE 8.14

The mean Graduate Record Examination (GRE) General Test scores among a sample of gifted students compared with the general population. Error bars indicate *SEM*.

In two sentences and a figure, we reported the value of the test statistic, *p* value, effect size, and the mean test scores. The error bars indicate the standard error of the mean for this study.

CHAPTER SUMMARY ORGANIZED BY LEARNING OBJECTIVE

LO 1: Identify the four steps of hypothesis testing.

- **Hypothesis testing,** or **significance testing,** a method of testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true. The four steps of hypothesis testing are as follows:
  - Step 1: State the hypotheses.
  - Step 2: Set the criteria for a decision.
  - Step 3: Compute the test statistic.
  - Step 4: Make a decision.

LO 2: Define null hypothesis, alternative hypothesis, level of significance, test statistic, *p* value, and statistical significance.

- The **null hypothesis (H_0),** stated as the **null,** is a statement about a population parameter, such as the population mean, that is assumed to be true.
- An **alternative hypothesis (H_1)** is a statement that directly contradicts a null hypothesis by stating that the actual value of a population parameter, such as the mean, is less than, greater than, or not equal to the value stated in the null hypothesis.
- **Level of significance** refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis.
- The **test statistic** is a mathematical formula that allows researchers to determine the likelihood or probability of obtaining sample outcomes if the null hypothesis were true. The value of a test statistic can be used to make inferences concerning the value of population parameters stated in the null hypothesis.
- A ***p* value** is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. The *p* value of a sample outcome is compared to the level of significance.
- **Significance,** or **statistical significance,** describes a decision made concerning a value stated in the null hypothesis. When a null hypothesis is rejected, a result is significant. When a null hypothesis is retained, a result is not significant.

LO 3: Define Type I error and Type II error, and identify the type of error that researchers control.

- We can decide to retain or reject the null hypothesis, and this decision can be correct or incorrect. Two types of errors in hypothesis testing are called Type I and Type II errors.
- A **Type I error** is the probability of rejecting a null hypothesis that is actually true. The probability of this type of error is determined by the researcher and stated as the level of significance or alpha level for a hypothesis test.
- A **Type II error** is the probability of retaining a null hypothesis that is actually false.

LO 4: Calculate the one–independent sample *z* test and interpret the results.

- The **one–independent sample *z* test** is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance. The test statistic for this hypothesis test is

$$z_{obt} = \frac{M - \mu}{\sigma_M}, \text{where } \sigma_M = \frac{\sigma}{\sqrt{n}}.$$

- **Critical values,** which mark the cutoffs for the **rejection region,** can be identified for any level of significance. The value of the test statistic is compared to the critical values. When the value of a test statistic exceeds a critical value, we reject the null hypothesis; otherwise, we retain the null hypothesis.

LO 5: Distinguish between a one-tailed and two-tailed test, and explain why a Type III error is possible only with one-tailed tests.

- **Nondirectional (two-tailed) tests** are hypothesis tests where the alternative hypothesis is stated as *not equal to* (≠). So we are interested in any alternative from the null hypothesis.
- **Directional (one-tailed) tests** are hypothesis tests where the alternative hypothesis is

stated as *greater than* (>) or *less than* (<) some value. So we are interested in a specific alternative from the null hypothesis.

- A **Type III error** occurs for one-tailed tests where a result would have been significant in one tail, but the researcher retains the null hypothesis because the rejection region was placed in the wrong or opposite tail.

LO 6: Explain what effect size measures and compute a Cohen's *d* for the one–independent sample *z* test.

- **Effect size** is a statistical measure of the size of an observed effect in a population, which allows researchers to describe how far scores shifted in the population, or the percent of variance that can be explained by a given variable.
- **Cohen's *d*** is used to measure how far scores shifted in a population and is computed using the following formula:

$$\text{Cohen's } d = \frac{M - \mu}{\sigma}.$$

- To interpret the size of an effect, we refer to **Cohen's effect size conventions,** which are standard rules for identifying small, medium, and large effects based on typical findings in behavioral research.

LO 7: Define power and identify six factors that influence power.

- The **power** in hypothesis testing is the probability that a randomly selected sample will show that the null hypothesis is false when the null hypothesis is in fact false.
- To increase the power of detecting an effect in a given population:
  a. Increase effect size (d), sample size (n), and alpha (α).
  b. Decrease beta error (β), population standard deviation (σ), and standard error (σ_M).

APA LO 8: Summarize the results of a one–independent sample *z* test in American Psychological Association (APA) format.

- To report the results of a *z* test, we report the test statistic, *p* value, and effect size of a hypothesis test. In addition, a figure or table is usually provided to summarize the means and standard error or standard deviation measured in a study.

KEY TERMS

alpha (α)
alternative hypothesis (H_1)
beta (β) error
Cohen's *d*
Cohen's effect size conventions
critical values
directional (one-tailed) tests
effect
effect size
hypothesis
hypothesis testing
level of significance
nondirectional (two-tailed) tests
null
null hypothesis (H_0)
obtained value
one–independent sample z test
power
p value
rejection region
significance
significance testing
statistical significance
test statistic
Type I error
Type II error
Type III error
z statistic

END-OF-CHAPTER PROBLEMS

Factual Problems

1. State the four steps of hypothesis testing.
2. What are two decisions that a researcher makes in hypothesis testing?
3. What is a Type I error (α)?
4. What is a Type II error (β)?
5. What is the power in hypothesis testing?
6. What are the critical values for a one–independent sample nondirectional (two-tailed) *z* test at a .05 level of significance?

7. Explain why a one-tailed test is associated with greater power than a two-tailed test.

8. How are the rejection region, probability of a Type I error, level of significance, and alpha level related?

9. Alpha (α) is used to measure the error for decisions concerning true null hypotheses. What is beta (β) error used to measure?

10. What three factors can be increased to increase power?

11. What three factors can be decreased to increase power?

12. Distinguish between the significance of a result and the size of an effect.

Concepts and Application Problems

13. Explain why the following statement is true: The population standard deviation is always larger than the standard error when the sample size is greater than one ($n > 1$).

14. A researcher conducts a hypothesis test and concludes that his hypothesis is correct. Explain why this conclusion is never an appropriate decision in hypothesis testing.

15. The weight (in pounds) for a population of school-aged children is normally distributed with a mean equal to 135 ± 20 pounds ($\mu \pm \sigma$). Suppose we select a sample of 100 children (n = 100) to test whether children in this population are gaining weight at a .05 level of significance.
    a. What are the null and alternative hypotheses?
    b. What is the critical value for this test?
    c. What is the mean of the sampling distribution?
    d. What is the standard error of the mean for the sampling distribution?

16. A researcher selects a sample of 30 participants and makes the decision to retain the null hypothesis. She conducts the same study testing the same hypothesis with a sample of 300 participants and makes the decision to reject the null hypothesis. Give a likely explanation for why the two samples led to different decisions.

17. A researcher conducts a one–independent sample z test and makes the decision to reject the null hypothesis. Another researcher selects a larger sample from the same population, obtains the same sample mean, and makes the decision to retain the null hypothesis using the same hypothesis test. Is this possible? Explain.

18. Determine the level of significance for a hypothesis test in each of the following populations given the specified standard error and critical values. Hint: Refer to the values given in Table 8.4:
    a. $\mu = 100$, $\sigma_M = 8$, critical values: 84.32 and 115.68
    b. $\mu = 100$, $\sigma_M = 6$, critical value: 113.98
    c. $\mu = 100$, $\sigma_M = 4$, critical value: 86.8

19. For each p value stated below: (1) What is the decision for each if $\alpha = .05$? (2) What is the decision for each if $\alpha = .01$?
    a. $p = .1000$
    b. $p = .0250$
    c. $p = .0050$
    d. $p = .0001$

20. For each obtained value stated below: (1) What is the decision for each if $\alpha = .05$ (one-tailed test, upper-tail critical)? (2) What is the decision for each if $\alpha = .01$ (two-tailed test)?
    a. $z_{obt} = 2.10$
    b. $z_{obt} = 1.70$
    c. $z_{obt} = 2.75$
    d. $z_{obt} = -3.30$

21. Will each of the following increase, decrease, or have no effect on the value of a test statistic for the one–independent sample z test?
    a. The sample size is increased.
    b. The population variance is decreased.
    c. The sample variance is doubled.
    d. The difference between the sample mean and population mean is decreased.

22. The police chief selects a sample of 49 local police officers from a population of officers with a mean physical fitness rating of 72 ± 7.0 ($\mu \pm \sigma$) on a 100-point physical endurance rating scale. He measures a sample mean physical fitness rating on

this scale equal to 74. He conducts a one–independent sample z test to determine whether physical endurance increased at a .05 level of significance.

a. State the value of the test statistic and whether to retain or reject the null hypothesis.

b. Compute effect size using Cohen's *d*.

23. A cheerleading squad received a mean rating (out of 100 possible points) of 75 ± 12 ($\mu \pm \sigma$) in competitions over the previous three seasons. The same cheerleading squad performed in 36 local competitions this season with a mean rating equal to 78 in competitions. Suppose we conduct a one–independent sample z test to determine whether mean ratings increased this season (compared to the previous three seasons) at a .05 level of significance.

a. State the value of the test statistic and whether to retain or reject the null hypothesis.

b. Compute effect size using Cohen's *d*.

24. A local school reports that its average GPA is 2.66 ± 0.40 ($\mu \pm \sigma$). The school announces that it will be introducing a new program designed to improve GPA scores at the school. What is the effect size (d) for this program if it is expected to improve GPA by:

a. .05 points?

b. .10 points?

c. .40 points?

25. Will each of the following increase, decrease, or have no effect on the value of Cohen's *d?*

a. The sample size is decreased.

b. The population variance is increased.

c. The sample variance is reduced.

d. The difference between the sample and population mean is increased.

26. State whether the effect size for a 1-point effect ($M - \mu = 1$) is small, medium, or large given the following population variances:

a. $\sigma = 1$

b. $\sigma = 2$

c. $\sigma = 4$

d. $\sigma = 6$

27. As α increases, so does the power to detect an effect. Why, then, do we restrict α from being larger than .05?

28. Will increasing sample size (n) and decreasing the population standard deviation (σ) increase or decrease the value of standard error? Will this increase or decrease power?

Problems in Research

29. **Directional vs. nondirectional hypothesis testing.** In an article reviewing directional and nondirectional tests, Leventhal (1999) stated the following hypotheses concerning the difference between two population means.

A	B
$\mu1 - \mu2 = 0$	$\mu1 - \mu2 = 0$
$\mu1 - \mu2 > 0$	$\mu1 - \mu2 \neq 0$

a. Which did he identify as nondirectional?

b. Which did he identify as directional?

30. **The one-tailed tests.** In their book, *Common Errors in Statistics (and How to Avoid Them),* Good and Hardin (2003) wrote, "No one will know whether your [one-tailed] hypothesis was conceived before you started or only after you'd examined the data" (p. 347). Why do the authors state this as a concern for one-tailed tests?

31. **The hopes of a researcher.** Hayne Reese (1999) wrote, "The standard method of statistical inference involves testing a null hypothesis that the researcher usually hopes to reject" (p. 39). Why does the researcher usually hope to reject the null hypothesis?

32. **Describing the *z* test.** In an article describing hypothesis testing with small sample sizes, Collins and Morris (2008) provided the following description for a z test: "*Z* is considered significant if the difference is more than roughly two standard deviations above or below zero (or more

precisely, $|Z| > 1.96$)" (p. 464). Based on this description:

a. Are the authors referring to critical values for a one- or two-tailed *z* test?

b. What alpha level are the authors referring to?

33. Sample size and power. Collins and Morris (2008) simulated selecting thousands of samples and analyzed the results using many different test statistics. With regard to the power for these samples, they reported that "generally speaking, all tests became more powerful as sample size increased" (p. 468). How did increasing the sample size in this study increase power?

34. Describing hypothesis testing. Blouin and Riopelle (2004) made the following statement concerning how scientists select test statistics: "[This] test is the norm for conducting a test of H_0, when . . . the population(s) are normal with known variance(s)" (p. 78). Based on this description, what test statistic are they describing as the norm? How do you know this?

APPENDIX C

Chapter Solutions for Even-Numbered End-of-Chapter Problems

CHAPTER 8

2. Reject the null hypothesis and retain the null hypothesis.

4. A Type II error is the probability of retaining a null hypothesis that is actually false.

6. Critical values = ±1.96.

8. All four terms describe the same thing. The level of significance is represented by alpha, which defines the rejection region or the region associated with the probability of committing a Type I error.

10. Alpha level, sample size, and effect size.

12. In hypothesis testing, the significance of an effect determines whether an effect exists in some population. Effect size is used as a measure for how big the effect is in the population.

14. All decisions are made about the null hypothesis and not the alternative hypothesis. The only appropriate decisions are to retain or reject the null hypothesis.

16. The sample size in the second sample was larger. Therefore, the second sample had more power to detect the effect, which is likely why the decisions were different.

18.
   a. $\alpha = .05$.
   b. $\alpha = .01$.
   c. $\alpha = .001$.

20.
   1a. Reject the null hypothesis.
   1b. Reject the null hypothesis.
   1c. Reject the null hypothesis.
   1d. Retain the null hypothesis.
   2a. Retain the null hypothesis.
   2b. Retain the null hypothesis.
   2c. Reject the null hypothesis.
   2d. Reject the null hypothesis.

22.
   a. $\sigma_M = \frac{7}{\sqrt{49}} = 1.0$; hence, $z_{obt} = \frac{74-72}{1} = 2.00$.

   The decision is to reject the null hypothesis.

   b. $d = \frac{74-72}{7} = .29$. A medium effect size.

24.
   a. $d = \frac{0.05}{0.4} = 0.125$. A small effect size.
   b. $d = \frac{0.1}{0.4} = 0.25$. A medium effect size.
   c. $d = \frac{0.4}{0.4} = 1.00$. A large effect size.

26.

a. $d = \frac{1}{1} = 1.00$. Large effect size.

b. $d = \frac{1}{2} = 0.50$. Medium effect size.

c. $d = \frac{1}{4} = 0.25$. Medium effect size.

d. $d = \frac{1}{6} = .17$. Small effect size.

28. This will decrease standard error, thereby increasing power.

30. The point Good and Hardin (2003) are making is that it is possible with the same data to retain the null for a two-tailed test and reject the null for a one-tailed test where the entire rejection region is placed in a single tail.

32.

a. Two-tailed *z* test.

b. $\alpha = .05$.

34. We would use the *z* test because the population variance is known.