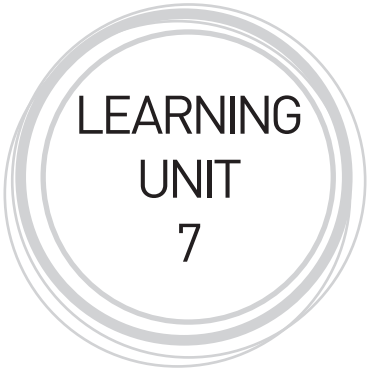


Hypothesis Testing: Significance, Effect Size, and Confidence Intervals

Making observations is something all of us are familiar with. We may observe friends "in love" or athletes "in the zone" or maybe you observe people "overreacting" to a situation. In each case, simply making the observation is not sufficient for science. From a scientific viewpoint, we must first structure our observations such that other people could observe the same things we did. In other words, how do you structure your observations to know what constitutes friends showing that they are "in love" or that an athlete is "in the zone" or that people are "overreacting" to a situation—in such a way that you could gain consensus from others using the same set of procedures to make those observations.

To address this question, we need to know what we expect to observe. In other words, we need to begin with hypotheses that help us to structure our observations. An equally critical step will be in how we test our hypotheses by analyzing the data we collect from our observations. By analyzing data, we can draw conclusions from the observations we make—we can understand the nature of the "effects" we observe. This learning unit provides an essential introduction to understanding the context of and logic for hypothesis testing, which is applied in Sections IV and V. Although we will not use Excel to introduce the nature of hypothesis testing in this learning unit, we will make extensive use of Excel in Sections IV and V, where we apply the general procedures described here for hypothesis testing.

Inferential Statistics and Hypothesis Testing

Inferential statistics allow us to observe samples to learn about behavior in populations that are often too large or inaccessible to observe. 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 150. The sample mean is an *unbiased estimator* of the population mean—if we select a random sample from a population, then on average the value of the sample mean will equal the value of the population mean. In our example, if we select a random sample from this population with a mean of 150, then, on average, the value of a sample mean will equal 150. 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. This was one of the major themes in Learning Unit 6.

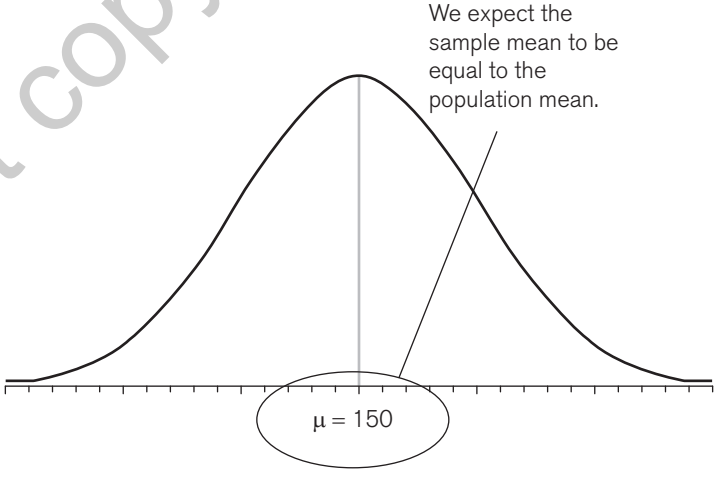
In behavioral research, we select samples to learn more about populations of interest to us. In terms of the mean, we measure a sample mean to learn more about the mean in a population. Therefore, we will use the sample mean to describe the population mean. We begin by stating a **hypothesis** about 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 that of the population mean. The larger the difference or discrepancy between the sample mean and population mean, the less likely it will be that the value of the population mean we hypothesized is correct. This type of experimental situation, using the example of standardized exam scores, is illustrated in Figure 7.1. Although subsequent learning units will cover a variety of questions, research designs, and statistical tests, the underlying reasoning used here applies to research designs associated with various statistical tests.

A **hypothesis** is a statement about or proposed explanation for an observation, a phenomenon, or a scientific problem that can be tested using the research method. A hypothesis is often a statement about the value for a parameter in a population.

The method of evaluating 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, let us use a simple example concerning social media use. According to estimates reported by Mediakix (2016),

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 a hypothesis by determining the likelihood that a sample statistic would be selected if the hypothesis regarding the population parameter were true.

FIGURE 7.1 ● The sampling distribution for a population with a mean equal to 150.



If 150 is the correct population mean, then the sample mean will equal 150, on average, with outcomes farther from the population mean being less and less likely to occur.

the average consumer spends roughly 120 minutes (or 2 hours) a day on social media. Suppose we want to test how the social media use of premillennial consumers (i.e., those born before the millennial generation) compares to that of the average consumer. To make a test, we record the time (in minutes) that a sample of older consumers uses social media per day, and compare this to the average of 120 minutes per day that all consumers (the population) use social media. The mean we measure for these premillennial consumers is a sample mean. We can then compare the mean in our sample to the population mean for all consumers ($\mu = 120$ minutes).

The method of hypothesis testing can be summarized in four steps. We describe each of these four steps in greater detail in the next section. These four steps guide us through various statistical tests in Sections IV and V of this book, whether the statistical tests evaluate means or evaluate variance.

1. To begin, we identify a hypothesis or claim that we feel should be tested. For example, we decide to test whether the mean number of minutes per day that premillennial consumers spend on social media is 120 minutes per day (i.e., the average for all consumers).
2. We select a criterion upon which we decide whether the hypothesis being tested should be accepted or not. For example, the hypothesis is whether or not premillennial consumers spend 120 minutes using social media per day. If premillennial consumers' use of social media is similar to that of the average consumer, then we expect the sample mean will be about 120 minutes. If premillennial consumers spend more or less than 120 minutes using social media per day, then we expect the sample mean will be some value much lower or higher than 120 minutes. However, at what point do we decide that the discrepancy between the sample mean and 120 minutes (i.e., the population mean) is so big that we can reject the notion that premillennial consumers' use of social media is similar to that of the average consumer? In Step 2 of hypothesis testing, we answer this question.
3. Next, we select a sample from the population and measure the sample mean. For example, we can select a sample of 1,000 premillennial consumers and measure the mean time (in minutes) that they use social media per day.
4. Finally, we compare what we observe in the sample to what we expect to observe if the claim we are testing—that premillennial consumers spend 120 minutes using social media per day—is true. We expect the sample mean will be around 120 minutes. The smaller the discrepancy between the sample mean and population mean, the more likely we are to decide that premillennial consumers' use of social media is similar to that of the average consumer (i.e., about 120 minutes per day). The larger the discrepancy between the sample mean and population mean, the more likely we are to decide to reject that claim.

Four Steps to Hypothesis Testing

The goal of hypothesis testing is to determine the likelihood that a sample statistic would be selected if the hypothesis regarding a population parameter were true. In

this section, we describe the four steps of hypothesis testing that were briefly introduced in the previous section:

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 example of social media use, we can state the null hypothesis (H_o) that premillennial consumers use an average of 120 minutes of social media per day: ($\mu = 120$ minutes).

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, and a hypothesis test is structured to decide whether or not to reject this assumption.

This is a starting point so that we can decide whether or not the null hypothesis 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.

The key 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**. In a courtroom, the defendant is assumed to be innocent (this is the null hypothesis, so to speak), so 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.

An **alternative hypothesis (H_1)** is a statement that directly contradicts a null hypothesis by stating 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 null and alternative hypotheses must encompass all possibilities for the population mean. For the example of social media use, we can state that the value in the null hypothesis is equal to 120 minutes. In this way, the null hypothesis value ($\mu = 120$ minutes) and the alternative hypothesis (H_1) value ($\mu \neq 120$ minutes) encompass all possible values for the population mean. If we believe that premillennial consumers use more than (>) or less than (<) 120 minutes of social media per day, then we can make a "greater than" or "less than" statement in the alternative hypothesis—this type of alternative is described in Step 2. Regardless of the decision alternative, the null and alternative hypotheses must encompass all possibilities for the value of the population mean.

Level of significance, or **significance level**, is 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.

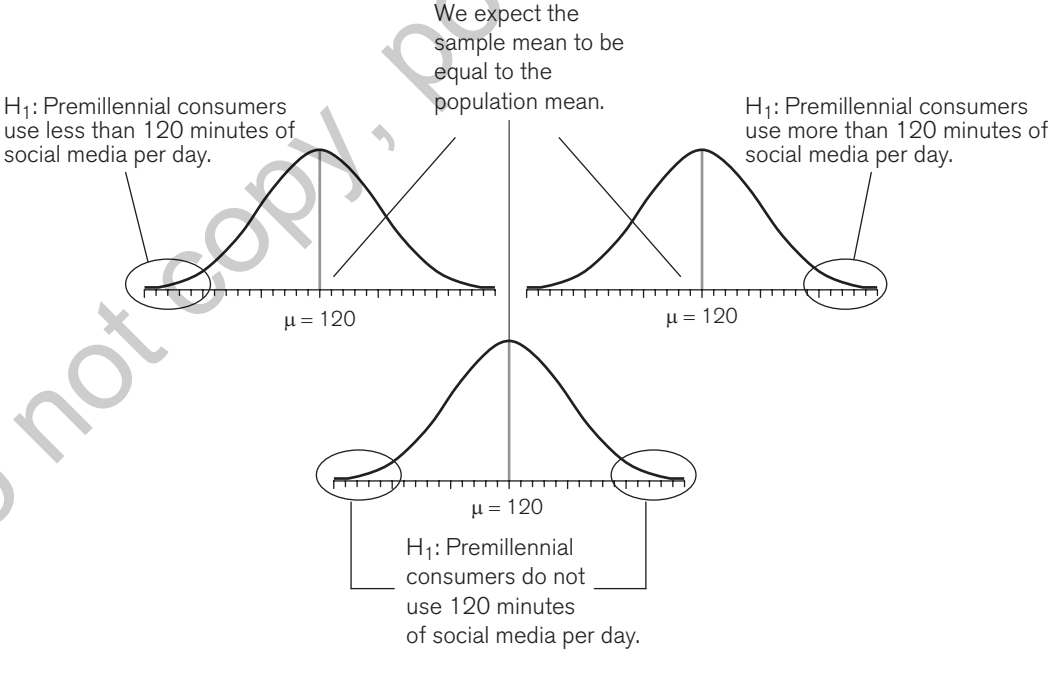
Step 2: Set the criteria for a decision. To set the criteria for a decision, we state the **level of significance** for a hypothesis test. This is similar to the criterion that jurors use in a criminal trial. Jurors decide whether the evidence presented shows guilt *beyond a reasonable doubt* (this is the criterion). Likewise, in hypothesis testing, we collect data to test whether or not the null hypothesis is retained, 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 would be less than 5% if the null hypothesis were true, then we conclude that the sample we selected is too unlikely, and thus we reject the null hypothesis.

The alternative hypothesis is identified so that the criterion can be specifically stated. Remember 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 sample mean that is beyond approximately 2 *SD* from the population mean. For the example of social media use, we can look for the probability of obtaining a sample mean beyond 2 *SD* in the upper tail (greater than 120), the lower tail (less than 120), or both tails (not equal to 120). Figure 7.2 shows the three decision alternatives for a hypothesis test; to conduct a hypothesis test, you choose only one alternative. How to choose an alternative is described in this learning unit. No matter what test you compute, the null and alternative hypotheses must encompass all possibilities for the population mean.

Step 3: Compute the test statistic. Suppose we observe the sample and record a sample mean equal to 100 minutes ($M = 100$) that premillennial consumers use social media per day. Of course, we did not observe everyone in the population, so to make a decision, we need to evaluate how likely this sample outcome is if the

FIGURE 7.2 ● The three decision alternatives for a hypothesis test.



Although a decision alternative can be stated in only one tail, the null and alternative hypotheses should encompass all possibilities for the population mean.

population mean stated in the null hypothesis (120 minutes per day) is true. To determine this likelihood, we use a **test statistic**, which 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 farther the distance, or number of standard deviations, a sample mean outcome 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.

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 or equal to 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:

The **test statistic** is a mathematical formula that identifies how far or how many standard deviations a sample outcome is from the value stated in a null hypothesis. It 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 a null hypothesis.

1. Reject the null hypothesis. The sample mean is associated with a low probability of occurrence when the null hypothesis is true. For this decision, we conclude that the value stated in the null hypothesis is wrong; it is rejected.
2. Retain the null hypothesis. The sample mean is associated with a high probability of occurrence when the null hypothesis is true. For this decision, we conclude that there is insufficient evidence to reject the null hypothesis; this does not mean that the null hypothesis is correct. It is not possible to *prove* the null hypothesis.

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 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.

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 or criterion for making a decision.

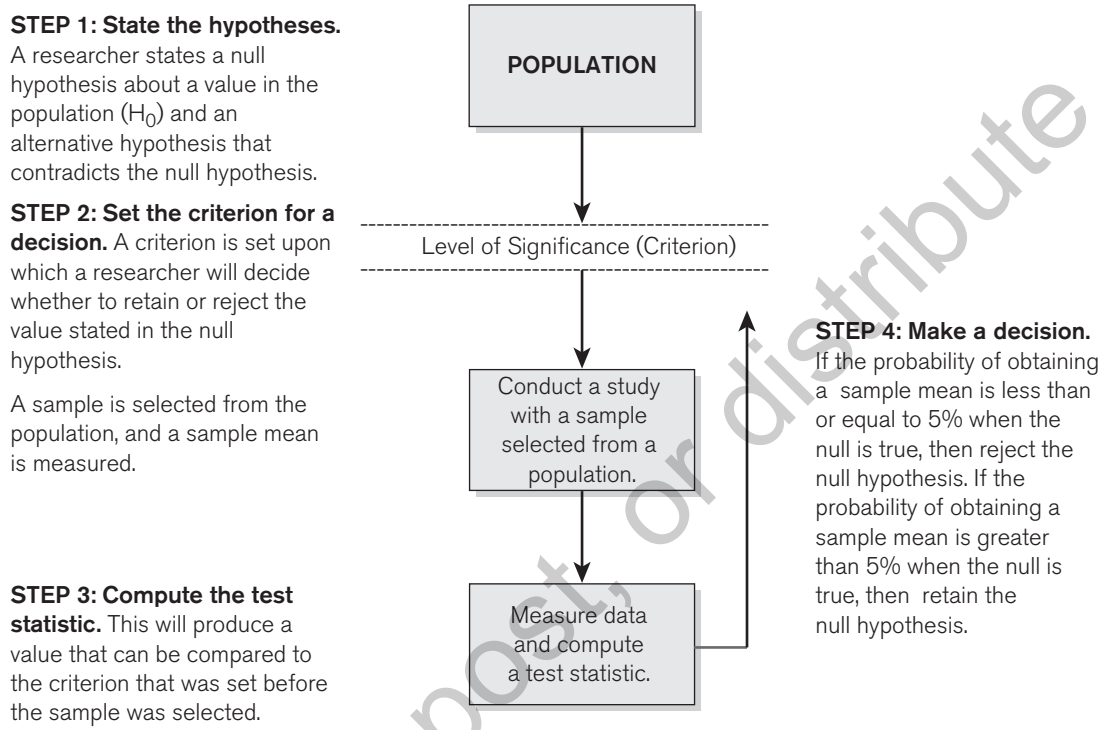
When the *p* value is less than 5% ($p < .05$), we reject the null hypothesis, and when $p = .05$, the decision is also to reject the null hypothesis. When the *p* value is greater than 5% ($p > .05$), we retain the null hypothesis. The decision to reject or retain the null hypothesis is called **significance**. When the *p* value is less than or equal to .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 7.3 summarizes the four steps of hypothesis testing.

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.

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 our decision about a null hypothesis is wrong. Table 7.1 shows that there are four decision alternatives regarding the truth and falsity of the decision we make about a null hypothesis:

FIGURE 7.3 ● A summary of the four steps of hypothesis testing.



If 150 is the correct population mean, then the sample mean will equal 150, on average, with outcomes farther from the population mean being less and less likely to occur.

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

We investigate each decision alternative in this section. Because 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 7.1. In this section, we 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. For this reason, a null result alone is rarely published in scientific journals for behavioral research.

TABLE 7.1 ● Four outcomes for making a decision.

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

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

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 by stating an alpha level.

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.

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.

The incorrect decision is to retain a false null hypothesis: a "false negative" finding. This decision is an example of a **Type II error**, or **beta (β) error**. With each test we make, there is always some probability that the decision is a Type II error. In this decision, we decide not to reject previous notions of truth that are in fact false. While this type of error is often regarded as less problematic than a Type I error (defined in the next paragraph), it can be problematic in many fields, such as in medicine, where testing of treatments could mean life or death for patients.

Decision: Reject 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: a "false positive" finding. 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. Using the courtroom analogy, making this type of error is analogous to finding an innocent person guilty. To minimize this error, we therefore place the burden on the researcher to demonstrate evidence that the null hypothesis is indeed false.

Because 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 ($\alpha = .05$) in behavioral research. To make a decision, we compare the alpha level (or criterion) to the *p* value (the actual likelihood of obtaining a sample mean, if the null were true). When the *p* value is less than the criterion of $\alpha = .05$, we decide to reject the null hypothesis; otherwise, we retain the null hypothesis.

The correct decision is to reject a false null hypothesis. In other words, we decide that the null hypothesis is false when it is indeed false. This decision is called the **power** of the decision-making process, 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 about them. This decision is the most published result in behavioral research.

Nondirectional and Directional Alternatives to the Null Hypothesis

Nondirectional tests, or **two-tailed tests**, are hypothesis tests in which the alternative hypothesis is stated as *not equal to* (≠) a value stated in the null hypothesis. Hence, the researcher is interested in any alternative to the null hypothesis.

Directional tests, or **one-tailed tests**, are hypothesis tests in which 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 to the null hypothesis.

A **Type III error** is a type of error possible with one-tailed tests in which a decision would have been to reject the null hypothesis, but 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.

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 in, as illustrated in Figure 7.2.

For a **nondirectional**, or **two-tailed, test**, the alternative hypothesis is stated as *not equal to* (≠) the null hypothesis. For this test, we divide the level of significance, $p \leq .05$, into both tails of the sampling distribution. Now each tail has a rejection region of less than or equal to .025. We therefore stay neutral in terms of the alternative to the null hypothesis; we are interested in any alternative to the null hypothesis. This is the most common alternative hypothesis tested in behavioral science. In Figure 7.2, this test is illustrated in the bottom figure, where the rejection region is in both tails.

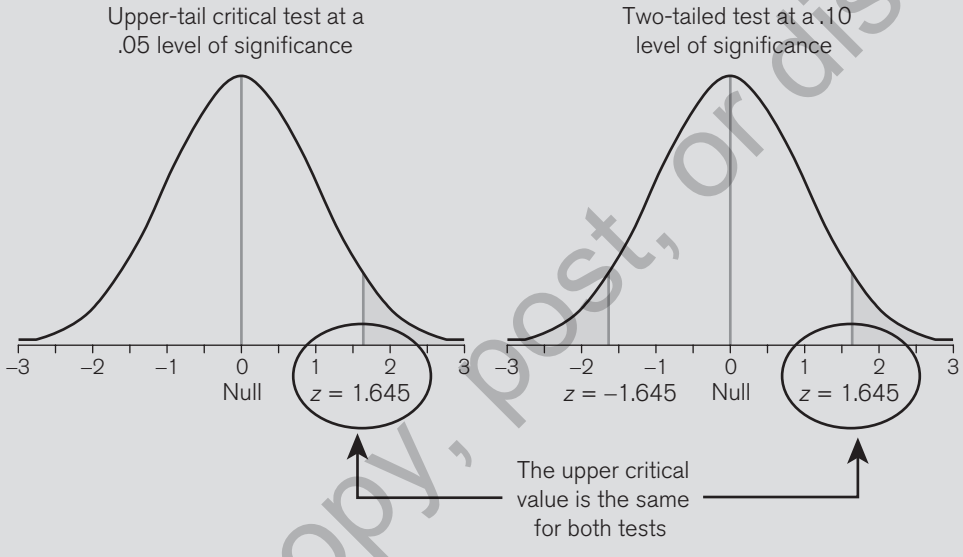
An alternative to the nondirectional test is a **directional**, or **one-tailed**, **test**, where the alternative hypothesis is stated as *greater than* (>) the null hypothesis or *less than* (<) the null hypothesis. For an upper-tail critical test, or a "greater than" statement, we place the level of significance, $p \leq .05$, 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 should only be used when it is impossible or highly unlikely that a sample mean will fall below the population mean stated in the null hypothesis. In Figure 7.2, this test is illustrated in the top right figure where the rejection region is in the upper tail only.

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 should only be used when it is impossible or highly unlikely that a sample mean will fall above the population mean stated in the null hypothesis. In Figure 7.2, this test is illustrated in the top left figure where the rejection region is in the lower tail only.

For directional or one-tailed testing, it is important to consider that this testing creates the unique possibility of committing a **Type III error**. This type of error occurs when a decision would have been to reject the null hypothesis, but the researcher decides to retain the null hypothesis because the rejection region was located in the "wrong tail"—meaning that the effect or difference observed occurred in the opposite tail from where the rejection region was located. This type of error is not possible with a two-tailed test, because the rejection region is located in both tails for such tests. We take a closer look at one- versus two-tailed testing in the next section, where we further evaluate the strengths and limitations of such tests.

TAKING A CLOSER LOOK AT ONE-TAILED AND TWO-TAILED TESTING

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 $\alpha = .10$. As shown in Figure 7.8, 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.



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. Note that the normal distribution is symmetrical, so the cutoff in the lower tail is the same distance below the mean (–1.645; the upper tail is +1.645).

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.

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.

Effect Size

An **effect** is a difference or disparity between what is thought to be true in a population and what is observed in a sample. In hypothesis testing, an effect is not significant when we retain the null hypothesis; an effect is significant when we reject the null hypothesis.

A decision to reject the null hypothesis means that an **effect** is significant. Hypothesis testing identifies whether or not an effect exists in a population. When a sample mean would be likely to occur if the null hypothesis were true ($p > .05$), we decide that an effect does not exist in a population; the effect is not significant. When a sample mean would be unlikely to occur if the null hypothesis were true (typically less than a 5% likelihood, $p < .05$), 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.

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 percentage of variance that can be explained by a given variable.

- *A change or shift in the population,* typically reported in standard deviation units (e.g., a seasonal promotion during an event increased the number of volunteers at the event 0.50 standard deviations above expected rates)
- *A proportion or percentage of variance accounted for,* typically reported as a proportion from 0 to 1.0 or as a percentage from 0% to 100% (e.g., 10% of the variance in academic achievement can be accounted for by the quality of instruction)

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 practical sense to compute the size of an effect that we just concluded does not exist.

Estimation and Confidence Intervals

Estimation is a statistical procedure in which a sample statistic is used to estimate the value of an unknown population parameter. Two types of estimation are point estimation and interval estimation.

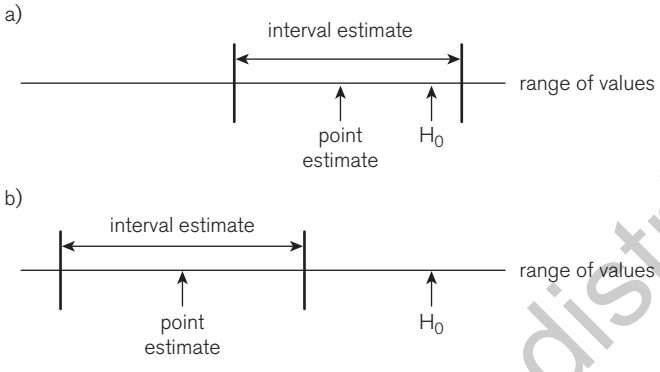
Beyond hypothesis testing, we can also learn more about the mean in a population using a different procedure without ever deciding to retain or reject a null hypothesis. An alternative approach requires only that we set limits for a population parameter within which it is likely to be contained. The goal of this alternative approach, called **estimation**, is the same as that in hypothesis testing—to learn more about the value of a mean in a population of interest.

There are two types of estimates: a point estimate and an interval estimate. When using one sample, a **point estimate** is the sample mean we measure. The advantage of using point estimation is that the *point estimate*, or sample mean, is an unbiased estimator—that is, the sample mean will equal the population mean on average. The disadvantage of using point estimation is that we have no way of knowing for sure whether a sample mean equals the population mean. One way to resolve this disadvantage is to identify a range of values (instead of giving just one value) within which we can be confident that any one of those values is equal to the population mean. The interval or range of possible values within which a population parameter is likely to be contained is called the **interval estimate**. Most often, the point estimate and interval estimate are given together. Thus, researchers report the sample mean (a point estimate) and give an interval within which a population mean is likely to be

A **point estimate** is the use of a sample statistic (e.g., a sample mean) to estimate the value of a population parameter (e.g., a population mean).

An **interval estimate**, often reported as a **confidence interval**, is an interval or range of possible values within which a population parameter is likely to be contained.

FIGURE 7.4 ● Point estimates and interval estimates on a range of values. (a) The interval estimate overlaps H_0. (b) The interval estimate does not overlap H_0.



contained (an *interval estimate*). The interval estimate, often reported as a **confidence interval**, is stated within a given **level of confidence**, which is the likelihood that an interval contains an unknown population mean.

Using estimation, we use the sample mean as a point estimate, and we use the variability as an estimate of the interval. Using the variability helps find a range of sample means within which the population mean is likely to be contained. While we do not "make a decision" per se using estimation, we can use the confidence limits to determine what the decision would have been using hypothesis testing. In terms of the decisions we make in hypothesis testing,

1. If the value stated by a null hypothesis is inside a confidence interval (Figure 7.4a), the decision is to retain the null hypothesis (not significant).
2. If the value stated by a null hypothesis is outside the confidence interval (Figure 7.4b), the decision is to reject the null hypothesis (significant).

Delineating Statistical Effects for Hypothesis Testing

A **level of confidence** is the probability or likelihood that an interval estimate will contain an unknown population parameter (e.g., a population mean).

At a macro level, hypothesis testing provides four core levels of information that can be used not only to make decisions about hypotheses, but also to help describe the nature of the effects being tested. Significance, effect size, and confidence intervals are three levels of information. Table 7.2 summarizes the information provided by each level, and adds a fourth level that is introduced in the next learning unit (Learning Unit 8).

- *Significance*. Is there an effect in the population?
- *Effect size*. What is the size of the effect in the population?
- *Confidence intervals*. Where is the effect likely to be in the population?
- *Power*. What is the likelihood of detecting the effect, if it exists?

When properly understood, addressing each of these levels of information—by answering each of the questions identified—can substantially bolster the comprehensiveness and informativeness of decisions made across hypothesis testing.

TABLE 7.2 ● Delineating significance, effect size, confidence intervals, and power.

Type of Analysis	Informativeness	
Significance	Question Answered	Is there an effect in the population?
	Decision	Reject or retain the null hypothesis
Effect Size	Question Answered	What is the size of an effect in the population?
	Decision	Shift in the population (in standard deviations), or proportion of variance accounted for (as a proportion from 0 to 1.0)
Confidence Intervals	Question Answered	Where is the effect likely to be in the population?
	Decision	Interval or range of possible values within which the parameter we are estimating is likely to be contained
Power	Question Answered	What is the likelihood of detecting an effect, if it exists?
	Decision	Likelihood that an effect, if it exists, will lead to a "reject the null hypothesis" decision