The Single-Sample *t* Test

LEARNING OBJECTIVES

- Choose when to use a single-sample *t* test.
- Calculate the test statistic for a single-sample *t* test.
- Interpret the results of a single-sample *t* test.

CHAPTER OVERVIEW

In the last chapter, the logic of hypothesis testing was introduced with the single-sample *z* test. The single-sample *z* test was great as a prototype of hypothesis testing, but it is rarely used because it requires knowing the population standard deviation of the dependent variable. For most psychological variables—for example, reaction time in a driving simulator, number of words recalled from a list of 20, how close one is willing to get to a snake—the population standard deviation isn't known. In this chapter, we'll learn the single-sample *t* test, a test that like the single-sample *z* test allows one to compare a sample mean to a population mean. The single-sample *t* test, unlike the single-sample *z* test, can be used when a researcher doesn't know the population standard deviation. Thus, it is a more commonly used test.

7.1 Calculating the Single-Sample *t* Test

The **single-sample *t* test** is used to compare a sample mean to a specific value like a population mean. As it deals with means, it is used when the dependent variable is measured at the interval or ratio level. It does the same thing as a single-sample *z* test, but it is more commonly used because it doesn't require knowing the population standard deviation, σ. Instead, the single-sample *t* test uses the sample standard deviation, *s*. As long as there is access to the sample, it is possible to calculate *s*.

For an example, imagine the following scenario. Dr. Farshad, a clinical psychologist, wondered whether adults with attention deficit hyperactivity disorder (ADHD) had reflexes that differed in speed from those of the general population. She located a test for reaction time that was normed on adults in the United States. The average reaction time was 200 milliseconds (msec). This is a population mean, so the Greek letter mu is used to represent it and one would write $\mu = 200$.

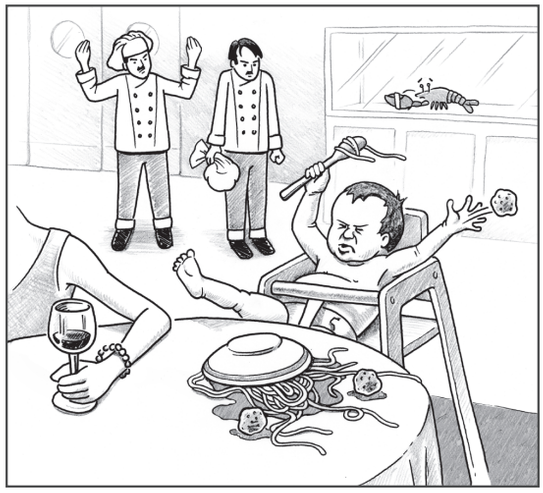
From the various ADHD treatment centers in her home state of Illinois, Dr. Farshad obtained a random sample of 141 adults who had been diagnosed with

ADHD. Each adult was tested individually on the reaction-time test. The mean reaction time for the *sample* of 141 was 220 msec ($M = 220$), with a standard deviation of 27 msec ($s = 27$). (Note that because these are values calculated from a sample, the abbreviations for the mean and standard deviation are Roman, not Greek, letters.)

Dr. Farshad's question is whether adults with ADHD differ in reaction time from the general population. 220 msec is definitely different from 200 msec, so the mean reaction time of this sample of adults with ADHD is different from the population mean. However, Dr. Farshad doesn't know if the difference is a statistically significant one! It's possible that the sample mean is different from the population mean due to sampling error. To answer her question, she's going to need hypothesis testing.

The Six Steps of Hypothesis Testing

Let's follow Dr. Farshad through the six steps of hypothesis testing for the single-sample *t* test: (1) picking a *test;* (2) checking the *assumptions;* (3) listing the *hypotheses;* (4) setting the *decision rule;* (5) *calculating* the test statistic; and (6) *interpreting* the results. Remember the mnemonic: "Tom and Harry despise crabby infants."



The six steps for hypothesis testing are captured in the mnemonic "**T**om **a**nd **H**arry **d**espise **c**rabby **i**nfants."

Step 1 Pick a Test

The first step in hypothesis testing involves selecting the appropriate statistical test. Dr. Farshad is comparing the mean of a sample to the mean of a population, so she could use either a single-sample *z* test or a single-sample *t* test. However, she doesn't know the population standard deviation. She'll have to use the single-sample *t* test to figure out whether the mean reaction time of the sample of adults with ADHD is statistically different from the mean reaction time of the general population.

Step 2 Check the Assumptions

To determine if the sample mean is statistically different from the population mean, Dr. Farshad plans to use a single-sample *t* test. However, a single-sample *t* test can only be used if its assumptions are met. The three assumptions for the single-sample *t* test are the same as they were for the single-sample *z* test and are listed in Table 7.1.

TABLE 7.1 Assumptions for the Single-Sample *t* Test

Assumption	Explanation	Robustness
Random sample	The sample is a random sample from the population.	Robust if violated.
Independence of observations	Cases within the sample don't influence each other.	Not robust to violations.
Normality	The dependent variable is normally distributed in the population.	Robust to violations if the sample size is large.

Note: If a nonrobust assumption is violated, a researcher needs to use a different statistical test, one with different assumptions.

The first assumption involves whether the sample is a random sample from the population. This is a robust assumption. So if it is violated—and it often is—the analysis can still be completed. One just has to be careful about the population to which one generalizes the results. In this example, the sample is a random sample from the population of adults with ADHD in one state. This means that Dr. Farshad should only generalize her results to that state.

The second assumption is that the observations within the sample are independent. This assumption is not robust. If it is violated, one cannot proceed with the single-sample *t* test. In this example, each participant is in the sample only once and the reaction time of each case is not influenced by any other case. The cases were selected randomly and tested individually, so they are independent and the second assumption is not violated.

The third assumption is that the dependent variable, reaction time, is normally distributed in the population. This assumption is robust and the analysis can be completed if the assumption is violated as long as the sample size is large, say, 30 or more, and the deviation from normality is not too large. How does one test this assumption? One way is simply assuming that it is true—it is generally accepted that psychological characteristics (like personality) and physical characteristics (like height) are normally distributed. So, Dr. Farshad is willing to assume that reaction time is normally distributed. Another way is to make a graph of the data. In **Figure 7.1**, the histogram that Dr. Farshad made for the frequency distribution of the reaction-time data can be seen. Though not perfectly normal, it has a normal-ish shape. That, combined with a large sample size, leads Dr. Farshad to feel sure that this assumption has not been violated.

A Common Question

Q Why does the normality assumption exist?

A Hypothesis testing works by comparing the calculated value of the statistic to the expected value. The expected value comes from the sampling distribution, so the normality assumption is really about the shape of the sampling distribution. We know from the central limit theorem that the sampling distribution of the mean will be normal if *N* is large. If *N* is small, the sampling distribution will be normal if the population from which the samples are drawn is normal. And that is why the normality assumption exists.

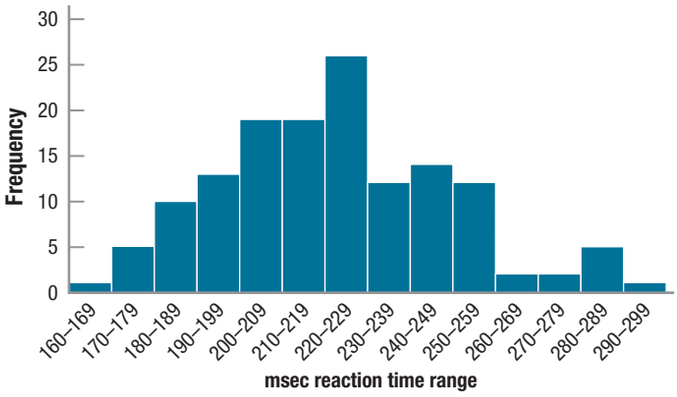


Figure 7.1 Histogram Showing Reaction Time for 141 Adults with Diagnosis of ADHD One way to check the normality assumption, if the sample size is large enough, is to make a histogram and examine it to see if it appears normally distributed. This graph of Dr. Farshad's reaction-time data looks reasonably normal-ish.

The assumptions have been met, so she can proceed with the planned statistical test, the single-sample *t* test.

Step 3 List the Hypotheses

Before writing the hypotheses, the researcher has to know if he or she is doing a one-tailed test or a two-tailed test. Two-tailed tests, called nondirectional tests, are more commonly used and they are a bit more conservative, in the sense that they make it harder to reject the null hypothesis. So, a two-tailed test is the "default" option—if in doubt, choose to do a two-tailed test. But, here's the difference between the two types of tests:

- A two-tailed test is used when the researcher is testing for a difference in either direction. Before collecting her data, Dr. Farshad didn't know whether adults with ADHD had faster or slower reaction times. This calls for a nondirectional test and nondirectional hypotheses.
- A one-tailed test is used when the researcher has a prediction about the direction of the difference *in advance* of collecting the data. As long as the difference is in the expected direction, a one-tailed test makes it easier to obtain statistically significant results.

The hypotheses for the nondirectional (two-tailed) single-sample *t* test for Dr. Farshad's study are

$$H_0: \mu_{\text{ADHD-Adults}} = 200$$

$$H_1: \mu_{\text{ADHD-Adults}} \neq 200$$

The null hypothesis (H_0) says that the mean reaction time for the population of adults with ADHD is *not* different from some specified value. In this case, that value is the mean reaction time, 200 msec, of Americans in general. (The null hypothesis could be phrased as $\mu_{\text{ADHD-Adults}} = \mu_{\text{Americans}} = 200$.)

The alternative hypothesis (H_1) says that the mean reaction time for the population of adults with ADHD is something other than the specified value of 200 msec. It doesn't say whether it is faster or slower, just that the mean reaction time is different. This means that the observed sample mean should be different enough from 200 msec that sampling error does not explain the difference. (The alternative hypothesis could be phrased as $\mu_{\text{ADHD-Adults}} \neq \mu_{\text{Americans}}$.)

If Dr. Farshad had believed that adults with ADHD had, for example, slower reaction times than the general population, then she would have planned a one-tailed test. In such a situation, the hypotheses would have been

$$H_0: \mu_{\text{ADHD-Adults}} \leq 200$$

$$H_1: \mu_{\text{ADHD-Adults}} > 200$$

Never forget that the alternative hypothesis expresses what the researcher believes to be true and that the researcher wants to be forced to reject the null hypothesis, so he or she has to accept the alternative hypothesis. In this one-tailed test example, the researcher's belief is that the reaction time is longer (i.e., slower) for people with ADHD. Once the alternative hypothesis is formed, the null hypothesis is written to make both hypotheses all-inclusive and mutually exclusive. Here, the null hypothesis would be that people with ADHD have reaction times the same as or faster than those of the general population.

Step 4 Set the Decision Rule

Setting the decision rule involves determining when to reject the null hypothesis and when to fail to reject the null hypothesis. To set the decision rule, find the critical value of the test statistic, the boundary between the rare zone and the common zone of the sampling distribution. For the single-sample *t* test, this will be a critical value of *t*.

What is *t*? *t* is the statistic that is calculated in the next step of the six-step hypothesis test procedure. The statistic *t* is a lot like the statistic *z* in the single-sample *z* test:

- If the null hypothesis is true and the sample mean is *exactly* equal to the specified value, *t* will equal zero.
- As the difference between the sample mean and the specified value grows, so does the *t* value.
- When the value of *t* that is calculated (the observed value of *t*) differs enough from zero (the expected value of *t*), then the null hypothesis is rejected.

The Critical Value of *t*

The point that separates "differs enough from zero" from "doesn't differ enough from zero" is called the **critical value of *t*,** abbreviated t_{cv}. To find t_{cv}, three pieces of information are needed:

1. Is the test one-tailed or two-tailed?
2. How willing is one to make a Type I error?
3. How large is the sample size?

The first question, whether a one-tailed or a two-tailed test is being done, was already answered when writing the hypotheses. Dr. Farshad is examining whether adults with ADHD have faster *or* slower reaction times than the general population. The reaction-time study calls for a two-tailed test because the hypotheses didn't specify a direction for the difference.

The second question in determining t_{cv} involves how willing one is to make a Type I error. A Type I error occurs when the researcher concludes, mistakenly, that the null hypothesis should be rejected. A common convention in statistics is to choose to

have no more than a 5% chance of making a Type I error. As alpha (α) is the probability of a Type I error, statisticians phrase this as "setting alpha at .05" or as "$\alpha = .05$." Dr. Farshad has chosen to follow convention and set alpha at .05.

Overall, she wants her chance of making a Type I error to be no more than 5%. She has chosen to do a two-tailed test, so she will need to split that 5% in two and put 2.5% in each tail of the sampling distribution. This means Dr. Farshad will have two critical values of *t,* one positive and one negative. (If the test were one-tailed, all 5% of the rare zone would fall on one side and there would be only one critical value of *t.*)

The third question for determining t_{cv}, how large the sample size is, matters because the shape of the *t* distribution changes as the sample size changes. The tail of a *t* distribution is larger when the sample size is smaller. As a result, the critical value of *t* is farther away from zero when the sample size is small.

This is difficult to visualize without a concrete example, so look at **Figure 7.2** in which two sampling distributions of *t* are superimposed—one for $N = 6$ (the dotted line) and one for $N = 60$ (the solid line). Which line is on top—the dotted line or the solid line—depends on whether one is looking at the center of the distribution or the tails.

Look at the tails. There, the dotted line (the one for the *t* distribution when $N = 6$) is above the solid line for the *t* distribution when $N = 60$. What does it mean for one line to be on top of the other? The *y*-axis measures frequency, so when one line is above another, this means it has a higher frequency at that point.

Focus on the positive side, the right-hand side, of the distribution. (Because the *t* distribution is symmetric, both sides are the same.) Note that around $t = 1.5$, the two lines cross. From that point on, the frequencies at each *t* value are higher for the $N = 6$ distribution than for the $N = 60$ distribution. Which distribution has more cases with *t* values above 1.50, $N = 6$ or $N = 60$? The answer is that the $N = 6$ distribution has a higher percentage of cases in the tail.

The implication of this is important: the total frequency of scores in the tail is higher for the distribution with the smaller sample size ($N = 6$). To cut off the extreme 2.5% of the scores for each distribution, the cut-off point will fall farther away from

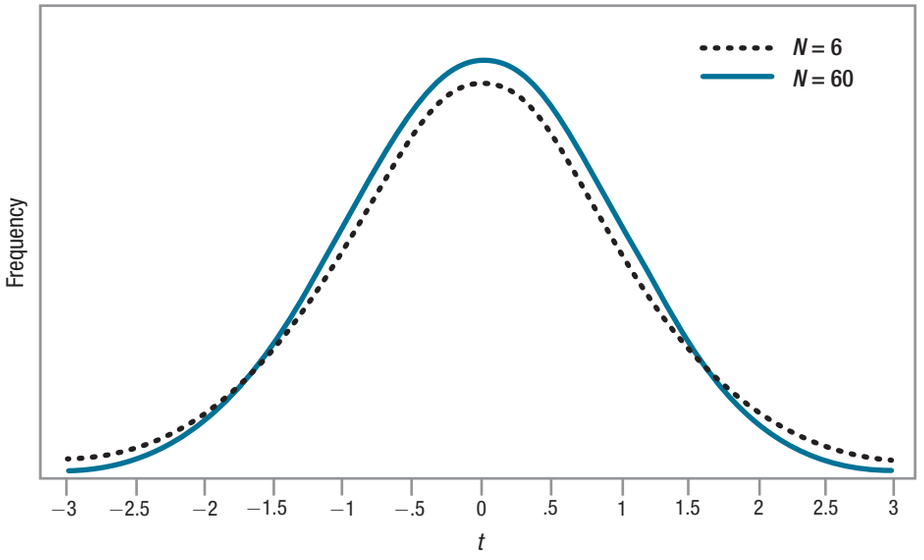


Figure 7.2 Shape of Sampling Distribution of *t* as a Function of Sample Size This figure shows how the shape of the sampling distribution of *t* changes with the sample size. When the sample size is small, there is more variability and more samples have means that fall farther out in a tail. As a result, the critical value of *t* falls farther away from the midpoint when the sample size is small, making it more difficult to reject the null hypothesis.

zero for the $N = 6$ distribution than for the $N = 60$ distribution. In this example, the cut-off points (the critical values of *t*) are 2.57 for $N = 6$ and 2.00 for $N = 60$. The critical value falls closer to zero when the sample size is larger, and it falls farther away from zero when the sample size is smaller.

Sample size affects the critical value of *t*, and the critical value determines the rare and common zones in the sampling distribution. (Remember, when a result falls in the rare zone, the null hypothesis is rejected; when a result falls in the common zone, the null hypothesis is not rejected.) So, sample size affects the ability to reject the null hypothesis:

- When the sample size is small, the critical value of *t* falls farther away from zero, making the rare zone harder to reach, making it more difficult to reject the null hypothesis.
- Larger sample sizes have the opposite effect. When a sample size is large, the critical value of *t* falls closer to zero. As a result, the rare zone is easier to reach, which makes it easier to reject the null hypothesis.

The goal of research almost always is to reject the null hypothesis, so having a larger sample size is an advantage.

To find the critical value of *t* for a given sample size, use Appendix Table 3, a table of critical values of *t*. A portion of Appendix Table 3 is shown in **Table 7.2**.

There are several things to note in Table 7.2. First, note that there are different *rows* for the critical values. The rows represent different critical values of *t* based on sample size. The heading for the rows is "*df*," which stands for *degrees of freedom*. **Degrees of freedom** represent the number of values in a sample that are free to vary. For example, if the mean of three cases is 10, then the values for only two cases are

TABLE 7.2 Critical Values of *t* (Appendix Table 3)

	Critical values of *t*			
df	$\alpha = .05$, one-tailed or $\alpha = .10$, two-tailed	$\alpha = .025$, one-tailed or $\alpha = .05$, two-tailed	$\alpha = .01$, one-tailed or $\alpha = .02$, two-tailed	$\alpha = .005$, one-tailed or $\alpha = .01$, two-tailed
1	6.314	**12.706**	31.821	63.657
2	2.920	**4.303**	6.965	9.925
3	2.353	**3.182**	4.541	5.841
4	2.132	**2.776**	3.747	4.604
5	2.015	**2.571**	3.365	4.032
6	1.943	**2.447**	3.143	3.707
7	1.895	**2.365**	2.998	3.499
8	1.860	**2.306**	2.896	3.355
9	1.833	**2.262**	2.821	3.250
10	1.812	**2.228**	2.764	3.169

Note: The critical value of *t* is the boundary that separates the rare zone of the sampling distribution of *t* from the common zone. For two-tailed tests, the critical values are both positive and negative values. The α level is the probability of making a Type I error. A one-tailed test is used with directional hypotheses and a two-tailed test with nondirectional hypotheses. *df* stands for degrees of freedom. For a single-sample *t* test, $df = N - 1$. The bold numbers represent the critical values of *t* most commonly used, those for a nondirectional (two-tailed) test with a 5% chance ($\alpha = .05$) of making a Type I error.

free to vary. If one case has a score of 9 and another case has a score of 11, then the third case has to have a score of 10. In that example, there are three values, and 2 degrees of freedom—once two values are known, the third is determined.

Degrees of freedom and sample size are yoked together. As the sample size becomes larger, the degrees of freedom increase. And, other things being equal, larger sample sizes are better because there is a greater likelihood that the sample represents the population. In fact, when the sample size is infinitely large, the *t* distribution is the same as the *z* distribution because the whole population is being sampled.

For a single-sample *t* test, the degrees of freedom are calculated as the sample size minus 1. This is shown in Equation 7.1.

Equation 7.1 Degrees of Freedom (*df*) for a Single-Sample *t* Test

$$df = N - 1$$

where df = degrees of freedom
N = sample size

For the reaction-time study, there are 141 participants in the sample, so degrees of freedom are calculated like this:

$$\begin{aligned} df &= 141 - 1 \\ &= 140 \end{aligned}$$

Look at Table 7.2 again. The second thing to note in the table of critical values of *t* is that there are four columns of critical values. Which column to use depends on two factors: (1) if a one-tailed test or two-tailed test is being done, and (2) where alpha, the willingness to make a Type I error, is set.

The critical values of *t* in the column for a two-tailed test with a 5% chance of Type I error (i.e., $\alpha = .05$) have been bolded to make them easier to find because they are the most commonly used.

With the reaction-time study as a two-tailed test, Dr. Farshad would reject the null hypothesis if adults with ADHD had slower reaction times than the general population *or* if adults with ADHD had faster reaction times than the general population. That is what a two-tailed test means.

Under these conditions ($df = 140$, $\alpha = .05$, two-tailed), Dr. Farshad finds that the critical value of *t* is ± 1.977 (i.e., -1.977 and 1.977). These values are marked in Figure 7.3, along with the rare and common zones.

The critical value of *t* would be different if this were a one-tailed test. If Dr. Farshad had reason, in advance of collecting data, to believe that adults with ADHD had slower reaction times, then she could do a one-tailed test. In such a case, she would only reject the null hypothesis if adults with ADHD had slower reaction times than the general population. For $df = 140$ and with $\alpha = .05$, the one-tailed critical value of *t* would be 1.656. This critical value of 1.656 is closer to zero, the midpoint of the sampling distribution, than was the critical value (1.977) for the two-tailed test. This means the rare zone is more easily reached and makes it easier to reject the null hypothesis for the one-tailed test.

Figure 7.4 shows the larger rare zone on the right for the one-tailed test. The rare zone for the two-tailed test is marked with /// and the rare zone for the one-tailed test is marked with \\\. Though the total *percentage* of the curve that is the rare zone is the same for the two tests, notice how more *area* of the rare zone is

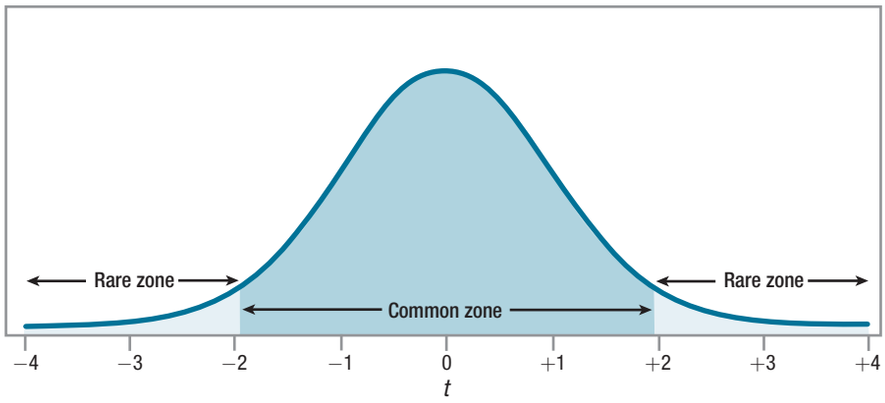


Figure 7.3 Setting the Decision Rule: Two-Tailed, Single-Sample *t* Test In this sampling distribution of *t* values ($df = 140$, $\alpha = .05$, two-tailed), the border between the rare and common zones is ± 1.977. If the observed value of *t* falls in the rare zone, the null hypothesis is rejected; one fails to reject it if the observed value falls in the common zone. Note that the rare zone is split into two parts, half in each tail of the distribution.

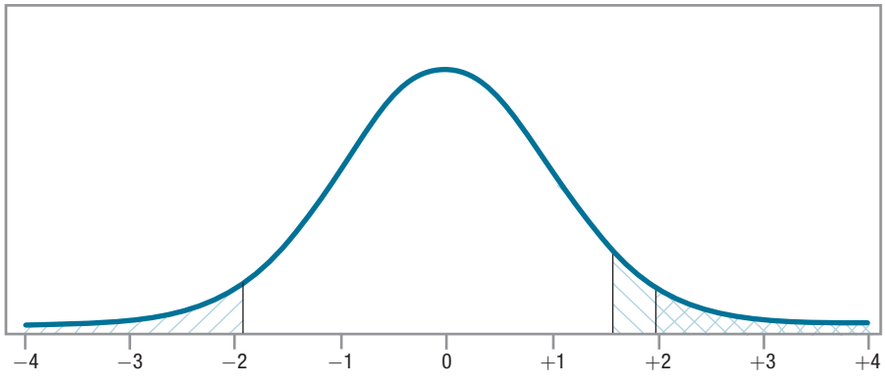


Figure 7.4 Comparing Rare Zones for One-Tailed and Two-Tailed Tests The rare zone of the sampling distribution of *t* with $df = 140$ and $\alpha = .05$ for a one-tailed test is the section marked \\\ to the right of $t = 1.656$. For a two-tailed test, if the difference between the sample and population mean falls in the same direction, it is the area marked /// to the right of $t = 1.977$ (The rare zone for the two-tailed test also includes the /// area on the left). Notice the area of the rare zone is larger for the one-tailed test on one side of the curve, making it that much easier to reject the null hypothesis for a one-tailed test.

marked off on one side by the one-tailed test (\\\\) than by the two-tailed test (///). As a result, it is easier to reject the null hypothesis with a one-tailed test than a two-tailed test, as long as the difference is in the expected direction. This is an advantage for one-tailed tests. The advantage disappears if there is a difference, but it is not in the hypothesized direction.

Dr. Farshad was conducting an exploratory study to see if adults with ADHD differed in reaction time—in either direction—from the general population. So, she's doing a two-tailed test. Here's the general version of the decision rule for a two-tailed test:

If $t \leq -t_{cv}$ *or* if $t \geq t_{cv}$, reject H_0.
If $-t_{cv} < t < t_{cv}$, fail to reject H_0.

When the first statement is true, the observed value of *t* falls in the rare zone, and the null hypothesis is rejected. When the second statement is true, *t* falls in the common zone, and the researcher will fail to reject the null hypothesis.

For the reaction-time study, the specific decision rule is:

- Reject the null hypothesis if $t \leq -1.977$ *or* if $t \geq 1.977$.
- Fail to reject the null hypothesis if $-1.977 < t < 1.977$.

A Common Question

Q Appendix Table 3 doesn't contain degrees of freedom for all possible values. What does one do if $df = 54$, for example?

A In the TV game show *The Price Is Right*, the person who comes closest to guessing the price of an object without exceeding the price wins. Follow that rule and use the degrees of freedom that are closest without going over. If $df = 54$ and the table only contains critical values of *t* for 50 and 55 degrees of freedom, use t_{cv} for $df = 50$.

Step 5 Calculate the Test Statistic

It is time to calculate the *t* value that compares the sample mean for reaction time (220 msec) to the specified value, the population mean for a reaction time of 200 msec. The formula is shown in Equation 7.2.

Equation 7.2 Formula for Calculating a Single-Sample *t* Test

$$t = \frac{M - \mu}{s_M}$$

where t = *t* value
M = sample mean
μ = population mean (or a specified value)
s_M = estimated standard error of the mean (Equation 5.2)

The numerator in the single-sample *t* test formula subtracts the population mean (μ) from the sample mean (M). The difference is then divided by the estimated standard error of the mean (s_M). Before using Equation 7.2, one needs to know the estimated standard error of the mean. This is calculated using Equation 5.2:

$$\begin{aligned} s_M &= \frac{s}{\sqrt{N}} \\ &= \frac{27}{\sqrt{141}} \\ &= \frac{27}{11.8743} \\ &= 2.2738 \\ &= 2.27 \end{aligned}$$

Once the estimated standard error of the mean has been calculated, all the values needed to complete Equation 7.2 are available: $M = 220$, $\mu = 200$, and $s_M = 2.27$. Here are the calculations to find the *t* value:

$$\begin{aligned} t &= \frac{M - \mu}{s_M} \\ &= \frac{220 - 200}{2.27} \\ &= \frac{20.0000}{2.27} \\ &= 8.8106 \\ &= 8.81 \end{aligned}$$

Step 5 is done and Dr. Farshad knows the *t* value: $t = 8.81$. The sixth and final step of hypothesis testing is interpretation. We'll turn to that after a little more practice with the first five steps of hypothesis testing in Worked Example 7.1.

A Common Question

Q The equation for *t*, $t = \frac{M - \mu}{s_M}$ looks a lot like the equation for *z*, $z = \frac{X - \mu}{s}$. Are they similar?

A Yes they are. Both serve to standardize deviation scores, so we can tell how common they are.

Worked Example 7.1

Adjusting to the first year of college can be hard, especially if something traumatic happens back at home. A veterinarian in an imaginary college town, Dr. Richman, wondered if the GPA went down for students who lost a family pet while they were away at college. The vet found 11 students at the college who indicated that their pets at home had died during the year. He had each student report his or her GPA for the year and found the mean was 2.58, with a standard deviation of 0.50. From the college registrar, Dr. Richman learned that the mean GPA for all students for the year was 2.68. Thus, $M = 2.58$ and $\mu = 2.68$. Does losing a pet have a negative impact on GPA?

The answer appears to be yes, because the sample of students who have lost a pet has a GPA lower than the population's GPA. But, isn't it possible, due to sampling error, that a random sample of 11 people from a population where $\mu = 2.68$ could have a sample mean of 2.58? The vet is going to need hypothesis testing to determine if the difference between the sample and the population is statistically significant.

Step 1 Pick a Test. Either a single-sample *z* test or a single-sample *t* test can be used to compare the mean of a sample to a specified value like the mean of a population. However, when the population standard deviation is unknown, a single-sample *t* test must be used. This situation, with σ unknown, calls for a single-sample *t* test.

Step 2 Check the Assumptions. The population is college students who have lost their pets. The sample is not a random sample from the population because all participants self-selected and come from only one college. It is possible that the students who lost pets and who chose not to participate differed in some way from those who volunteered. So, the first assumption is violated. Luckily, the first assumption is robust and can be violated, though Dr. Richman will need to be careful about generalizing from the results.

There's no reason to believe that the second assumption is violated. The observations seem to be independent. No participants are siblings who would have lost the same pet. Further, participants did not come from a support group for people who had lost a pet, where people might have influenced each other in coping with their losses. And each participant was in the sample only once.

The third assumption is not violated because Dr. Richman is willing to assume that GPA, like intelligence, is normally distributed. (The sample size, 11, is too small to provide a meaningful histogram.) None of the nonrobust assumptions was violated, so Dr. Richman can proceed with the single-sample *t* test.

Step 3 List the Hypotheses. Dr. Richman believes that losing a pet will harm adjustment, so he is doing a one-tailed test. Here are the null and alternative hypotheses:

$$H_0: \mu_{\text{StudentswithPetLoss}} \geq 2.68$$

$$H_1: \mu_{\text{StudentswithPetLoss}} < 2.68$$

The null hypothesis (H_0) says that the mean GPA of the population of first-year students who have lost a pet during the year is the same or greater than the mean GPA of students in general for that year. If the null hypothesis is true, when the vet takes a sample of college students who have lost a pet, their mean should be close enough to 2.68 that the difference can be explained by sampling error.

The alternative hypothesis (H_1) says that what the veterinarian believes to be true is true: the population of students with pet loss has a mean GPA lower than the mean GPA of all students. Therefore, a mean for a sample of pet-loss students should be far enough below 2.68 that sampling error is not a plausible explanation for the difference.

Step 4 Set the Decision Rule. Dr. Richman has directional hypotheses and is using the default value of 5% for his willingness to make a Type I error. This means that it is a one-tailed test and that alpha is set at .05. The sample size is 11, and the vet will use Equation 7.1 to determine the degrees of freedom:

$$\begin{aligned} df &= N - 1 \\ &= 11 - 1 \\ &= 10 \end{aligned}$$

Using Appendix Table 3, Dr. Richman finds $t_{cv} = 1.812$. Because he is doing a one-tailed test, he has to decide whether the critical value of *t* is −1.812 or +1.812. If his theory is correct and losing a pet hurts students' GPAs, then the numerator of the single-sample *t* test formula (Equation 7.2), which is $M - \mu$, will be a negative number because M should be below 2.68. Hence, t_{cv} is a negative value, −1.812. Here, and in **Figure 7.5**, is the vet's decision rule:

- If $t \leq -1.812$, reject H_0.
- If $t > -1.812$, fail to reject H_0.

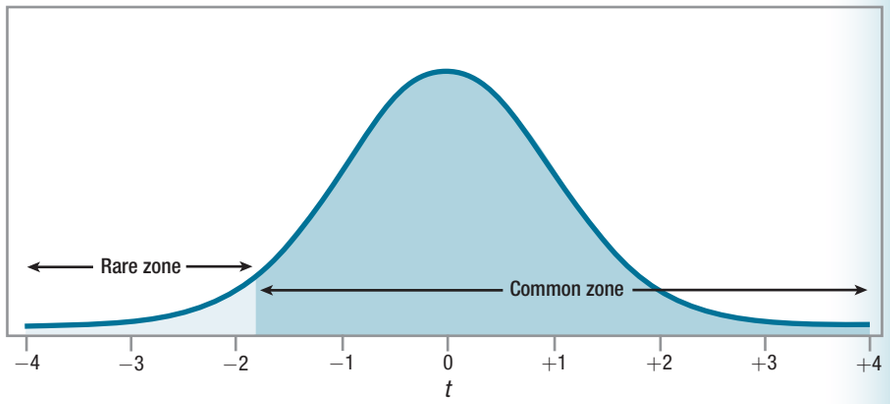


Figure 7.5 One-Tailed Decision Rule for Sampling Distribution of *t*, $df = 10$
This is a one-tailed test with 10 degrees of freedom. The critical value of t is −1.812. Note that the entire rare zone falls on one side of the sampling distribution.

Step 5 Calculate the Test Statistic. Before calculating *t,* the vet needs to calculate s_M using Equation 5.2:

$$\begin{aligned} s_M &= \frac{s}{\sqrt{N}} \\ &= \frac{0.50}{\sqrt{11}} \\ &= \frac{0.50}{3.3166} \\ &= 0.1508 \\ &= 0.15 \end{aligned}$$

Next, he'll use s_M in Equation 7.2 to calculate *t:*

$$\begin{aligned} t &= \frac{M - \mu}{s_M} \\ &= \frac{2.58 - 2.69}{0.15} \\ &= \frac{-0.1000}{0.15} \\ &= -0.6667 \\ &= -0.67 \end{aligned}$$

Step 5 is completed and the vet now knows the value of the test statistic: $t = -0.67$. In the next section, we'll learn how to explain what this means with regard to the impact of the loss of a pet on GPA.

Practice Problems 7.1

Review Your Knowledge

7.01 A researcher draws a sample from a population and finds that the sample mean is different from what he thought the population mean was. One explanation for the discrepancy is that he was wrong about what the population mean is. What's another explanation for the discrepancy?

7.02 What are the six steps to be followed in conducting a hypothesis test?

7.03 When should a single-sample *t* test be used?

7.04 What are the assumptions for a single-sample *t* test?

7.05 Which rare zone is larger if the observed difference is in the expected direction—for a one-tailed test or a two-tailed test?

Apply Your Knowledge

7.06 A researcher has a sample of Nobel Prize winners. She thinks that they may be smarter than average. If the average IQ is 100, what are the null and alternative hypotheses for a single-sample *t* test?

7.07 A researcher has drawn a random sample of 48 cases from a population. He plans to use a single-sample *t* test to compare the sample mean to the population mean. He has set α at .05 and is doing a two-tailed test. Write out the decision rule regarding the null hypothesis.

7.08 If $N = 17$ and $s = 6$, what is s_M?

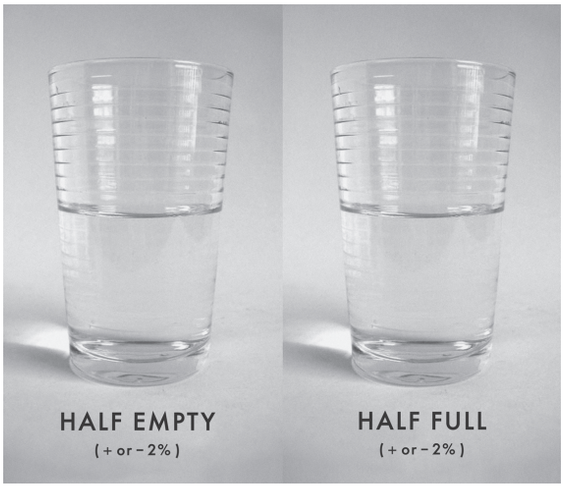
7.09 If $M = 24$, $\mu = 30$, and $s_M = 8$, what is t?

7.2 Interpreting the Single-Sample *t* Test

Computers are great for statistics. They can crunch numbers and do math faster and more accurately than humans can. But even a NASA supercomputer couldn't do what we are about to do: use common sense to explain the results of a single-sample *t* test in plain English. Interpretation is the most human part of statistics.

Interpretation is the most human part of statistics.

To some degree, interpretation is a subjective process. The researcher takes objective facts—like the means and the test statistic—and explains them in his or her own words. How one researcher interprets the results of a hypothesis test might differ from how another researcher interprets them. Reasonable people can disagree. But, there are guidelines that researchers need to follow. An interpretation needs to be supported by facts. One person may perceive a glass of water as half full and another as half empty, but they should agree that it contains roughly equal volumes of water and air.



Interpretation is subjective, but should be based on facts. Confidence intervals, introduced later in the chapter, provide a little wiggle room. (Photo courtesy of Paul Sahre.)

Interpreting the results of a hypothesis test starts with questions to be answered. The answers will provide the material to be used in a written interpretation. For a single-sample *t* test, our three questions are:

1. Was the null hypothesis rejected?
2. How big is the effect?
3. How wide is the confidence interval?

The questions are sequential. Each one gives new information that is useful in understanding the results from a different perspective. Answering only the first question, as was done in the last chapter, will provide enough information for completing a basic interpretation. However, answering all three questions leads to a more nuanced understanding of the results and a better interpretation.

Let's follow Dr. Farshad as she interprets the results of her study, question by question. But, first, let's review what she did. Dr. Farshad wondered if adults with ADHD differed in reaction time from the general population. To test this, she obtained a random sample, from her state, of 141 adults who had been diagnosed with ADHD. She had each participant take a reaction time test and found the sample mean and sample standard deviation ($M = 220$ msec, $s = 27$ msec). She planned to use a single-sample *t* test to compare the sample mean to the population mean for adult Americans ($\mu = 200$ msec).

Dr. Farshad's hypotheses for the single-sample *t* test were nondirectional as she was testing whether adults with ADHD had faster or slower reaction times than the general population. She was willing to make a Type I error no more than 5% of the time, she was doing a two-tailed test, and she had 140 degrees of freedom, so the critical value of *t* was ± 1.977. The first step in calculating *t* was to find the estimated standard error of the mean ($s_M = 2.21$) and she used that in order to go on and find $t = 8.81$.

Step 6 Interpret the Results

Was the Null Hypothesis Rejected?

In Dr. Farshad's study, the *t* value was calculated as 8.81. Now it is time to decide whether the null hypothesis was rejected or not. To do so, Dr. Farshad puts the *t* value, 8.81, into the decision rule that was generated for the critical value of *t*, ± 1.977, in Step 4. Which of the following statements is true?

- Is $8.81 \leq -1.977$ *or* is $8.81 \geq 1.977$?
- Is $-1.977 < 8.81 < 1.977$?

The second part of the first statement is true: 8.81 is greater than or equal to 1.977. The observed value of *t*, 8.81, falls in the rare zone of the sampling distribution, so the null hypothesis is rejected (see **Figure 7.6**). She can conclude that there is a statistically significant difference between the sample mean of the adults with ADHD and the general population mean.

Rejecting the null hypothesis means that Dr. Farshad has to accept the alternative hypothesis and conclude that the mean reaction time for adults with ADHD is *different* from the mean reaction time for the general population. Now she needs to determine the *direction* of the difference. By comparing the sample mean (220) to the population mean (200), she can conclude that adults with ADHD take a longer time to react and so have a *slower* reaction time than adults in general.

Dr. Farshad should also report the results in APA format. APA format provides five pieces of information: (1) what test was done, (2) the number of cases, (3) the value

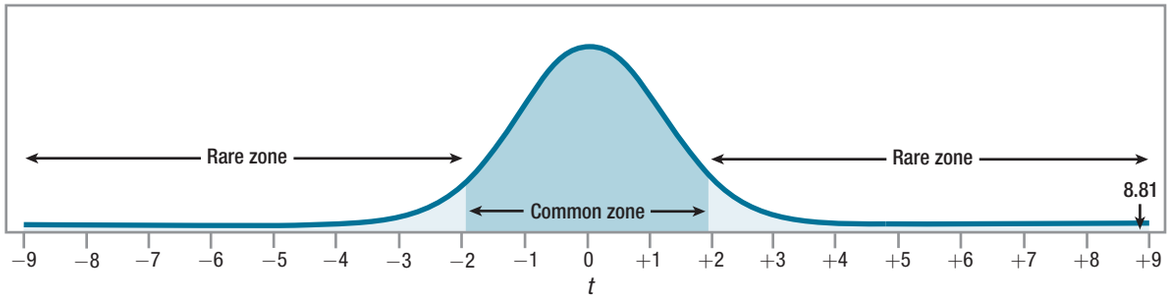


Figure 7.6 Single-Sample *t* Test: Results from ADHD Reaction-Time Study The value of the test statistic (t) is 8.81. This is greater than or equal to the critical value of t of 1.977, so it falls in the rare zone. The null hypothesis is rejected.

of the test statistic, (4) the alpha level used, and (5) whether the null hypothesis was or wasn't rejected.

In APA format, the results would be $t(140) = 8.81$, $p < .05$:

1. The initial t says that the statistical test was a t test.
2. The sample size, 141, is present in disguised form. The number 140, in parentheses, is the degrees of freedom for the t test. For a single-sample t test, $df = N - 1$. That means $N = df + 1$. So, if $df = 140$, $N = 140 + 1 = 141$.
3. The observed t value, 8.81, is reported. This number is the value of t calculated, *not* the critical value of t found in Appendix Table 3. Note that APA format requires the value to be reported to two decimal places, no more and no fewer.
4. The .05 tells that alpha was set at .05.
5. The final part, $p < .05$, reveals that the null hypothesis was rejected. It means that the observed result (8.81) is a rare result—it has a probability of less than .05 of occurring when the null hypothesis is true.

If Dr. Farshad stopped after answering only the first interpretation question, this is what she could write for an interpretation:

> There is a statistically significant difference between the mean reaction time of adults in Illinois with ADHD and the reaction time found in the general U.S. population, $t(140) = 8.81$, $p < .05$. Adults with ADHD ($M = 220$ msec) have a slower mean reaction time than does the general public ($\mu = 200$).

Practice Problems 7.2

Apply Your Knowledge

For these problems, use $\alpha = .05$, two-tailed.

7.10 If $N = 19$ and $t = 2.231$, write the results in APA format.

7.11 If $N = 7$ and $t = 2.309$, write the results in APA format.

7.12 If $N = 36$ and $t = 2.030$, write the results in APA format.

7.13 If $N = 340$ and $t = 3.678$, write the results in APA format.

How Big Is the Effect?

All we know so far is that there is a statistically significant difference, such that we can conclude that adults with ADHD in Illinois have slower reaction times than the general American adult population does. We know there's a 20-msec difference, but we are hard-pressed to know how much of a difference that really is.

Thus, the next question to address when interpreting the results is **effect size,** how large the impact of the explanatory variable is on the outcome variable. With the reaction-time study, the explanatory variable is ADHD status (adults with ADHD vs. the general population) and the outcome variable is reaction time. Asking what the effect size is, is asking how much impact ADHD has on a person's reaction time.

In this section, we cover two ways to measure effect size: Cohen's d and r^2. Both can be used to determine if the effect is small, medium, or large. Both standardize the effect size, so outcome variables measured on different metrics can be compared. A 2-point change in GPA would be huge, while a 2-point change in total SAT score would be trivial. Both Cohen's d and r^2 take the unit of measurement into account. And, finally, both will be used as measures of effect size for other statistical tests in other chapters.

Cohen's *d*

The formula for **Cohen's *d*** is shown in Equation 7.3. It takes the difference between the two means and standardizes it by dividing it by the sample standard deviation. Thus, d is like a z score—it is a standard score that allows different effects measured by different variables in different studies to be expressed—and compared—with a common unit of measurement.

Equation 7.3 Formula for Cohen's *d* for a Single-Sample *t* Test

$$d = \frac{M - \mu}{s}$$

where d = the effect size
M = sample mean
μ = hypothesized population mean
s = sample standard deviation

To calculate d, Dr. Farshad needs to know the sample mean, the population mean, and the sample standard deviation. For the reaction-time data, $M = 220$, $\mu = 200$, and $s = 27$. Here are her calculations for d:

$$\begin{aligned} d &= \frac{M - \mu}{s} \\ &= \frac{220 - 200}{27} \\ &= \frac{20.0000}{27} \\ &= 0.7407 \\ &= 0.74 \end{aligned}$$

Cohen's $d = 0.74$ for the ADHD reaction-time study. Follow APA format and use two decimal places when reporting *d*. And, because *d* values can be greater than 1, values of *d* from −0.99 to 0.99 get zeros before the decimal point.

In terms of the *size* of the effect, it doesn't matter whether *d* is positive or negative. A *d* value of 0.74 indicates that the two means are 0.74 standardized units apart. This is an equally strong effect whether it is 0.74 or −0.74. But the sign associated with *d* is important for knowing the direction of the difference, so make sure to keep it straight.

Here's how Cohen's *d* works:

- A value of 0 for Cohen's *d* means that the explanatory variable has absolutely no effect on the outcome variable.
- As Cohen's *d* gets farther away from zero, the size of the effect increases.

Cohen (1988), the developer of *d,* has offered standards for what small, medium, and large effect sizes are in the social and behavioral sciences. Cohen's *d* values for these are shown in **Table 7.3**. In **Figure 7.7**, the different effect sizes are illustrated by comparing the IQ of a control group ($M = 100$) to the IQ of an experimental group that is smarter by the amount of a small effect (3 IQ points), a medium effect (7.5 IQ points), or a large effect (12 IQ points).

TABLE 7.3 Effect Sizes in the Social and Behavioral Sciences

Size of effect	Cohen's *d*
None	≈0.00
Small	≈0.20
Medium	≈0.50
Large	>0.80

Note: Cohen's *d* is calculated with Equation 7.3. The sign of Cohen's *d* doesn't matter. A Cohen's *d* of −0.50 has the same degree of effect as a Cohen's *d* of 0.50, just in the opposite direction.

Cohen describes a small effect as a *d* of around 0.20. This occurs when there is a small difference between means, a difference in performance that would not be readily apparent in casual observation. Look at the top panel in Figure 7.7. That's what a small effect size looks like. Though the two groups differ by 3 points on intelligence ($M = 100$ vs. $M = 103$), it wouldn't be obvious that one group is smarter than the other without using a sensitive measure like an IQ test (Cohen, 1988).

Another way to visualize the size of the effect is to consider how much overlap exists between the two groups. If the curve for one group fits exactly over the curve for the other, then the two groups are exactly the same and there is no effect. As the amount of overlap decreases, the effect size increases. For the small effect in Figure 7.7, about 85% of the total area overlaps.

A medium effect size, a Cohen's *d* of around 0.50, is large enough to be observable, according to Cohen. Look at the middle panel in Figure 7.7. One group has a mean IQ of 100, and the other group has a mean IQ of 107.5, a 7.5 IQ point difference. Notice the amount of differentiation between groups for a medium effect size. In this example, about two-thirds of the total area overlaps, a decrease from the 85% overlap seen with a small effect size.

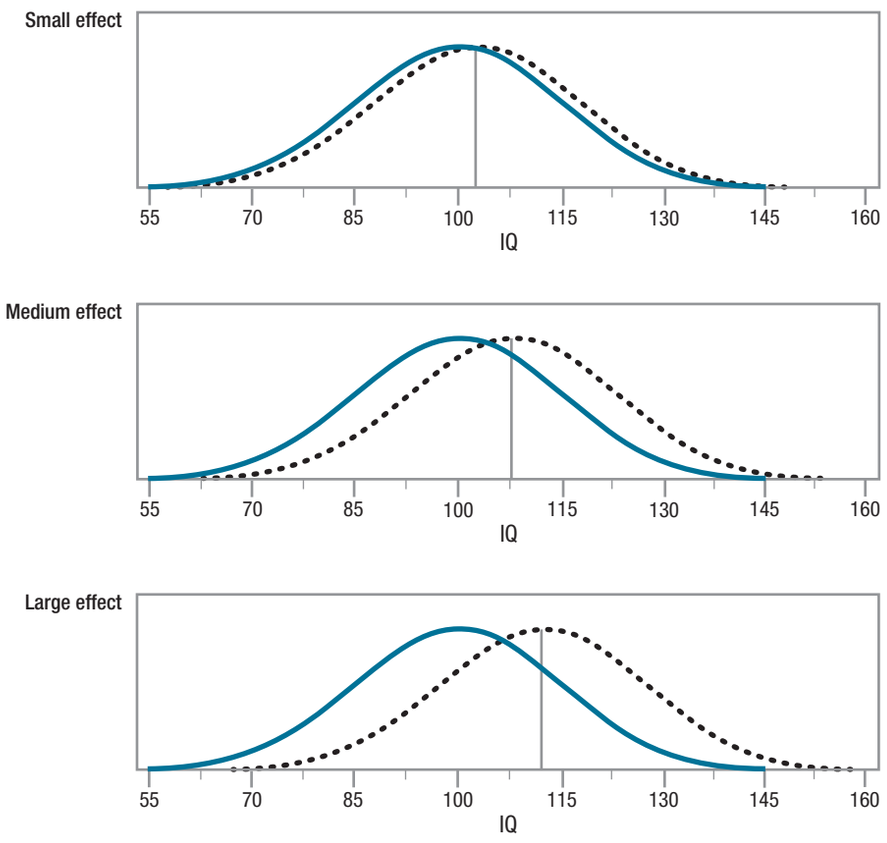


Figure 7.7 Examples of Small, Medium, and Large Effect Sizes This figure uses the distribution of IQ scores for two different groups to show effect sizes. In each panel, the solid-line curve shows the set of scores for the control group and the dotted-line curve shows the set of scores for the experimental group. In each panel, the experimental group has a higher mean IQ. The top panel (Small effect) shows what Cohen (1988) calls a small effect size ($d = 0.20$), the middle panel (Medium effect) a medium effect size ($d = 0.50$), and the bottom panel (Large effect) a large effect size ($d = 0.80$). These are mean differences, respectively, of 3, 7.5, and 12 IQ points. Notice how the differentiation between the two groups in each panel increases as the size of the effect increases, both in terms of increasing distance between the two means and decreasing overlap between the two distributions.

A large effect size is a Cohen's d of 0.80 or larger. This is shown in the bottom panel of Figure 7.7. One group has a mean IQ of 100, and the other group has a mean IQ of 112, a 12 IQ point difference. Here, there is a lot of differentiation between the two groups with only about half of the area overlapping. The increased differentiation can be seen in the large section of the control group that has IQs lower than the experimental group and the large section of the experimental group that has IQs higher than the control group.

The Cohen's d calculated for the reaction-time study by Dr. Farshad was 0.74, which means it is a medium to large effect. If she stopped her interpretation after calculating this effect size, she could add the following sentence to her interpretation: "The effect of ADHD status on reaction time falls in the medium to large range, suggesting that the slower mean reaction time associated with having ADHD may impair performance."

r Squared

The other commonly used measure of effect size is r^2. Its formal name is coefficient of determination, but everyone calls it *r* squared. $\boldsymbol{r^2}$**,** like *d*, tells how much impact the explanatory variable has on the outcome variable.

Equation 7.4 Formula for r^2, the Percentage of Variability in the Outcome Variable Accounted for by the Explanatory Variable

$$r^2 = \frac{t^2}{t^2 + df} \times 100$$

where r^2 = the percentage of variability in the outcome variable that is accounted for by the explanatory variable
t^2 = the squared value of *t* from Equation 7.2
df = the degrees of freedom for the *t* value

This formula says that r^2 is calculated as the squared *t* value divided by the sum of the squared *t* value plus the degrees of freedom for the *t* value. Then, to turn it into a percentage, the ratio is multiplied by 100. r^2 can range from 0% to 100%. These calculations reveal the percentage of variability in the outcome scores that is accounted for (or predicted) by the explanatory variable. For the ADHD data, these calculations would lead to the conclusion that $r^2 = 36\%$:

$$\begin{aligned} r^2 &= \frac{t^2}{t^2 + df} \times 100 \\ &= \frac{8.81^2}{8.81^2 + 140} \times 100 \\ &= \frac{77.6161}{77.6161 + 140} \times 100 \\ &= \frac{77.6161}{217.6161} \times 100 \\ &= .3567 \times 100 \\ &= 35.67\% \end{aligned}$$

r^2 tells the percentage of variability in the outcome variable that is accounted for (or predicted) by the explanatory variable.

What does r^2 tell us? Imagine that someone took a large sample of people and timed them running a mile. There would be a lot of variability in time, with some runners taking 5 minutes and others 20 minutes or more. What are some factors that influence running speed? Certainly, physical fitness and age play important roles, as do physical health and weight. Which of these four factors has the most influence on running speed? That could be determined by calculating r^2 for each variable. The factor that has the largest r^2, the one that explains the largest percentage of variability in time, has the most influence.

The question addressed by r^2 in the reaction-time study is how much of the variability in reaction time is accounted for by the explanatory variable (ADHD status) and how much is left unaccounted for:

- The closer r^2 is to 100%, the stronger the effect of the explanatory variable is and the less variability in the outcome variable remains to be explained by other variables.

- The closer r^2 is to 0%, the weaker the effect of the explanatory variable and the more variability in the dependent variable exists to be explained by other variables.

Cohen (1988), who provided standards for *d* values, also provides standards for r^2:

- A small effect is an $r^2 \approx 1\%$.
- A medium effect is an $r^2 \approx 9\%$.
- A large effect is an $r^2 \approx 25\%$.

By Cohen's standards, an r^2 of 36%, as is the case in our ADHD study, is a large effect. This means that as far as explanatory variables in the social and behavioral sciences go, this one accounts for a lot of the variability in the outcome variable. **Figure 7.8** is a visual demonstration of what explaining 36% of the variability means. Note that although a majority of the variability, 64%, in reaction time remains unaccounted for, this is still considered a large effect. Here is what Dr. Farshad could add to her interpretation now that r^2 is known: "Having ADHD has a large effect on one's reaction time. In the present study, knowing ADHD status explains more than a third of the variability in reaction time."

One important thing to note is that even though Dr. Farshad went to the trouble of calculating both *d* and r^2, it is not appropriate to report both. These two provide overlapping information and it does not make a case stronger to say that the effect was a really strong one, because both *d* and r^2 are large. This would be like testing boiling water and reporting that it was really, really boiling because it registered 212 degrees on a Fahrenheit thermometer and 100 degrees on a Celsius thermometer.

Here's a heads up for future chapters and for reading results sections in psychology articles—there are other measures of effect size that are similar to r^2. Both η^2 (eta squared) and ω^2 (omega squared) provide the same information, how much of the variability in the outcome variable is explained by the explanatory variable.

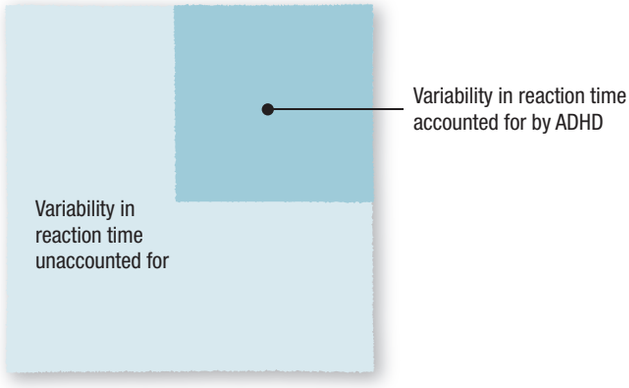


Figure 7.8 Percentage of Variability in Reaction Time Accounted for by ADHD Status The darkly shaded section of this square represents the 36% of the variability in reaction time that is accounted for by ADHD status. The nonshaded part of the square represents the 64% of variability that is still not accounted for.

Practice Problems 7.3

Apply Your Knowledge

7.14 If $M = 66$, $\mu = 50$, and $s = 10$, what is d?

7.15 If $M = 93$, $\mu = 100$, and $s = 30$, what is d?

7.16 If $N = 18$ and $t = 2.37$, what is r^2?

7.17 If $r = .32$, how much of the variability in the outcome variable, Y, is accounted for by the explanatory variable, X?

How Wide Is the Confidence Interval?

In Chapter 5, we encountered confidence intervals for the first time. There, they were used to take a sample mean, M, and estimate a range of values that was likely to capture the population mean, μ. Now, for the single-sample t test, the confidence interval will be used to calculate the range for how big the difference is between two population means. Two populations? Dr. Farshad's confidence interval, in the ADHD reaction-time study, will tell how large (or small) the mean difference in reaction time might be between the *population of adults with ADHD* and the general population of Americans.

The difference between population means can be thought of as an effect size. If the distance between two population means is small, then the size of the effect (i.e., the impact of ADHD on reaction time) is not great. If the two population means are far apart, then the size of the effect is large.

Equation 7.5 is the formula for calculating a confidence interval for the difference between two population means. It calculates the most commonly used confidence interval, the 95% confidence interval.

Equation 7.5 95% Confidence Interval for the Difference Between Population Means

$$95\%\ CI_{\mu_{Diff}} = (M - \mu) \pm (t_{cv} \times s_M)$$

where $95\%\ CI_{\mu_{Diff}}$ = the 95% confidence interval for the difference between two population means
M = sample mean from one population
μ = mean for other population
t_{cv} = critical value of t, two-tailed, $\alpha = .05$, $df = N - 1$ (Appendix Table 3)
s_M = estimated standard error of the mean (Equation 5.2)

Here are all the numbers Dr. Farshad needs to calculate the 95% confidence interval:

- M, the mean for the sample of adults with ADHD, is 220.
- μ, the population mean for adults in general, is 200.
- $s_M = 2.27$, a value obtained earlier via Equation 5.2 for use in Equation 7.2.
- t_{cv}, the critical value of t, is 1.977.

Applying Equation 7.5, she would calculate

$$\begin{aligned} 95\%\ \text{CI}_{\mu_{\text{Diff}}} &= (M - \mu) \pm (t_{cv} \times s_M) \\ &= (220 - 200) \pm (1.977 \times 2.27) \\ &= 20.0000 \pm 4.4878 \\ &= \text{from } 15.5122 \text{ to } 24.4878 \\ &= \text{from } 15.51 \text{ to } 24.49 \end{aligned}$$

The 95% confidence interval for the difference between population means ranges from 15.51 msec to 24.49 msec. In APA format, this confidence interval would be reported as 95% CI [15.51, 24.49].

What information does a confidence interval provide? The technical definition is that if one drew sample after sample and calculated a 95% confidence interval for each one, 95% of the confidence intervals would capture the population value. But, that's not very useful as an interpretative statement. Two confidence interval experts have offered their thoughts on interpreting confidence intervals (Cumming and Finch, 2005). With a 95% confidence interval, one interpretation is that a researcher can be 95% confident that the population value falls within the interval. Another way of saying this is that the confidence interval gives a range of *plausible* values for the population value. That means it is possible, but unlikely, that the population value falls outside of the confidence interval. A final implication is that the ends of the confidence interval are reasonable estimates of how large or how small the population value is. The end of the confidence interval closer to zero can be taken as an estimate of how small the population value might be, while the end of the confidence interval further from zero estimates how large the population value might be.

Dr. Farshad's confidence interval tells her that there's a 95% chance that the difference between the two population means falls somewhere in the interval from 15.51 msec to 24.49 msec. It means that the best prediction of how much slower reaction times are for the population of adults with ADHD than for the general population of Americans ranges from 15.51 msec slower to 24.49 msec slower (**Figure 7.9**).

Our interpretation of confidence intervals will focus on three aspects: (1) whether the value of 0 falls within the confidence interval; (2) how close the confidence interval comes to zero or how far from zero it goes, and (3) how wide the confidence interval is. These three aspects are explained below and summarized in **Table 7.4**.

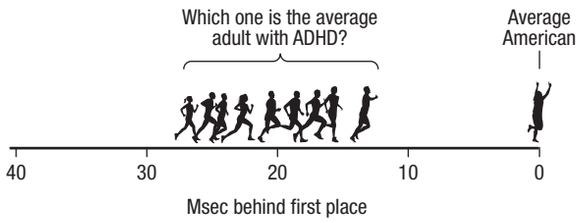


Figure 7.9 95% Confidence Interval for the Difference Between Population Means for the Reaction-Time Study Imagine a race in which a person representing the average American crosses the finish line first, followed by a person representing the average adult with ADHD. How much slower is the average adult with ADHD? The confidence interval says that the average adult with ADHD *probably* trails the average American by anywhere from 15.51 msec to 24.49 msec, but it doesn't specify where in that range the average adult with ADHD is.

TABLE 7.4 How to Choose: Interpreting a Confidence Interval for the Difference Between Population Means

	Confidence Interval Captures Zero	Confidence Interval Is Near Zero	Confidence Interval Is Far from Zero
Confidence Interval Is Narrow	There is not enough evidence to conclude an effect exists. A researcher can't say the two population means are different.	The effect is likely weak. It is plausible that the two population means are different. Cohen's *d* or r^2 represents the size of the effect.	The effect is likely strong. It is plausible that the two population means are different. Cohen's *d* or r^2 represents the size of the effect.
Confidence Interval Is Wide	There is not enough evidence to conclude an effect exists. There is little information about whether the two population means are different. Replicate the study with a larger sample.	The effect is likely weak to moderate. It is plausible that the two population means are different. Calculate Cohen's *d* for both ends of the CI. Replicate the study with a larger sample.	The effect is likely moderate to strong. It is plausible that the two population means are different. Calculate Cohen's *d* for both ends of the CI. Replicate the study with a larger sample.

Note: A narrow confidence interval gives a precise estimate of the size of the difference between population means. A wide confidence interval doesn't provide much guidance on whether the difference between population means is large or small. Whether the confidence interval is "near zero" or falls "far from zero" provides information regarding how large the effect might be in the population.

1. Does the confidence interval capture zero? The 95% confidence interval for the difference between population means gives the range within which it is likely that the actual difference between two population means lies. If the null hypothesis is rejected for a single-sample *t* test, the conclusion reached is that this sample probably did not come from that population; rather, that there are two populations with two different means. In such a situation, the difference between the two population means is not thought to be zero, and the confidence interval shouldn't include zero. When the null hypothesis is rejected, the 95% confidence interval won't capture zero. (This assumes that the researcher was using a two-tailed test with $\alpha = .05$.)

   However, when a researcher fails to reject the null hypothesis, the confidence interval should include zero in its range. All that a zero falling in the range of the confidence interval means is that it is *possible* the difference between the two population means is zero. It doesn't mean that the difference *is* zero, just that it may be. In a similar fashion, when one fails to reject the null hypothesis, the conclusion is that there isn't enough evidence to say the null hypothesis is wrong. It may be right, it may be wrong; the researcher just can't say.

   Having already determined earlier that the null hypothesis was rejected and the results were statistically significant, Dr. Farshad knew that her confidence interval wouldn't capture zero. And, as the interval ranged from 15.51 to 24.49, she was right.

2. How close does the confidence interval come to zero? How far away from zero does it go? If the confidence interval doesn't capture zero, how close one end of the confidence interval comes to zero can be thought of as providing information about how weak the effect may be. If one end of the confidence interval is very close to zero, then the effect size may be small, as there could be little difference between the population means.

   Similarly, how far away the other end of the confidence interval is from zero can be thought of as providing information about how strong the effect may be. The farther away from zero the confidence interval ranges, the larger the effect size may be.

3. How wide is the confidence interval? The width of the confidence interval also provides information. Narrower confidence intervals provide a more precise estimate of the population value and are more useful.

   When the confidence interval is wide, the size of the effect in the population is uncertain. It might be a small effect, a large effect, or anywhere in between. In these instances, replication of the study with a larger sample size will yield a narrower confidence interval.

   When evaluating the width of a confidence interval, one should take into account the variable being measured. A confidence interval that is 1 point wide would be wide if the variable were GPA, but narrow if the variable were SAT. Interpretation relies on our human common sense.

   Dr. Farshad's confidence interval ranges from 15.51 to 24.49. The width can be determined by subtracting the lower limit from the upper limit, as shown in Equation 7.6.

Equation 7.6 Formula for Calculating the Width of a Confidence Interval

$$CI_W = CI_{UL} - CI_{LL}$$

where CI_W = the width of the confidence interval
CI_{UL} = the upper limit of the confidence interval
CI_{LL} = the lower limit of the confidence interval

Applying Equation 7.6 to her data, Dr. Farshad calculates $24.49 - 15.51 = 8.98$. The confidence interval is almost 9 msec wide. Is this wide or narrow?

Evaluating the width of a confidence interval often requires expertise. It takes a reaction-time researcher like Dr. Farshad to tell us whether a 9-msec range for a confidence interval is a narrow range and sufficiently precise. Here, given the width of the confidence interval, she feels little need to replicate the study with a larger sample size to narrow the confidence interval.

Putting It All Together

Dr. Farshad has completed her study using a single-sample *t* test to see if adults with ADHD had a reaction time that differed from that for the general population. By addressing the three questions in Table 7.5, Dr. Farshad has gathered all the pieces she needs to write an interpretation. There are four points she addresses in her interpretation:

1. Dr. Farshad starts with a brief explanation of the study.
2. She presents some facts, such as the means of the sample and the population. But, she does not report all the values she calculated just because she calculated them. Rather, she is selective and only reports what she believes is most relevant.
3. She explains what she believes the results mean.
4. Finally, she offers some suggestions for future research. Having been involved in the study from start to finish, she knows its strengths and weaknesses better than anyone else. She is in a perfect position to offer advice to other researchers about ways to redress the limitations of her study. In her suggestions, she uses the word "replicate." To **replicate** a study is to repeat it, usually introducing some change in procedure to make it better.

TABLE 7.5 Three Questions for Interpreting a Single-Sample *t* Test

1. Was the null hypothesis rejected?
   - Decide by comparing the calculated value of t to the critical value of t, t_{cv}, using the decision rule generated in Step 4.
   - Was H_0 rejected?
     - If yes, (1) call the results statistically significant, and (2) compare M to μ to determine the direction of the difference.
     - If no, (1) say the results are not statistically significant, and (2) conclude there is not enough evidence to say a difference exists.
   - Report the results in APA format:
     - If H_0 is rejected, report the results as "$p < .05$."
     - If H_0 is not rejected, report the results as "$p > .05$."
2. How big is the effect?
   - Calculate Cohen's d (Equation 7.3) or r^2 (Equation 7.4).
     - No effect: $d = 0.00$; $r^2 = 0$.
     - Small effect: $d = 0.20$; $r^2 = 1\%$.
     - Medium effect: $d = 0.50$; $r^2 = 9\%$.
     - Large effect: $d \geq 0.80$; $r^2 \geq 25\%$.
3. How wide is the confidence interval?
   - Calculate the 95% confidence interval for the difference between population means (Equation 7.5).
     - Interpret the confidence interval based on (1) whether it captures zero, (2) how close to/far from zero it comes, and (3) how wide it is. (See Table 7.4.)

Here is Dr. Farshad's interpretation:

A study compared the reaction time of a random sample of Illinois adults who had been diagnosed with ADHD ($M = 220$ msec) to the known reaction time for the American population ($\mu = 200$ msec). The reaction time of adults with ADHD was statistically significantly slower than the reaction time found in the general population [$t(140) = 8.81$, $p < .05$]. The size of the difference in the larger population probably ranges on this task from a 16- to a 24-msec decrement in performance. This is not a small difference—these results suggest that ADHD in adults is associated with a medium to large level of impairment on tasks that require fast reactions. If one were to replicate this study, it would be advisable to obtain a broader sample of adults with ADHD, not just limiting it to one state. This would increase the generalizability of the results.

Worked Example 7.2

For practice in interpreting the results of a single-sample *t* test when the null hypothesis is not rejected, let's return to Dr. Richman's study.

In his study, Dr. Richman studied the effect of losing a pet during the year on college performance. Dr. Richman located 11 students who had lost a pet. Their mean GPA for the year was 2.58 ($s = 0.50$) compared to a population mean of 2.68. Because he expected pet loss to have a negative effect, Dr. Richman used a one-tailed test. With the alpha set at .05, t_{cv} was -1.812. He calculated $s_M = 0.15$ and found $t = -0.67$. Now it is time for Dr. Richman to interpret the results.

Was the null hypothesis rejected? The vet was doing a one-tailed test and his hypotheses were

$$H_0: \mu_{\text{StudentsWithPetLoss}} \geq 2.68$$
$$H_1: \mu_{\text{StudentsWithPetLoss}} < 2.68$$

Inserting the observed value of *t*, −0.67, into the decision rule he had generated in Step 4, he has to decide which statement is true:

Is $-0.67 \leq -1.812$? If so, reject H_0.
Is $-0.67 > -1.812$? If so, fail to reject H_0.

As the second statement is true, Dr. Richman has failed to reject the null hypothesis. **Figure 7.10** shows how the *t* value of −0.67 falls in the common zone of the sampling distribution.

Having failed to reject the null hypothesis, the results are called not statistically significant. Just like finding a defendant not guilty doesn't mean that the defendant is innocent, failing to reject the null hypothesis does not mean that it is true. All the vet can conclude is that there's not enough evidence to conclude that college students who lose a pet do worse in school that year. Because he failed to reject the null hypothesis, there's no reason to believe a difference exists between the two populations, and there's no reason for him to worry about the direction of the difference between them.

In APA format, the results would be written as $t(10) = -0.67$, $p > .05$ (one-tailed):

- The *t* tells what statistical test was done.
- The 10 in parentheses, which is the degrees of freedom, reveals there were 11 cases as $N = df + 1$ for a single-sample *t* test.
- −0.67 is the observed value of the statistic.
- The .05 indicates that this was the alpha level selected.
- $p > .05$ indicates that the researcher failed to reject the null hypothesis. The observed *t* value is a common one when the null hypothesis is true. "Common" is defined as occurring more than 5% of the time.

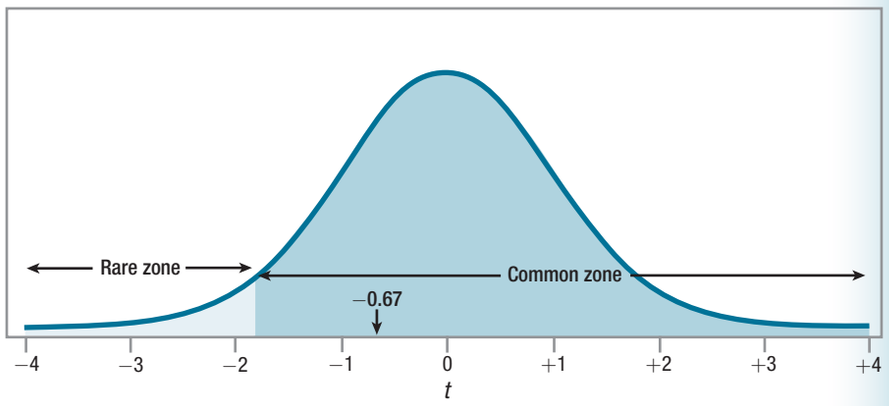


Figure 7.10 Failing to Reject the Null Hypothesis, One-Tailed Test The critical value of *t* (one-tailed, $\alpha = .05$, $df = 10$) is −1.812. This means that the value of the test statistic, −0.67, falls in the common zone, and the null hypothesis is not rejected.

- The final parenthetical expression (one-tailed) is something new. It indicates, obviously, that the test was a one-tailed test. Unless told that the test is a one-tailed test, assume it is two-tailed.

How big is the effect? Calculating an effect size like Cohen's *d* or r^2 when one has failed to reject the null hypothesis is a controversial topic in statistics. If one fails to reject the null hypothesis, not enough evidence exists to state there's an effect. Some researchers say that there is no need to calculate an effect size if no evidence that an effect exists is found. Other researchers believe that examining the effect size when one has failed to reject the null hypothesis alerts the researcher to the possibility of a Type II error (Cohen, 1994; Wilkinson, 1999). (Remember, a Type II error occurs when the null hypothesis should be rejected, but it isn't.) In this book, we're going to side with calculating an effect size when one fails to reject the null hypothesis as doing so gives additional information useful in understanding the results.

In the pet-loss study, the mean GPA for the 11 students who had lost a pet was 2.58 ($s = 0.50$) and the population mean for all first-year students was 2.68. Applying Equation 7.3, here are the calculations for *d:*

$$\begin{aligned} d &= \frac{M - \mu}{s} \\ &= \frac{2.58 - 2.68}{0.50} \\ &= \frac{-0.1000}{0.50} \\ &= -0.2000 \\ &= -0.20 \end{aligned}$$

And, applying Equation 7.4, here are the calculations for r^2:

$$\begin{aligned} r^2 &= \frac{t^2}{t^2 + df} \times 100 \\ &= \frac{-0.67^2}{-0.67^2 + 10} \times 100 \\ &= \frac{0.4489}{0.4489 + 10} \times 100 \\ &= \frac{0.4489}{10.4489} \times 100 \\ &= 0.0430 \times 100 \\ &= 4.30 \end{aligned}$$

The effect size *d*, −0.20, is a small effect according to Cohen, while r^2 of 4.30% would be classified as a small to medium effect. Why, when we fail to reject the null hypothesis, does there seem to be some effect? There are two explanations for this:

1. Even if the null hypothesis is true, in which case the size of the effect in the population is zero, it is possible, due to sampling error, that a nonzero effect would be observed in the sample. So, it shouldn't be surprising to find a small effect.

2. It is also possible that the observed effect size represents the size of the effect in the population. If this were true, it would mean a Type II error is being made.

When a researcher has concern about Type II error, he or she can suggest replication with a larger sample size. A larger sample size has a number of benefits. First, it makes it easier to reject the null hypothesis because a larger sample size, with more degrees of freedom, moves the rare zone closer to the midpoint. It is also easier to reject the null hypothesis because a larger sample makes the standard error of the mean smaller, which makes the *t* value larger. In fact, if the sample size had been 75 in the pet-loss study, not 11, *d* would still be −0.20, but the vet would have rejected the null hypothesis.

A statistician would call this study underpowered. To be **underpowered** means that a study doesn't include enough cases to have a reasonable chance of detecting an effect that exists. As was mentioned in Chapter 6, if power is low, then beta, the chance of a Type II error, is high. So, raising a concern about one (Type II error) raises a concern about the other (power).

How wide is the confidence interval? The final calculation Dr. Richman should make in order to understand his results is to calculate a 95% confidence interval for the difference between population means. This requires Equation 7.5:

$$\begin{aligned} 95\%\ CI_{\mu_{Diff}} &= (M - \mu) \pm (t_{cv} \times s_M) \\ &= (2.58 - 2.68) \pm (2.228 \times 0.15) \\ &= -0.1000 \pm 0.3342 \\ &= \text{from } -0.4342 \text{ to } 0.2342 \\ &= \text{from } -0.43 \text{ to } 0.23 \end{aligned}$$

Before interpreting this confidence interval, remember that the null hypothesis was not rejected and there is no evidence that pet loss has an effect on GPA. The confidence interval should show this, and it does.

This 95% confidence interval ranges from −0.43 to 0.23. First, note that it includes the value of 0. A value of 0 indicates no difference between the two population means and so it is a *possibility,* like the null hypothesis said, that there is no difference between the two populations. The confidence interval for the difference between population means should capture zero whenever the null hypothesis is not rejected. (This is true for a 95% confidence interval as long as the hypothesis test is two-tailed with $\alpha = .05$.)

Next, Dr. Richman looks at the upper and lower ends of the confidence interval, −0.43 and 0.23. This confidence interval tells him that it is possible the average GPA for the population of students who have lost a pet could be as low as 0.43 points worse than the general student average, or as high as 0.23 points better than the general student average. This doesn't provide much information about the effect of pet loss on academic performance—maybe it helps, maybe it hurts.

Finally, he looks at the width of the confidence interval using Equation 7.6:

$$\begin{aligned} CI_W &= CI_{UL} - CI_{LL} \\ &= 0.23 - (-0.43) \\ &= 0.23 + 0.43 \\ &= 0.6600 \\ &= 0.66 \end{aligned}$$

The confidence interval is 0.66 points wide. This is a fairly wide confidence interval for a variable like GPA, a variable that has a maximum range of 4 points.

A confidence interval is better when it is narrower, because a narrower one gives a more precise estimate of the population value. How does a researcher narrow a confidence interval? By increasing sample size. Dr. Richman should recommend replicating with a larger sample size to get a better sense of what the effect of pet loss is in the larger population. Increasing the sample size also increases the power, making it more likely that an effect will be found statistically significant.

Putting it all together. Here's Dr. Richman's interpretation. Note that (1) he starts with a brief explanation of the study, (2) reports some facts but doesn't report everything he calculated, (3) gives his interpretation of the results, and (4) makes suggestions for improving the study.

> This study explored whether losing a pet while at college had a negative impact on academic performance. The GPA of 11 students who lost a pet ($M = 2.58$) was compared to the GPA of the population of students at that college ($\mu = 2.68$). The students who had lost a pet did not have GPAs that were statistically significantly lower [$t(10) = -0.67$, $p > .05$ (one-tailed)].
>
> Though there was not sufficient evidence in this study to show that loss of a pet has a negative impact on college performance, the sample size was small and the study did not have enough power to find a small effect. Therefore, it would be advisable to replicate the study with a larger sample size to have a better chance of determining the effect of pet loss and to get a better estimate of the size of the effect.

Practice Problems 7.4

Apply Your Knowledge

7.18 If $M = 70$, $\mu = 60$, $t_{cv} = \pm 2.086$, and $s_M = 4.36$, what is the 95% confidence interval for the difference between population means?

7.19 If $M = 55$, $\mu = 50$, $t_{cv} = \pm 2.052$, and $s_M = 1.89$, what is the 95% confidence interval for the difference between population means?

7.20 A college president has obtained a sample of 81 students at her school. She plans to survey them regarding some potential changes in academic policies. But, first, she wants to make sure that the sample is representative of the school in terms of academic performance. She knows from the registrar that the mean GPA for the entire school is 3.02. She calculates the mean of her sample as 3.16, with a standard deviation of 0.36. She used a single-sample *t* test to compare the sample mean (3.16) to the population mean (3.02). Using her findings below, determine if the sample is representative of the population in terms of academic performance:

- $s_M = 0.04$
- $t = 3.50$
- $d = 0.39$
- $r^2 = 13.28\%$
- 95% CI$_{\mu_{Diff}}$ = from 0.06 to 0.22

Application Demonstration

Chips Ahoy is a great cookie. In 1997 Nabisco came out with a clever advertising campaign, the Chips Ahoy Challenge. Its cookies had so many chocolate chips that Nabisco guaranteed there were more than a thousand chips in every bag. And

the company challenged consumers to count. It's almost two decade later, but do Chips Ahoy cookies still have more than a thousand chocolate chips in every bag? Let's investigate.

It is easier to count the number of chips in a single cookie than to count the number of chips in a whole bag of cookies. If a bag has at least 1,000 chips and there are 36 cookies in a bag, then each cookie should have an average of 27.78 chips. So, here is the challenge rephrased: "Chips Ahoy cookies are so full of chocolate chips that each one contains an average of 27.78 chips. We challenge you to count them."

Step 1 This challenge calls for a statistical test. The population value is known: $\mu = 27.78$. All that is left is to obtain a sample of cookies; find the sample mean, *M;* and see if there is a statistically significant difference between M and μ. Unfortunately, Nabisco has not reported σ, the population standard deviation, so the single-sample z test can't be used. However, it is possible to calculate the standard deviation (s) from a sample, which means the single-sample t test can be used.

Ten cookies were taken from a bag of Chips Ahoy cookies. Each cookie was soaked in cold water, the cookie part washed away, and the chips counted. The cookies contained from 22 to 29 chips, with a mean of 26.10 and a standard deviation of 2.69.

Step 2 The next step is to check the assumptions. The random sample assumption was violated as the sample was not a random sample from the population of Chips Ahoy cookies being manufactured. Perhaps the cookies being manufactured on the day this bag was produced were odd. Given the focus on quality control that a manufacturer like Nabisco maintains, this seems unlikely. So, though this assumption is violated, it still seems reasonable to generalize the results to the larger population of Chips Ahoy cookies.

We'll operationalize the independence of observations assumption as most researchers do and consider it not violated as no case is in the sample twice and each cookie is assessed individually. Similarly, the normality assumption will be considered not violated under the belief that characteristics controlled by random processes are normally distributed.

Step 3 The hypotheses should be formulated before any data are collected. It seems reasonable to do a nondirectional (two-tailed) test in order to allow for the possibilities of more chips than promised as well as fewer chips than promised. Here are the two hypotheses:

$$H_0: \mu_{\text{Chips}} = 27.28$$

$$H_1: \mu_{\text{Chips}} \neq 27.78$$

Step 4 The consequences of making a Type I error in this study are not catastrophic, so it seems reasonable to use the traditional alpha level of .05. As already decided, a two-tailed test is being used. Thus, the critical value of *t,* with 9 degrees of freedom, is ± 2.262. Here is the decision rule:

$$\text{If } t \leq -2.262 \textit{ or if } t \geq 2.262, \text{ reject } H_0.$$

$$\text{If } -2.262 < t < 2.262, \text{ fail to reject } H_0.$$

Step 5 The first step in calculating a single-sample *t* test is to calculate what will be used as the denominator, the estimated standard error of the mean:

$$s_M = \frac{s}{\sqrt{N}} = \frac{2.69}{\sqrt{10}} = 0.85$$

The estimated standard error of the mean is then used to calculate the test statistic, *t:*

$$t = \frac{M - \mu}{s_M} = \frac{26.10 - 27.78}{0.85} = -1.98$$

Step 6 The first step in interpreting the results is to determine if the null hypothesis should be rejected. The second statement in the decision rule is true: $-2.262 < -1.98 < 2.262$, the null hypothesis is not rejected. The difference between 26.10, the mean number of chips found in the cookies, and 27.78, the number of chips expected per cookie, is not statistically significant. From this study, even though the cookies in the bag at $M = 26.10$ chips per cookie fell short of the expected 27.78 chips per cookie, not enough evidence exists to question Nabisco's assertion that there are 27.78 chips per cookie and a thousand chips in every bag. And, there's not enough evidence to suggest that, since 1997, Nabisco's recipe has changed and/or its quality control has slipped.

Having answered the challenge, it is tempting to stop here. But forging ahead—going on to find Cohen's *d*, r^2, and the 95% confidence interval—will give additional information and help clarify the results.

Finding Cohen's *d,* the size of the effect, provides a different perspective on the results:

$$\begin{aligned} d &= \frac{M - \mu}{s} \\ &= \frac{26.10 - 27.78}{2.69} \\ &= \frac{-1,6800}{2.69} \\ &= -0.6245 \\ &= -0.62 \end{aligned}$$

Though the *t* test says there is not enough evidence to find an effect, Cohen's *d* says $d = -0.62$, a medium effect. This seems contradictory—is there an effect or isn't there? Perhaps the study doesn't have enough power and a Type II error is being made. This means that the study should be replicated with a larger sample size before concluding that Chips Ahoy hasn't failed the thousand chip challenge. Calculating $r^2 = 30\%$ leads to a similar conclusion.

Does the confidence interval tell a similar story? Here are the calculations for the confidence interval:

$$\begin{aligned} 95\%\ \text{CI}_{\mu_{\text{Diff}}} &= (M - \mu) \pm (t_{cv} \times s_M) \\ &= (26.10 - 27.78) \pm (2.262 \times 0.85) \\ &= -1.6800 \pm 1.9227 \\ &= \text{from } -3.6027 \text{ to } 0.2427 \\ &= \text{from } -3.60 \text{ to } 0.24 \end{aligned}$$

The confidence interval says that the range from −3.60 to 0.24 probably contains the difference between 27.78 (the expected number of chips per cookie) and the number really found in each cookie. As this range captures zero, it is possible that there is no difference, the null hypothesis is true, and bags of cookies really do contain 1,000 chips.

Positive numbers in the confidence interval suggest that the difference lies in the direction of there being more than 27.78 chips per cookie. Negative numbers suggest the difference is in the direction of there being fewer than 27.78 chips per cookie. Most of the confidence interval lies in negative territory. A betting person would wager that the difference is more likely to rest with bags containing fewer than a thousand chips.

The *t* test left open the possibility that Nabisco still ruled the 1,000 chip challenge. But, thanks to going beyond *t,* to calculating *d* and a confidence interval, it no longer seems clear that Nabisco would win the Chips Ahoy challenge. For a complete interpretation of the results of this Chips Ahoy challenge, see **Figure 7.11**.

PENNSTATE

Erie The Behrend College

School of Humanities and Social Sciences
Penn State Erie, The Behrend College
170 Irvin Kochel Center
4951 College Drive
Erie, PA 16563-1501

814-898-6108
Fax: 814-898-6032
behrend.psu.edu

Irene Rosenfeld
Chairman and Chief Executive Officer, Kraft Foods
3 Lakes Drive
Northfield, IL 60093

June 1, 2012

Dear Ms. Rosenfeld:

I am writing to you because Kraft Foods owns Nabisco, the maker of Chips Ahoy cookies. My letter contains both good news and bad news

Back in 1997, Nabisco issued the Chips Ahoy Challenge. They said that each bag of Chips Ahoy cookies contained at least 1,000 chips and they dared consumers to count.

I'm a college professor, working on a revision of a statistics textbook, and I decided to use this challenge as an application of a single-sample *t* test. I was curious if 15 years later there were still 1,000 chips per bag.

I bought a bag of Chips Ahoy cookies, found that it contained 36 cookies, and calculated that if each cookie contained an average of 27.78 chips, the bag would contain 1,000 chips. I then selected ten cookies, dissolved them in water, and counted the number of chips in each. I found a mean of 26.10 chips with a standard deviation of 2.69.

First, the good news. Using a single-sample *t* test, I found that 26.10 was not statistically, significantly, different from 27.78. In plain language, this means that, based on this study, it's plausible that a bag of Chips Ahoy cookies contains 1,000 chips.

Unfortunately, bad news follows the good. By calculating something called an effect size and something called a confidence interval, I became concerned that my conclusion that the difference was not statistically significant was erroneous. My study was what statisticians call "underpowered." It didn't have enough cookies in the sample to have a fair chance of finding a difference if there were a difference. With more cookies in my sample, I probably would have found that it is unlikely that bags contains 1,000 chips.

I plan to replicate the study with a larger sample size.

But, until I do, I am curious – do Kraft and Nabisco still stand behind the Chips Ahoy Challenge?

Sincerely,

Eric W. Corty, Ph.D.
Professor of Psychology

Figure 7.11 Interpretation of Results of the Chips Ahoy Challenge

SUMMARY

Choose when to use a single-sample *t* test.

- A single-sample *t* test is used to compare a sample mean to some specified value, like a population mean, when the population standard deviation is not known. Like all statistical tests, it has assumptions that must be met, hypotheses to be listed, and a decision rule to be set in advance of calculating the value of the test statistic.

Calculate the test statistic for a single-sample *t* test.

- The single-sample *t* test calculates a *t* value. *t* is distributed very much like *z*—if the null hypothesis for a two-tailed test is true, the *t* distribution is symmetrical and centered on zero. The *t* value is based on the size of the difference between the sample mean and the specified value. If the difference, when standardized as a *t* score, is too large to be explained by sampling error, the null hypothesis is rejected.

Interpret the results of a single-sample *t* test.

- Interpretation explains the results of a statistical test in plain language. An interpretation is subjective, but it is based on facts. Interpretation proceeds by asking questions of the data (e.g., was the null hypothesis rejected, how big is the effect) and then using the answers to address four points: (1) what was the study about, (2) what were the results, (3) what do the results mean, and (4) what should be done in future research.
- Cohen's *d* and r^2 were introduced in this chapter and confidence intervals made a return appearance. All can be used to measure the size of the effect, the impact of the explanatory variable on the outcome variable. A confidence interval uses a sample value to estimate the range within which a population value falls; narrower confidence intervals give a more precise estimate. Cohen's *d* takes the difference between two means and standardizes it. Cohen has suggested *d* values, for the social and behavioral sciences, representing small, medium, and large effects. r^2 tells how much of the variability in the outcome variable is explained by or accounted for by, the explanatory variable.

DIY

Every year the Centers for Disease Control (CDC) conducts a survey of health practices and risk behaviors in American adults. The survey is called the BRFSS, the Behavioral Risk Factor Surveillance System. In 2013, the most recent year for which data are available, the sample consisted of almost 500,000 Americans, 18 and older, from all 50 states, Washington, D.C., Guam, and Puerto Rico.

I picked three variables in the data set—number of hours of sleep per night, weight in pounds, and number of alcoholic drinks consumed per day during the past month, on days that at least one drink was consumed—and found the mean for each for men, for women, and for both sexes combined.

	Hours of Sleep	Weight in Pounds	Drinks per Drinking Day
Men and women	7.05	176.54	2.21
Men	7.03	196.88	2.66
Women	7.06	161.79	1.80

Pick one of these variables and survey about 10 of your friends. Calculate the mean and the standard deviation. Then complete a single-sample *t* test to see whether your sample differs in its behavior from the U.S. population.

KEY TERMS

Cohen's *d* – a standardized measure of effect used to measure the difference between means.

critical value of *t* – value of *t* used to determine whether a null hypothesis is rejected or not; abbreviated t_{cv}.

degrees of freedom (*df*) – the number of values in a sample that are free to vary.

effect size (*d*) – a measure of the degree of impact of the independent variable on the dependent variable.

replicate – to repeat a study, usually introducing some change in procedure to make it better.

r^2 – an effect size that calculates how much of the variability in the outcome variable is accounted for by the predictor variable.

single-sample *t* test – a statistical test that compares a sample mean to a population mean when the population standard deviation is not known.

underpowered – term for a study with a sample size too small for the study to have a reasonable chance to reject the null hypothesis given the size of the effect.

CHAPTER EXERCISES

Answers to the odd-numbered exercises appear at the back of the book.

Review Your Knowledge

7.01 In order to use a single-sample *t* test, one *does* / *does not* need to know the population standard deviation.

7.02 The single-sample *t* test compares a sample ____ to a population ____.

7.03 A sample, selected at random from a population, may have a sample mean that differs from the population mean due to ____.

7.04 Tom ____ Harry ____ ____ infants.

7.05 One assumption of the single-sample *t* test is that the sample is a random sample from the population. This *is* / *is not* robust to violation.

7.06 The population that the sample comes from determines the population to which the results can be ____.

7.07 A second assumption of the single-sample *t* test is that observations within the sample are ____.

7.08 The third assumption of the single-sample *t* test is called the ____ assumption for short.

7.09 If a ____ assumption is violated, a researcher can still proceed with the test as long as the violation is not too great.

7.10 In order to write the null and alternative hypotheses for a single-sample *t* test, the researcher needs to know whether the test has one or two ____.

7.11 The default option in hypothesis testing is a ____-tailed test.

7.12 If a researcher is doing a one-tailed test, he or she should predict the ____ of the results before collecting any data.

7.13 The null hypothesis for a nondirectional, single-sample *t* test says that the sample *does* / *does not* come from the population.

7.14 The null hypothesis for a two-tailed, single-sample *t* test could be written as ____ $= \mu_2$.

7.15 The alternative hypothesis for a two-tailed, single-sample *t* test could be written as μ_1 ____ μ_2.

7.16 If the null hypothesis for a two-tailed, single-sample *t* test is true, $t =$ ____.

7.17 As the distance between the sample mean and the population mean grows, the value of *t* ____.

7.18 The abbreviation for the critical value of *t* is ____.

7.19 Determining the critical value depends on (a) how many ____ the test has, (b) how willing one is to make a Type ____ error, and (c) the ____.

7.20 If the hypotheses are nondirectional, then the researcher is doing a ____-tailed test.

7.21 If the alternative hypothesis is $\mu > 173$, then the test is a ____-tailed test.

7.22 If the researcher wants a 5% chance of Type I error, then alpha is set at ____.

7.23 For a two-tailed test with $\alpha = .05$, the rare zone has ____% of the sampling distribution in each tail.

7.24 When the sample size is large, the rare zone gets ____ and it is ____ to reject the null hypothesis.

7.25 Degrees of freedom, for a single-sample *t* test, equal ____ minus 1.

7.26 For a two-tailed, single-sample *t* test, the null hypothesis is rejected if $t \leq$ ____ or if $t \geq$ ____.

7.27 *t* for a single-sample *t* test is calculated by dividing the difference between *M* and μ by ____.

7.28 Interpretation uses human ____ to make meaning out of the results.

7.29 Interpretation is subjective, but needs to be supported by ____.

7.30 The first question asked in interpretation is about whether the ____ hypothesis is ____.

7.31 For the second interpretation question, one calculates ____, and for the third, one calculates a ____.

7.32 A decision is made about rejecting the null hypothesis by comparing the ____ value of *t* to the value calculated from the sample mean.

7.33 If a researcher rejects the null hypothesis, then he or she is forced to accept the ____.

7.34 If the null hypothesis is rejected, it is concluded that the mean for the population the ____ came from differs from the hypothesized value.

7.35 To determine the direction of a statistically significant difference, compare the ____ to the ____.

7.36 Sample size is reported in APA format for a single-sample *t* test by reporting ____.

7.37 APA format uses the inequality ____ to indicate that the null hypothesis was rejected when $\alpha = .05$.

7.38 If a result is reported in APA format as $p < .05$, that means the observed value of the test statistic fell in the ____ zone.

7.39 If one fails to reject the null hypothesis, one can say that there is ____ to conclude a difference exists between the population means.

7.40 APA format uses the inequality ____ to indicate that the null hypothesis was not rejected when $\alpha = .05$.

7.41 Effect sizes are used to quantify the impact of the ____ on the ____.

7.42 Cohen's *d* is an ____.

7.43 A value of 0 for Cohen's *d* means that the independent variable had ____ impact on the dependent variable.

7.44 Cohen considers a *d* of ____ a small effect, ____ a medium effect, and ____ or higher a large effect.

7.45 As the effect size *d* increases, the degree of overlap between the distributions for two populations ____.

7.46 When one fails to reject the null hypothesis, one should ____ calculate Cohen's *d*.

7.47 Calculating *d* when one has failed to reject the null hypothesis alerts one to the possibility of a Type ____ error.

7.48 To ____ a study is to repeat it.

7.49 r^2 measures how much variability in the _____ is explained by the _____.

7.50 *d* and r^2 should lead to similar conclusions about the size of an effect. Though both may have been calculated, it is _____ to report both.

7.51 If a researcher calculates a 95% confidence interval, he or she can be ____% confident that it captures the ____ value.

7.52 The 95% confidence interval for the difference between population means tells how far apart or how close the two ____ means might be.

7.53 The size of the difference between population means can be thought of as another ____.

7.54 If a 95% confidence interval for the difference between population means does not capture zero, and the researcher is doing a two-tailed test with $\alpha = .05$, then the null hypothesis *was / was not* rejected.

7.55 If the 95% confidence interval for the difference between population means falls close to zero, this means the size of the effect may be ____.

7.56 A wide 95% confidence interval for the difference between population means leaves a researcher unsure of ____.

Apply Your Knowledge

Picking the right test

7.57 A researcher has a sample ($N = 38$, $M = 35$, $s = 7$) that he thinks came from a population where $\mu = 42$. What statistical test should he use?

7.58 A researcher has a sample ($N = 52$, $M = 17$, $s = 3$) that she believes came from a population where $\mu = 20$ and $\sigma = 4$. What statistical test should she use?

Checking the assumptions

7.59 A researcher wants to compare the mean weight of a convenience sample of students from a college to the national mean weight of 18- to 22-year-olds. (a) Check the assumptions and decide whether it is OK to proceed with a single-sample *t* test. (b) Can the researcher generalize the results to all the students at the college?

7.60 There is a random sample of students from a large public high school. Each person, individually, takes a paper-and-pencil measure of introversion. (a) Check the assumptions and decide whether it is OK to proceed with a single-sample *t* test to compare the sample mean to the population mean of introversion for U.S. teenagers. (b) To what population can one generalize the results? (The same introversion measure was used for the national sample.)

Writing nondirectional hypotheses

7.61 The population mean on a test of paranoia is 25. A psychologist obtained a random sample of nuns and found $M = 22$. Write the null and alternative hypotheses.

7.62 A researcher wants to compare a random sample of left-handed people in terms of IQ to the population mean of IQ. Given $M = 108$ and assuming $\mu = 100$, write the null and alternative hypotheses.

Writing directional hypotheses

7.63 In America, the average length of time the flu lasts is 6.30 days. An infectious disease physician has developed a treatment that he believes will treat the flu more quickly. Write the null and alternative hypotheses.

7.64 An SAT-tutoring company claims that its students perform above the national average on SAT subtests. If the national average on SAT subtests is 500 and the tutoring company obtains SAT scores from a random sample of 626 of its students, write the null and alternative hypotheses.

Finding t_{cv} (assume the test is two-tailed and alpha is set at .05)

7.65 If $N = 17$, find t_{cv} and use it to draw a *t* distribution with the rare and common zones labeled.

7.66 If $N = 48$, find t_{cv} and use it to draw a *t* distribution with the rare and common zones labeled.

Writing the decision rule (assume the test is two-tailed and alpha is set at .05)

7.67 If $N = 64$, write the decision rules for a single-sample *t* test.

7.68 If $N = 56$, write the decision rules for a single-sample *t* test.

Calculating s_M

7.69 If $N = 23$ and $s = 12$, calculate s_M.

7.70 If $N = 44$ and $s = 7$, calculate s_M.

Given s_M, calculating t

7.71 If $M = 10$, $\mu = 12$, and $s_M = 1.25$, what is *t*?

7.72 If $M = 8$, $\mu = 6$, and $s_M = 0.68$, what is *t*?

Calculating t

7.73 If $N = 18$, $M = 12$, $\mu = 10$, and $s = 1$, what is *t*?

7.74 If $N = 25$, $M = 18$, $\mu = 13$, and $s = 2$, what is *t*?

Was the null hypothesis rejected? (Assume the test is two-tailed.)

7.75 If $t_{cv} = \pm 2.012$ and $t = -8.31$, is the null hypothesis rejected?

7.76 If $t_{cv} = \pm 2.030$ and $t = 2.16$, is the null hypothesis rejected?

7.77 If $t_{cv} = \pm 2.776$ and $t = 1.12$, is the null hypothesis rejected?

7.78 If $t_{cv} = \pm 1.984$ and $t = -1.00$, is the null hypothesis rejected?

Writing results in APA format (Assume the test is two-tailed and alpha is set at .05.)

7.79 Given $N = 15$ and $t = 2.145$, write the results in APA format.

7.80 Given $N = 28$ and $t = 2.050$, write the results in APA format.

7.81 Given $N = 69$ and $t = 1.992$, write the results in APA format.

7.82 Given $N = 84$ and $t = 1.998$, write the results in APA format.

Calculating Cohen's d

7.83 Given $M = 90$, $\mu = 100$, and $s = 15$, calculate Cohen's *d*.

7.84 Given $M = 98$, $\mu = 100$, and $s = 15$, calculate Cohen's *d*.

Calculating r^2

7.85 Given $N = 17$ and $t = 3.45$, what is r^2?

7.86 If $N = 29$ and $t = 1.64$, what is r^2?

Calculating a 95% confidence interval

7.87 Given $M = 45$, $\mu = 50$, $t_{cv} = \pm 2.093$, and $s_M = 2.24$, calculate the 95% confidence interval for the difference between population means.

7.88 Given $M = 55$, $\mu = 50$, $t_{cv} = \pm 2.093$, and $s_M = 2.24$, calculate the 95% confidence interval for the difference between population means.

Given effect size and confidence interval, interpret the results. Be sure to (1) tell what was done, (2) present some facts, (3) interpret the results, and (4) make a suggestion for future research. (Assume the test is two-tailed and $\alpha = .05$.)

7.89 Given the supplied information, interpret the results. A nurse practitioner has compared the blood pressure of a sample ($N = 24$) of people who are heavy salt users ($M = 138$, $s = 16$) to blood pressure in the general population ($\mu = 120$) to see if high salt consumption were related to raised or lowered blood pressure. She found:

- $t_{cv} = 2.069$
- $t = 5.50$
- $d = 1.13$
- $r^2 = 57\%$
- 95% CI [11.23, 24.77]

7.90 Given the supplied information, interpret the results. A dean compared the GPA of a sample ($N = 20$) of students who spent more than two hours a night on homework ($M = 3.20$, $s = 0.50$) to the average GPA at her college ($\mu = 2.80$). She was curious if homework had any relationship, either positive or negative, with GPA. She found:

- $t_{cv} = 2.093$
- $t = 3.64$
- $d = 0.80$
- $r^2 = 41\%$
- 95% CI [0.17, 0.63]

Completing all six steps of a hypothesis test

7.91 An educational psychologist was interested in time management by students. She had a theory that students who did well in school spent less time involved with online social media. She found out, from the American Social Media

Research Collective, that the average American high school student spends 18.68 hours per week using online social media. She then obtained a sample of 31 students, each of whom had been named the valedictorian of his or her high school. These valedictorians spent an average of 16.24 hours using online social media every week. Their standard deviation was 6.80. Complete the analyses and write a paragraph of interpretation.

7.92 A psychology professor was curious how psychology majors fared economically compared to business majors. Did they do better, or did they do worse five years after graduation? She did some research and learned that the national mean for the salary of business majors five years after graduation was $55,000. She surveyed 22 recent psychology graduates at her school and found $M = \$43{,}000$, $s = 12{,}000$. Complete the analyses and write a paragraph of interpretation.

Expand Your Knowledge

7.93 Imagine a researcher has taken two random samples from two populations (A and B). Each sample is the same size ($N = 71$), has the same sample mean ($M = 50$), and comes from a population with the same mean ($\mu = 52$). The two populations differ in how much variability exists. As a result, one sample has a smaller standard deviation ($s = 2$) than the other ($s = 12$). The researcher went on to calculate single-sample t values for each sample. Based on the information provided below, how does the size of the sample standard deviation affect the results of a single-sample t test?

	s_M	t_{cv}	t	d	r^2	CI	Width of CI
A: Less variability ($s = 2$)	0.24	1.994	−8.33	−1.00	50%	−2.48 to −1.52	0.96
B: More variability ($s = 12$)	1.42	1.994	−1.41	−0.17	3%	−4.83 to 0.83	5.66

7.94 Another researcher selected two random samples from one population. This population has a mean of 63 ($\mu = 63$). It turned out that each sample had the same mean ($M = 60$) and standard deviation ($s = 5$). The only way the two samples differed was in terms of size: one, C, was smaller ($N = 10$) and one, D, was larger ($N = 50$). The researcher went on to conduct a single-sample t test for each sample. Based on the information provided below, how does sample size affect the results of a single-sample t test?

	s_M	t_{cv}	t	d	r^2	CI	Width of CI
C: Smaller N ($N = 10$)	1.58	2.262	−1.90	−0.60	29%	−6.57 to 0.57	7.14
D: Larger N ($N = 50$)	0.71	2.010	−4.23	−0.60	27%	−4.43 to −1.57	2.86

7.95 A third researcher obtained two random samples from another population. The mean for this population is 50 ($\mu = 50$). Each sample was the same size ($N = 10$) and each had the same standard deviation ($s = 20$). But one, sample E, had a mean of 80 ($M = 80$) and the other, sample F, a mean of 60 ($M = 60$). Based on the information provided below, how does the distance from the sample mean to the population mean affect the results of a single-sample t test?

	s_M	t_{cv}	t	d	r^2	CI	Width of CI
E: M farther from μ($M = 80$)	6.32	2.262	4.75	1.50	71%	15.70 to 44.30	28.60
F: M closer to μ($M = 60$)	6.32	2.262	1.58	0.50	22%	−4.30 to 24.30	28.60

7.96 Based on the answers to Exercises 7.93 to 7.95, what factors have an impact on a researcher's ability to reject the null hypothesis? Which one(s) can he or she control?

7.97 A researcher is conducting a two-tailed, single-sample t test with alpha set at .05. What is the largest value of t that one could have that, no matter how big N is, will guarantee failing to reject the null hypothesis?

7.98 If t_{cv} for a two-tailed test with $\alpha = .05$ is ± 2.228, what would the alpha level be for the critical value of 2.228 as a one-tailed test?

7.99 If $N = 21$ and $s_M = 1$, write as much of the equation as possible for calculating the 90% $CI\mu_{Diff}$ and the 99% $CI\mu_{Diff}$. Use Equation 7.4 as a guide.

7.100 Which confidence interval in Exercise 7.95 is wider: 90% or 99%? Explain why.

SPSS

Let's use SPSS to analyze the Chips Ahoy data from the Application Demonstration at the end of the chapter. The first step is data entry. **Figure 7.12** shows the data for the 10 cookies. Note that the variable is labeled "Num_chips" at the top of the column, and that the value for each case appears in a separate row.

	Num_chips
1	27.00
2	23.00
3	28.00
4	29.00
5	27.00
6	28.00
7	22.00
8	22.00
9	28.00
10	27.00

Figure 7.12 Data Entry for Single-Sample *t* Test in SPSS The number of chips in each cookie is entered on its own row. (Source: SPSS)

SPSS calls a single-sample *t* test the "One-Sample T Test." It is found by clicking on "Analyze" on the top line (**Figure 7.13**). Then click on "Compare Means" and "One-Sample T Test. . . ."

This pulls up the box seen in **Figure 7.14**. Note that the variable "Num_chips," which we wish to analyze, has already been moved over from the box on the left

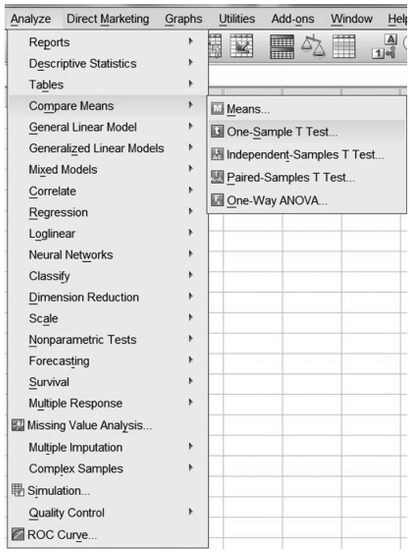


Figure 7.13 The Single-Sample *t* Test in SPSS SPSS calls the single-sample *t* test the "One-Sample T Test." (Source: SPSS)

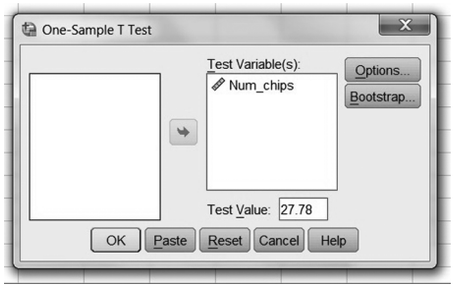


Figure 7.14 Selecting the Dependent Variable and Entering the Population Mean for a Single-Sample *t* Test in SPSS The dependent variable being tested is listed in the box labeled "Test Variable(s)." The population value it is being compared to is entered in the box labeled "Test Value." (Source: SPSS)

(which is now empty) to the box labeled "Test Variable(s)." The population mean, 27.78, has been entered into the box labeled "Test Value." Once this is done, it is time to click the "OK" button on the lower right.

Figure 7.15 shows the output that SPSS produces. There are a number of things to note:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Num_chips	10	26.1000	2.68535	.84918

One-Sample Test

	Test Value = 27.78					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Num_chips	-1.978	9	.079	-1.68000	-3.6010	.2410

Figure 7.15 SPSS Output for a Single-Sample *t* Test SPSS provides descriptive statistics, the *t* value, an exact significance level, and the 95% confidence interval for the difference between population means. It does not calculate Cohen's *d* or r^2 but provides enough information that these can be done by hand. (Source: SPSS)

- The first box provides descriptive statistics, including the standard error of the mean, s_M, the denominator for *t*.
- At the top of the second output box, "Test Value = 27.78" indicates that this is the value against which the sample mean is being compared.
- Next, SPSS reports the *t* value (−1.978) and the degrees of freedom (9).
- SPSS then reports what it calls "Sig. (2-tailed)" as .079.
  - The important thing for us is whether this value is $\leq .05$ or $> .05$.
    - If it is $\leq .05$, the null hypothesis is rejected and the results are reported in APA format as $p < .05$.
    - If it is $> .05$, the null hypothesis is not rejected and the results are reported in APA format as $p > .05$.
  - This value, .079, is the exact significance level for this test with these data. It indicates what the two-tailed probability is of obtaining a *t* value of −1.978 *or larger* if the null hypothesis is true. In the present situation, it says that a *t* value of −1.978 is a common one—it happens 7.9% of the time when 10 cases are sampled from a population where $\mu = 27.78$. APA format calls for the exact significance level to be used. Thus, these results should be reported as $t(9) = -1.98$, $p = .079$.
  - The next bit of output reports the mean difference, −1.68, between the sample mean (26.10) and the test value (27.78).
  - SPSS reports what it calls the "95% Confidence Interval of the Difference," what is called in the book the 95% confidence interval for the difference between population means, ranging from a "Lower" bound of −3.60 to an "Upper" bound of 0.24.
  - Note that SPSS does not report Cohen's *d*. Similarly, there is not enough information to calculate r^2. However, it provides enough information, the mean difference (−1.68) and the standard deviation (2.685), that it can be done by hand with Equation 7.3.