The Paired-Samples *t* Test

LEARNING OBJECTIVES

- Recognize the different types of dependent samples.
- Calculate and interpret a paired-samples *t* test.

CHAPTER OVERVIEW

This is the third, and final, chapter on *t* tests. Chapter 7 covered a *t* test for comparing a sample mean to a specified value, the single-sample *t* test. Chapter 8 moved to a *t* test for comparing the means of two *independent* samples. Now, in Chapter 9, we add a **paired-samples *t* test**, for comparing the means of two *dependent* samples.

9.1 Paired Samples

When samples are dependent, each case consists of a pair of data points, one data point from each of two samples. In dependent samples, also called paired samples, the data points may be paired in a variety of ways. One pairing, called a **repeated-measures design**, means the same participants provide data at two points in time. (This is also called **longitudinal research** because it follows participants over time or a **pre-post design** as participants are measured on the outcome variable before and after an intervention.) An example of a repeated-measures design would be measuring the level of anxiety in people before and after learning relaxation techniques.

Another type of pairing, called a **within-subjects design,** involves the same participants being measured in two different situations or under two different conditions. For example, a cognitive psychologist might measure how much information people retain when studying in silence and then measure information retention *for the same participants* when they study while listening to music.

Repeated-measures and within-subjects designs involve one sample of cases measured at two points in time or under two conditions. Find this confusing. "Why," they ask, "is it called a *two*-sample test when there is just *one* sample of cases?" Unfortunately, this is statistical terminology that just needs to be learned. To a statistician, each condition in a dependent samples study is considered a "sample."

Repeated-measures and within-subjects designs have a significant advantage over independent-samples designs. These dependent-samples designs control for **individual differences,** attributes that vary from case to case. Because the same participants are in both groups, the researcher can be sure that the two samples are comparable in terms of background characteristics. As a result of this, the researcher can be more confident that any observed difference between the groups on the outcome variable is due to the explanatory variable and not some confounding variable.

In an attempt to derive this benefit, researchers have developed a number of other paired-samples techniques in which different participants are in two conditions. In one, the pairs have some similarity because of some connection, either biological (such as that between two siblings), or formed (such as that between a romantic couple). Another is called *matched pairs.* In **matched pairs**, participants are grouped, by the researcher, into sets of two based on their being similar on potential confounding variables. For example, if a dean were comparing the GPAs of male and female students, she might want to match them, based on IQ, into male–female pairs. That way, a researcher couldn't argue that intelligence was a confounding variable if one sex had a higher GPA.

Because there are so many different types of paired samples, the paired-samples *t* test has more names than any other test in statistics. But, whether it is called a paired-samples *t* test, dependent-samples *t* test, correlated-samples *t* test, related-samples *t* test, matched-pairs *t* test, within-subjects *t* test, or repeated-measures *t* test, it is all the same test.

The wide number of different names reflects how commonly used the paired-samples *t* test is. It is a commonly used test for several reasons. One reason is that many experimental situations are of a pre-post design where the outcome variable is measured before and after the explanatory variable is applied. Another reason is that controlling individual differences makes studies that use paired samples more powerful than studies that use independent samples. In a statistical sense, being more powerful means that the probability of being able to reject the null hypothesis, when it is false, is higher. As a result, a researcher needs a smaller sample size for a paired-samples *t* test than for an independent-samples *t* test. This is a big advantage of paired-samples *t* tests. If a researcher is studying a rare phenomenon or one where participants are hard to come by, a dependent-samples design is the way to go.

If a researcher is studying a rare phenomenon or one where participants are hard to come by, a dependent-samples design is the way to go.

Here is an example of research that used paired samples to investigate how stress affects recovery from a physical wound (Kiecolt-Glaser et al., 1995). One sample, the people who were under stress, consisted of women who were caring for a husband or mother with Alzheimer's disease. Because the researchers believed that age and socioeconomic status might influence physical recovery, they matched each caregiver with a control participant of the same sex, age, and family income who was not a caregiver. So, the participants were matched pairs of women, one a caregiver (the experimental group) and one a control.

Using a dermatology procedure, the researchers made a small wound on each participant's forearm and timed how long it took to heal. The wound took almost 10 days longer on average to heal in the caregivers ($M = 48.7$) than in the controls ($M = 39.3$), and this difference was statistically significant. Why did this difference exist? Well, because the pairs were matched on age and socioeconomic status, it can't be argued that the caregivers were older or poorer. With these confounding variables removed, it seems more plausible that it is the stress of caring for someone who is deteriorating with a chronic illness that affects how quickly one heals from a physical wound.

Now, having seen paired-samples in action and observed their advantages, it's time to learn how to perform a paired-samples *t* test.

9.2 Calculating the Paired-Samples *t* Test

Here are some data that are appropriate for analysis with a paired-samples *t* test. Imagine that a sensory psychologist, Dr. Keim, wanted to examine the effect of humidity on perceived temperature. She obtained six volunteers at her college and tested them, individually, in a temperature- and humidity-controlled chamber. Each participant was tested twice and the tests were separated by 24 hours. For both tests, the temperature in the chamber was set at 76°F. For one test the humidity level was "low," and for the other test the humidity level was "high." In order to avoid any effects due to the order of the tests, which humidity level each participant would experience first was randomly determined. For a test, a participant spent 15 minutes in the chamber, after which he or she was asked what the temperature was inside it. This "perceived temperature" is the study's dependent variable.

The data from the humidity study are shown in Table 9.1, where each case appears on a row and each condition (sample) in a column. Table 9.1 also contains the means and standard deviations for both conditions. Remember, in both conditions the actual temperature was the same, 76°F. The mean perceived temperature on the low-humidity test was 75.00°F, and in the high-humidity condition it was 82.50°F. Between the two test conditions, there was a 7.50°F difference in the means.

TABLE 9.1 The Effect of Humidity Level on Perceived Temperature in °F

Participant	Low-Humidity Test (control condition)	High-Humidity Test (experimental condition)	Difference Score
1	76	81	5.00
2	80	90	10.00
3	78	85	7.00
4	72	82	10.00
5	76	82	6.00
6	68	75	7.00
$M =$	75.00	82.50	7.50
$s =$	4.34	4.93	2.07

The difference score is calculated for each pair (row) of scores by subtracting the value for one test condition from the value for the other test condition. It doesn't matter which value is subtracted from the other, as long as the same order is used consistently. Here, in order to end up with positive numbers, the low-humidity condition is subtracted from the high-humidity condition.

Figure 9.1 is a pair of box-and-whisker plots, showing the median, interquartile range, and minimum/maximum for each condition. When looking at the graph, the difference seems clear—temperature does appear to be perceived as higher when the humidity is higher. But, looks can be deceiving and the sample size is low, so hypothesis testing is necessary to find out if the difference is a statistically significant one or if it can be explained by sampling error.

Step 1 Pick a Test. Dr. Keim, using a within-subjects design, is comparing the means for a sample of people measured in two different conditions, low humidity and high humidity. Remember, when one sample is measured in two different conditions,

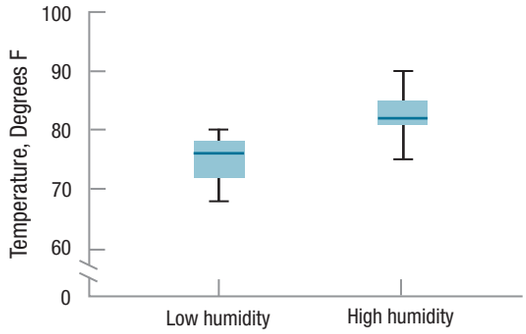

Figure 9.1 Box-and-Whisker Plots Showing Perceived Temperature in Low- and High-Humidity Conditions This graphic appears to show that temperature is perceived as hotter when the humidity is higher. To determine whether the difference between the two conditions is a statistically significant one, a paired-samples *t* test is needed.

as happens here, statisticians consider this *two* dependent samples. This situation, comparing the means of dependent samples, calls for a paired-samples *t* test.

Tom **a**nd **H**arry **d**espise **c**rabby **i**nfants

Step 2 Check the Assumptions. There are three assumptions for the paired-samples *t* test (Table 9.2), and they are familiar. The first assumption is that the sample is a random sample from the population to which the results will be generalized. Dr. Keim would like to be able to generalize her results to people in general, that is, to all the people in the world. She recognizes that she has a convenience sample not a random sample from this population, so this assumption is violated. The random samples assumption is robust, however, so she can continue with the test. She is aware of no one who suggests that Americans and/or college students perceive heat differently from others, so she will be willing to generalize her results more broadly than just to U.S. college students.

The second assumption is that the observations are independent within a sample. Be careful in assessing this assumption: it refers to independence *within* a

TABLE 9.2 Assumptions for the Paired-Samples *t* Test

Random sample: The sample is a random sample from the population.	Robust to violation
Independence of observations: Each case within a group or condition is independent of the other cases in that group or condition.	Not robust to violation
Normality: The population of difference scores is normally distributed.	Robust to violation

sample, not *between* samples. Since the same cases are in both samples, or conditions, *the two samples* are not independent. However, each person in each sample is tested individually and each person is only tested once in each condition. The independence of observations assumption is not violated.

The third assumption, the normality assumption, says that in the larger population *the difference scores* are normally distributed. Look at Table 9.1 and note that there's a third column listing the difference between the value a case had in one condition (high humidity) and its score in the other condition (low humidity). This is called a difference score, abbreviated *D*, and is needed to calculate a paired-samples *t* test. The formula for *D* is shown in Equation 9.1.

Equation 9.1 Formula for Calculating Difference Score, *D*

$$D = X_1 - X_2$$

where D = the difference score being calculated
X_1 = a case's score in Condition 1
X_2 = a case's score in Condition 2

In calculating *D*, it matters little which value is subtracted from the other, as long as the same order is followed for all cases. Most people prefer to work with positive numbers, so feel free to decide the order for the subtraction to maximize the number of positive values. Dr. Keim chose to subtract the low-humidity scores from the high-humidity scores in order to have positive difference scores.

For the first case, the person who perceived the low-humidity condition as 76° and the high-humidity condition as 81°F, the difference score is calculated as

$$\begin{aligned} D &= X_{\text{HighHumidity}} - X_{\text{LowHumidity}} \\ &= 81 - 76 \\ &= 5.0000 \\ &= 5.00 \end{aligned}$$

A sample size of 6 is a little small for making decisions about the shape of a parent population. However, based on her previous research, Dr. Keim is willing to assume that the difference scores are normally distributed in the population, so the normality assumption is not violated.

Step 3 List the Hypotheses. The hypotheses, which are statements about populations, are going to be the same as they were for the independent-samples *t* test. The null hypothesis is going to say that the two population means are the same. The alternative hypothesis will state that the two population means differ, but it won't indicate

whether the difference is large or small, positive or negative. These hypotheses are nondirectional or two-tailed, so they don't specify a direction for the difference. Thus, Dr. Keim is testing for two possibilities: (1) low humidity is perceived as hotter than high humidity, or (2) high humidity is perceived as hotter than low humidity. The generic form of two-tailed hypotheses is

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The specific form of the two hypotheses for this study is

$$H_0: \mu_{\text{LowHumidity}} = \mu_{\text{HighHumidity}}$$
$$H_1: \mu_{\text{LowHumidity}} \neq \mu_{\text{HighHumidity}}$$

Of course, it is possible to have directional hypotheses for a paired-samples *t* test. If Dr. Keim had predicted, in advance, that high humidity would lead to a higher perceived temperature, her hypotheses would have been

$$H_0: \mu_{\text{LowHumidity}} \geq \mu_{\text{HighHumidity}}$$
$$H_1: \mu_{\text{LowHumidity}} < \mu_{\text{HighHumidity}}$$

Step 4 Set the Decision Rule. The decision rule for a paired-samples *t* test is formulated the same way as it was for the independent-samples *t* test. The critical value of *t*, t_{cv}, found in Appendix Table 3, is based on (1) the number of tails for the test, (2) how willing one is to make a Type I error (i.e., the alpha level), and (3) how many degrees of freedom there are. The default option, or the most common form of the paired-samples *t* test, as for other statistical tests, is a two-tailed test with alpha set at .05.

Once the observed value of *t* is calculated (Step 5), it is compared to the critical value in order to decide whether or not to reject the null hypothesis. For a two-tailed test, the general form of the decision rule is:

- If $t \leq -t_{cv}$ *or* $t \geq t_{cv}$, reject the null hypothesis.
- If $-t_{cv} < t < t_{cv}$, fail to reject the null hypothesis.

Dr. Keim is doing a two-tailed test and is content to use the default alpha level of .05. All that she needs to do is calculate degrees of freedom. Equation 9.2 is the formula for calculating degrees of freedom for a paired-samples *t* test.

Equation 9.2 Degrees of Freedom (*df*) for a Paired-Samples *t* Test

$$df = N - 1$$

where df = the degrees of freedom
N = the number of *pairs* of cases

Dr. Keim's study involves six *pairs* of cases, so degrees of freedom are calculated as

$$df = 6 - 1$$
$$= 5$$

Look in Appendix 3 and find the intersection of the column for a two-tailed hypothesis test with the alpha set at .05 and the row for $df = 5$. There, the critical value of *t*, ±2.571, is found. **Figure 9.2** uses the critical values to mark the rare zone, where the null hypothesis is rejected, and the common zone, where it is not.

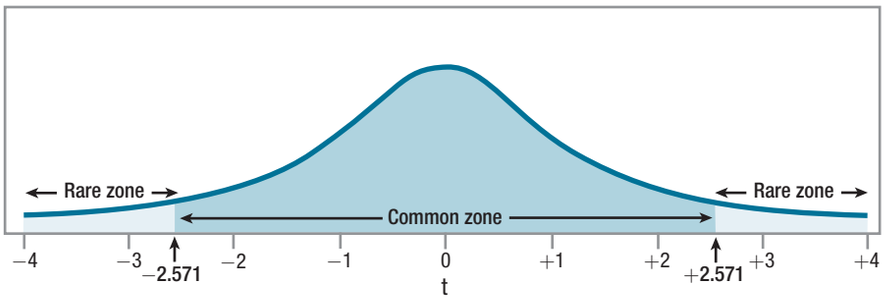

Figure 9.2 Critical Value of t, Two-Tailed, $\alpha = .05$, $df = 5$ This is the sampling distribution of *t* that would occur if the null hypothesis were true and $\mu_1 = \mu_2$. When the null hypothesis is true, the observed value will fall in the common zone 95% of the time and it will fall in the rare zone only 5% of the time.

Here is the decision rule for Dr. Keim's study:

- If $t \leq -2.571$ *or* $t \geq 2.571$, reject the null hypothesis.
- If $-2.571 < t < 2.571$, fail to reject the null hypothesis.

Step 5 Calculate the Test Statistic. The same general procedure is used for calculating the paired-samples *t* value as was used for the independent-samples *t:* divide the difference between the two sample means by the standard error of the difference. This standard error of the difference is the standard error of the difference for difference scores and will be abbreviated s_{M_D} to differentiate it from the other denominators for *t* tests. The **standard error of the mean difference for difference scores** is the standard deviation of the sampling distribution of difference scores.

Equation 9.3 Formula for the Standard Error of the Mean Difference for Difference Scores (s_{M_D})

$$s_{M_D} = \frac{s_D}{\sqrt{N}}$$

where s_{M_D} = the standard error of the mean difference for difference scores
s_D = the standard deviation (s) of the difference scores
N = the number of pairs of cases

Using Equations 3.6 and 3.7, Dr. Keim has calculated the standard deviation of the difference scores and found $s_D = 2.07$. There are six pairs of cases, so $N = 6$. Plugging these values into Equation 9.3 gives

$$\begin{aligned} s_{M_D} &= \frac{s_D}{\sqrt{N}} \\ &= \frac{2.07}{\sqrt{6}} \\ &= \frac{2.07}{2.4495} \\ &= 0.8451 \\ &= 0.85 \end{aligned}$$

The standard error of the mean difference for the difference scores is 0.85. This value will be used in Equation 9.4, the formula for calculating *t*, the value of the test statistic for the paired-samples *t* test.

Equation 9.4 Formula for Calculating *t*, the Value of the Test Statistic for a Paired-Samples *t* Test

$$t = \frac{M_1 - M_2}{s_{M_D}}$$

where t = the value of the test statistic for a paired-samples *t* test
M_1 = the mean of one sample
M_2 = the mean of the other sample
s_{M_D} = standard error of the mean difference for difference scores (Equation 9.3)

Given $s_{M_D} = 0.85$, $M_{\text{LowHumidity}} = 75.00$, and $M_{\text{HighHumidity}} = 82.50$, Dr. Keim is ready to calculate *t*. Again, it matters little which mean is called M_1 and which is M_2, so Dr. Keim has arranged the calculations to end up with a positive number in the numerator, assuring a positive *t* value:

$$t = \frac{M_{\text{HighHumidity}} - M_{\text{LowHumidity}}}{s_{M_D}}$$

$$= \frac{82.50 - 75.00}{0.85}$$

$$= \frac{7.5000}{0.85}$$

$$= 8.8235$$

$$= 8.82$$

Dr. Keim's *t* value is 8.82, and she is done with Step 5. In the next section of the chapter, after more practice with the first five steps, we'll follow Dr. Keim as she applies the decision rule and interprets the results.

Worked Example 9.1

A clinical psychologist, Dr. Althof, was studying the long-term effectiveness of psychodynamic psychotherapy for depression. He obtained a sample of 16 people with moderate to severe depression and assigned each to receive 20 sessions of psychodynamic therapy from a trained therapist. At the end of treatment, he administered a depression scale to each person and determined that the mean level of depression was 14.00. Scores on this depression scale can range from 0 to 50, and higher scores indicate greater depression. Six months later, Dr. Althof tracked down all 16 participants and readministered the depression scale, finding the mean was now 15.00. **Figure 9.3** is a pair of box-and-whisker plots showing the depression scores at the two points in time. A 1-point increase on a 50-point scale doesn't sound like very much, but the box-and-whisker plots suggest an increase in the depression level in the six months following the end of treatment. Dr. Althof needs hypothesis testing to find out if the change is a statistically significant one.

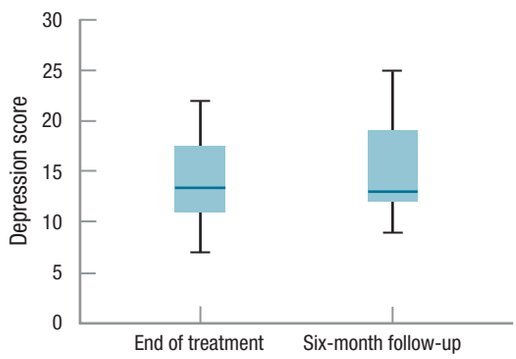

Figure 9.3 Box-and-Whisker Plots Showing Depression Level after Treatment and at Six-Month Follow-Up In this graph, there is some increase in the depression level in the six months following treatment. It will take a paired-samples *t* test to determine if the difference is statistically significant.

Step 1 Pick a Test. Dr. Althof is comparing the mean of a sample at one time to the mean of the sample at a second time. Though it is one sample of people, they are measured at two points in time, making it appropriate for a paired-samples *t* test.

Step 2 Check the Assumptions. The assumptions for the paired-samples *t* test are listed in Table 9.2.

- The random samples assumption is violated. It is not reported that the 16 cases were a random sample from the population of all the people in the world with moderate to severe depression, so it is safe to assume this is not a random sample from that population. When this robust assumption is violated, however, a researcher can still proceed with the test. Dr. Althof just has to be careful about the population to which he generalizes the results.
- The independence of observations assumption is not violated. Each participant received individual therapy, so the participants within a sample didn't influence each other.
- The assumption for the normality of difference scores is not violated. Dr. Althof knows from his review of the literature that depression scores are normally distributed. He is willing to assume that the difference scores (six-month follow-up depression score minus end-of-treatment depression score) will be normally distributed in the larger population.

Step 3 List the Hypotheses. Sometimes the effect of treatment increases over time, but more often the effect of treatment decreases over time, so Dr. Althof was open to both options when he planned the study. As a result, his hypotheses are nondirectional (two-tailed). The null hypothesis states that the two population means (end-of-treatment mean vs. six-month follow-up mean) are the same, and the alternative hypothesis says that the two population means differ:

$$H_0: \mu_{EOT} = \mu_{6MFU}$$
$$H_1: \mu_{EOT} \neq \mu_{6MFU}$$

Step 4 Set the Decision Rule. The critical value of t depends on the number of tails, alpha, and degrees of freedom. With nondirectional hypotheses, the test is two-tailed. Dr. Althof is comfortable setting the alpha at the usual level, .05, and having a 5% chance of Type I error. Finally, using Equation 9.2, he calculates degrees of freedom for the paired-samples t test:

$$\begin{aligned} df &= N - 1 \\ &= 16 - 1 \\ &= 15 \end{aligned}$$

Consulting Appendix Table 3 in the column for a two-tailed test with an alpha of .05 and the row with 15 degrees of freedom, Dr. Althof finds that the critical value of t is ± 2.131. The common and rare zones for his decision rule are shown in Figure 9.4. The decision rule is:

- If $t \leq -2.131$ *or* $t \geq 2.131$, reject the null hypothesis.
- If $-2.131 < t < 2.131$, fail to reject the null hypothesis.

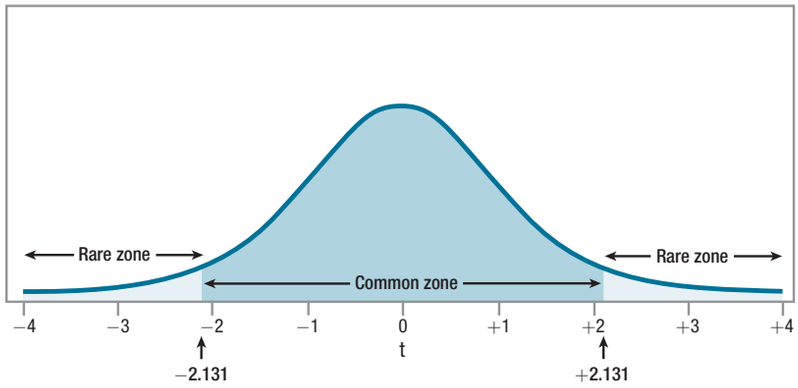

Figure 9.4 Critical Value of *t*, Two-Tailed, $\alpha = .05$, $df = 15$ This is the sampling distribution of t for 15 degrees of freedom. Compare this to the sampling distribution of t in Figure 9.2, where $df = 5$. Note that this sampling distribution is more peaked, packed more closely around zero. As a result, t_{cv} here, 2.131, is closer to zero, making the rare zone larger. When the rare zone is larger, which occurs when the sample size is larger, it is easier to reject the null hypothesis.

Step 5 Calculate the Test Statistic. The data set, with depression scores at the end of treatment and six months later, is shown in Table 9.3. Also shown are the difference scores and the standard deviation of the difference scores ($s_D = 2.31$).

The first step in the calculations is applying Equation 9.3 to find the standard error of the difference, s_{M_D}:

$$\begin{aligned} s_{M_D} &= \frac{s_D}{\sqrt{N}} \\ &= \frac{2.31}{\sqrt{16}} \\ &= \frac{2.31}{4.0000} \\ &= 0.5775 \\ &= 0.58 \end{aligned}$$

The value for the standard error of the difference, 0.58, is then used in Equation 9.4 to find t. It doesn't matter which mean is subtracted from which, so

TABLE 9.3 Data for Depression Score at the End of Treatment and at Six-Month Follow-Up

Participant	Six-month Follow-up	End of Treatment	Difference Score
1	11	12	−1.00
2	12	8	4.00
3	10	7	3.00
4	19	18	1.00
5	25	21	4.00
6	19	18	1.00
7	13	15	−2.00
8	19	17	2.00
9	13	10	3.00
10	15	12	3.00
11	9	12	−3.00
12	12	10	2.00
13	13	13	0.00
14	16	14	2.00
15	12	15	−3.00
16	22	22	0.00
M	15.00	14.00	1.00
s	4.58	4.37	2.31

Dr. Althof arranged them to end up with a positive value by subtracting the follow-up mean (14.00) from the end-of-treatment mean (15.00):

$$t = \frac{M_{6MFU} - M_{EOT}}{s_{M_D}}$$

$$= \frac{15.00 - 14.00}{0.58}$$

$$= \frac{1.0000}{0.58}$$

$$= 1.7241$$

$$= 1.72$$

Having found $t = 1.72$, Step 5 is complete. All that's left is the interpretation, which we'll turn to in the next section.

Practice Problems 9.1

Apply Your Knowledge

9.01 Given the following pairs of scores, calculate difference scores: 72 and 75; 69 and 45; 42 and 39; 47 and 46; 55 and 61; 50 and 61; 71 and 69; 55 and 69.

9.02 Given $s_D = 8.43$ and $N = 64$, calculate s_{M_D}.

9.03 Given $M_1 = 19.98$, $M_2 = 18.65$, and $s_{M_D} = 2.45$, calculate *t*.

9.3 Interpreting the Paired-Samples *t* Test

Interpreting a paired-samples *t* test starts by addressing the same questions that were used in interpreting the single-sample *t* test and the independent-samples *t* test:

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

And the interpretation ends the same way as well, with a written statement that covers four points:

1. What was the study about?
2. What were its main results?
3. What do these results mean?
4. Are there specific suggestions for future research?

Before Dr. Keim starts the interpretation, let's review her study on the effect of humidity on perceived temperature. She used six participants in a within-subjects design, where each person was tested twice in a 76° room, once at low humidity and once at high humidity. The mean perceived temperature in the low-humidity condition was 75.00°F ($s = 4.34$°F) and 82.50°F ($s = 4.93$°F) in the high-humidity condition. The mean difference score was 7.50 ($s_D = 2.07$°F). For a two-tailed test with $\alpha = .05$ and $df = 5$, t_{cv} was ±2.571°F. Dr. Keim calculated $s_{M_D} = 0.85$°F and $t = 8.82$°F.

Was the Null Hypothesis Rejected?

This first interpretation question can be answered by plugging the observed value of the test statistic, $t = 8.82$, into the decision rule Dr. Keim formulated in Step 4:

- Is $8.82 \leq -2.571$ *or* $8.82 \geq 2.571$?
- Or, is $-2.571 < 8.82 < 2.571$?

The first statement is true because 8.82 is greater than or equal to 2.571. This can be seen in Figure 9.5, where it is clear that the value of the test statistic falls in the rare zone. Dr. Keim has rejected the null hypothesis.

Rejecting the null hypothesis that the two population means are the same leads to accepting the alternative hypothesis that the two population means are different.

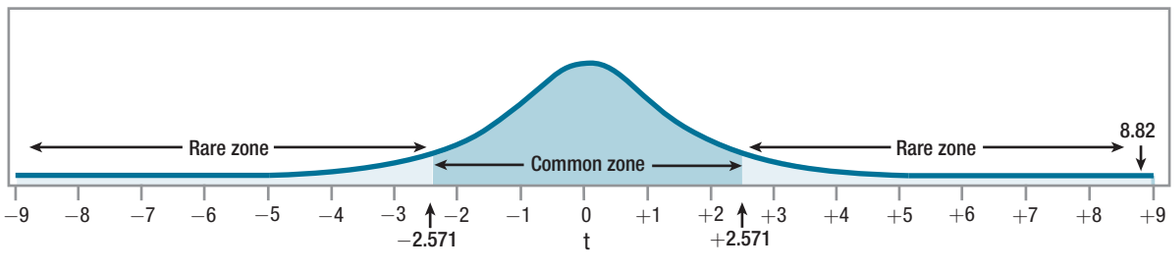

Figure 9.5 Observed Value of *t* from Humidity Study The observed value of *t*, 8.82, falls in the rare zone, so the null hypothesis is rejected.

Dr. Keim can say that a statistically significant difference exists between the perceived temperature in low humidity vs. high humidity.

The Direction of the Difference

Reporting that there is a difference is true, but it is not very useful. The logical follow-up question is, "What is the direction of the difference?" By examining the two sample means, 75.00° for the low-humidity condition vs. 82.50° for the high-humidity condition, Dr. Keim can conclude that 76° feels significantly hotter with high humidity than with low humidity.

A researcher only needs to worry about the direction of the difference when the null hypothesis is rejected. If Dr. Keim had failed to reject the null hypothesis, then there would not have been enough evidence to conclude that a difference exists between the two means, so there would have been no reason to consider the direction of the difference.

APA Format

Reporting the results in APA format for the humidity results, Dr. Keim would write

$$t(5) = 8.82,\ p < .05$$

For a *t*-test, APA format contains five pieces of information: (1) what test was done, (2) the number of cases, (3) the value of the test statistic, (4) the alpha level chosen, and (5) whether the null hypothesis was rejected.

1. The initial *t* says that this was a *t* test.
2. The number in parentheses, 5, is the degrees of freedom. By adding 1 to the number in the parentheses (i.e., 5 + 1 = 6), one can determine the number of pairs of cases.
3. The number after the equals sign, 8.82, gives the value of the test statistic.
4. The final value, .05, shows that there is a 5% chance of making a Type I error because alpha was set at .05.
5. $p < .05$ is how APA says that the null hypothesis has been rejected. It means that the results are rare, they happen less than 5% of the time, when the null hypothesis is true. (If alpha had been set at something other than .05, say .10, then write $p < .10$ instead.)

Cohen's *d* and r^2 for the Paired-Samples *t* Test

In the last two chapters, for the single-sample *t* test and the independent-samples *t* test, we calculated Cohen's *d* and r^2 in order to quantify how much impact the explanatory variable had on the outcome variable. It is possible to calculate Cohen's *d* and r^2 for a paired-samples *t* test, but doing so is *not* advisable.

The reason that these effect sizes are inappropriate for a paired-samples *t* test is that they measure more than just the impact of the explanatory variable on the dependent variable. Cohen's *d* and r^2 for a paired-samples *t* test also include the effect of individual differences. As a result, they overestimate the size of the effect of the independent variable on the dependent variable. For example, r^2 would be 94% for the humidity data if it were calculated. If true, it would mean that humidity status

explains 94% of the variability in perceived temperature. That would be a huge effect for the explanatory variable of humidity level. Unfortunately, it is not true. Here's why.

Look at Table 9.1, which presents the data for each case in each of the two conditions. Notice that there's variability within a condition: not everyone perceives the temperature the same way. As everyone in a condition receives the same treatment, the variability within a condition results from individual differences among the participants. Look at case 6 who perceived the low-humidity condition as the coldest, 68°. Case 6 also gives the lowest rating of the temperature in the high-humidity condition. Case 6 differs from the other individuals in this study in that he or she always feels cold. Case 2, in contrast, differs in the perception of temperature, feeling the warmest in both conditions.

The type of participant, cold-sensitive or heat-sensitive, has a large effect on the temperature the environment is perceived to be. Whether a person is cold-sensitive or heat-sensitive is an individual differences variable. When r^2 or Cohen's d is calculated for a paired-samples t test, they mix in the impact of these individual differences on temperature *with* the impact of the different conditions (high vs. low humidity) on temperature. This means that they give an inflated effect size and are not appropriate to estimate effect size for a paired-samples t test.

So, how does a researcher measure effect size for a paired-samples t test? One way is to calculate the confidence interval for the difference between population means and then use professional expertise—and common sense—to translate it into an effect size. Another approach, not possible until Chapter 12, is to use a test called the repeated-measures ANOVA instead of a paired-samples t test. The repeated-measures ANOVA, unlike the paired-samples t test, separates out the impact of individual differences from the effect of the explanatory variable, allowing the effect of the explanatory variable alone to be assessed.

How Wide Is the Confidence Interval and How Big Is the Effect?

The confidence interval for the difference between population means reveals how small or how large the difference between the population means might be. For the humidity data set, the confidence interval will tell, at a population level, how much hotter the temperature is perceived when the humidity rises. A confidence interval of any percentage could be calculated, but the most common is the 95% confidence interval. Equation 9.5 gives the formula for the 95% confidence interval for the difference between population means for the paired-samples t test.

Equation 9.5 Formula for the 95% Confidence Interval for the Difference Between Population Means for a Paired-Samples *t* Test

$$95\%\text{CI}_{\mu_{\text{Diff}}} = (M_1 - M_2) \pm (t_{cv} \times s_{M_D})$$

where $95\%\text{CI}_{\mu_{\text{Diff}}}$ = the 95% confidence interval for the difference between population means for a paired-samples t test
M_1 = the mean of one sample
M_2 = the mean of the other sample
t_{cv} = the critical value of t, two-tailed, $\alpha = .05$, $df = N - 1$
s_{M_D} = the standard error of the mean difference for the difference scores (Equation 9.3)

Applying this to her humidity data, Dr. Keim would calculate the 95% confidence interval as follows:

$$\begin{aligned} 95\%\text{CI}_{\mu_{\text{Diff}}} &= (82.50 - 75.00) \pm (2.571 \times 0.85) \\ &= 7.5000 \pm 2.1854 \\ &= \text{from } 5.3146 \text{ to } 9.6854 \\ &= [5.31, 9.69] \end{aligned}$$

Her confidence interval ranges from a lower limit of 5.31° to an upper limit of 9.69°. To make the interpretation more accessible, Dr. Keim has rounded each end to a whole number, making the confidence interval range from 5 to 10 degrees. In interpreting this confidence interval, Dr. Keim will pay attention to three points: (1) whether the confidence interval captures zero, (2) how close it comes to zero, and (3) how wide it is.

1. If a confidence interval captures zero, this means it is possible that the difference between the means of the populations is zero. When such happens, it is plausible that the null hypothesis is true. (This assumes that the confidence interval and the alpha level of the hypothesis test are synchronized.)

2. How close the confidence interval comes to zero provides information about the effect size in the larger population. If both ends of the confidence interval are close to zero, then the effect size is probably small. If both ends of the confidence interval are far from zero, then the effect size is probably large. If one end of the confidence interval is close to zero and the other end is far away, then the researcher will be left unsure of how strong or weak the effect actually is in the population.

3. The width of the confidence interval provides information about how precisely the effect can be specified in the population. Narrower confidence intervals are preferred because they give a more precise sense of the size of the effect in the population. If the confidence interval is wide, the researcher will usually recommend replicating with a larger sample size in order to determine the parameter value more precisely.

For the humidity study, the confidence interval, from 5 to 10 degrees, does not capture zero. This is to be expected as the null hypothesis was rejected. The confidence interval reiterates what is already known—it is unlikely, in the larger population, that there is no difference in how 76° is perceived in low humidity vs. how it is perceived in high humidity. Instead, in the population, the high-humidity condition is probably perceived as 5–10 degrees hotter, on average, than the same temperature at low humidity.

The second step, determining how small and how large the effect may be, takes more thought and expertise for the paired-samples *t* test than it did for the independent-samples *t* test. With an independent-samples *t* test, the researcher could calculate Cohen's *d* or r^2 and rely on Cohen's standards for small, medium, and large effects. The end of the confidence interval closer to zero (the lower limit) is 5 and the end farther away (the upper limit) is 10. In the larger population, will people perceive the temperature to be a lot hotter or a little hotter if they perceive 76° in conditions of high humidity as 5 degrees hotter than in low humidity? Dr. Keim believes this to be a meaningful effect. She reasons that, on a summer day, there is a noticeable difference in comfort level between being in an air-conditioned room at 71° and one at 76°, so a 5-degree difference is a meaningful one. If a 5-degree difference is meaningful, then a 10-degree difference is even more so. In understanding

what a 10-degree difference means, Dr. Keim thinks of how a 76° summer day feels pleasant and an 86° day feels hot.

Finally, she uses Equation 7.5 to calculate the width of the confidence interval. To avoid rounding error, she uses the real limits of the confidence interval, 5.31 to 9.69, not her rounded version of 5 to 10:

$$\begin{aligned} CI_W &= CI_{UL} - CI_{LL} \\ &= 9.69 - 5.31 \\ &= 4.38 \end{aligned}$$

The confidence interval is 4.38° wide. This seems sufficiently narrow to Dr. Keim. She would like to replicate the study in order to make sure the same effect is observed again, and she would like to increase the sample size in order to have a better sample, but she feels no need to replicate with a larger sample size in order to narrow the confidence interval. Given that both a 5-degree difference and a 10-degree difference seem meaningful and that the confidence interval is narrow, Dr. Keim is inclined to focus on the observed difference of 7.50 degrees in her interpretation.

Putting It All Together

Here is Dr. Keim's four-point interpretation in which she states: (1) what the study was about; (2) the main results; (3) what the results mean; and (4) her suggestions for future research. This interpretation is a little longer than previous ones. It takes Dr. Keim two paragraphs to say all that she wants to.

> The impact of humidity on perceived temperature was examined. Using a within-subjects design, six participants judged the temperature after being in a 76° room under conditions of low humidity and high humidity. In the low-humidity condition, they judged the temperature fairly accurately ($M = 75.00$°F), but they judged it as hotter ($M = 82.50$°F) under conditions of high humidity. This 7.50°F difference was statistically significant [$t(5) = 8.82$°F, $p < .05$]. Humidity appears to make people feel hotter. According to this study, people feel almost 8 degrees hotter under conditions of high humidity. This is a noticeable increase in perceived temperature and can move a person from feeling comfortable to being uncomfortable.
>
> There were several limitations to this study that can be rectified in future research. All participants were college students in the United States, people who have experience with heated and cooled environments. It would be interesting to see if the same effect were observed among people with less control over their indoor environments. A second limitation is that only one temperature, 76°, was tested. The effect of humidity on perceived temperature should be observed at both higher and lower temperatures. Finally, there were only six participants in this study. Replicating the study would increase confidence in the robustness of the finding that humidity level affects the perception of temperature.

Worked Example 9.2 For more practice with interpretation, let's return to the study where Dr. Althof followed patients with depression for six months following treatment. At the end of treatment, the mean depression level for the 16 participants was 14.00, and six months later it had climbed to 15.00. The mean difference score was 1.00 ($s_D = 2.31$). Dr. Althof had nondirectional hypotheses and a 5% chance of making a Type I error. t_{cv} was 2.131, s_{M_D} was calculated to be 0.58, and t was 1.72.

Was the null hypothesis rejected? The first step in interpretation is to determine if the null hypothesis is rejected. To do so, Dr. Althof substitutes the observed value of *t,* 1.72, into the decision rule he had generated in Step 4:

- Is $1.72 \leq -2.131$ *or* $1.72 \geq 2.131$?
- Or, is $-2.131 < 1.72 < 2.131$?

The second statement, is true: 1.72 falls between −2.131 and 2.131. Look at **Figure 9.6**, where it is clear that the value of the test statistic falls in the common zone. Dr. Althof has failed to reject the null hypothesis. There is not enough evidence to conclude that depression level changes, in either a positive or negative direction, in the six months after the end of psychodynamic therapy. In APA format, the results would be written as

$$t(15) = 1.72,\ p > .05$$

Remember, "$p > .05$" signifies that the result ($t = 1.72$) is an expected, or common, occurrence when the null hypothesis is true. It means that the null hypothesis was not rejected.

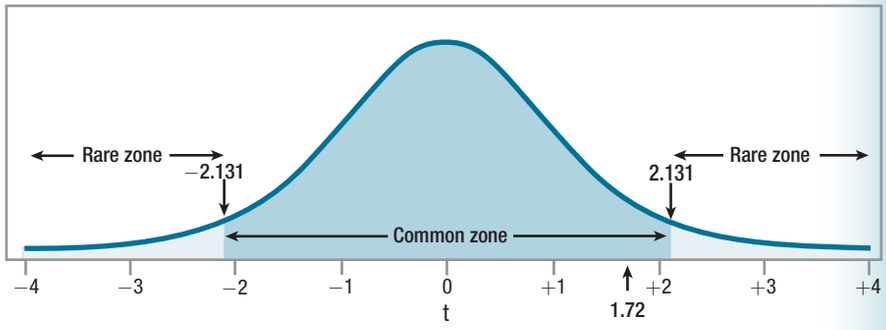

Figure 9.6 Observed Value of *t* from the Depression Follow-Up Study The observed value of *t,* 1.72, falls in the common zone. The null hypothesis is not rejected.

This failure to reject the null hypothesis can be taken as good news. Dr. Althof thinks of it as lack of evidence that relapse to depression occurs within six months of ending psychodynamic treatment for depression. The effectiveness of this treatment seems fairly long-lasting.

How wide is the confidence interval? How big is the effect? The second step in interpretation for a paired-samples *t* test is to use the confidence interval to evaluate the size of the effect. Dr. Althof used Equation 9.5 to calculate the confidence interval:

$$\begin{aligned} 95\%\text{CI}\mu_{\text{Diff}} &= (M_1 - M_2) \pm (t_{cv} \times s_{M_D}) \\ &= (15.00 - 14.00) \pm (2.131 \times 0.58) \\ &= 1.0000 \pm 1.2360 \\ &= \text{from } -0.2360 \text{ to } 2.2360 \\ &= [-0.24, 2.24] \end{aligned}$$

For the larger population of people with moderate to severe depression, this confidence interval tells Dr. Althof what the mean difference is in their level of depression from the end of treatment to six months later. It tells him that the mean difference could be anywhere from an average of being 2.24 points

more depressed six months after treatment to being 0.24 points *less* depressed six months after treatment. And, there is a 5% chance that this interval does not capture the actual mean difference.

There are three points to consider in interpreting a confidence interval: (1) whether it captures zero, (2) how close it comes to zero, and (3) how wide it is.

1. As expected, because Dr. Althof had failed to reject the null hypothesis, zero falls within this interval. This means that the value of zero is a viable option for how much difference occurs in the mean depression level from the end of treatment to six months later. In the larger population, there may be no loss in the effect of treatment in the six months following treatment.
2. One end of the confidence interval (−0.24) falls quite near zero and would be a trivial decrease in depression if it were true. The other end (2.24) doesn't fall far away from zero, considering that the depression scale ranges over 50 points. If the effect in the population were a mean increase of 2.24 depression points on a 50-point scale over six months, that is not much of an effect.
3. The confidence interval is not very wide, being a total of 2.48 points wide, from −0.24 to 2.24. Based on this, Dr. Althof does not feel a great need to replicate the study with a larger sample size. Note, however, that the vast majority of the confidence interval falls on one side of zero. If the sample size had been larger and the confidence interval narrower, then it would have failed to capture zero and Dr. Althof would have rejected the null hypothesis. This makes him worry that a Type II error may have occurred. Perhaps there is a small effect that was not found. For this reason, Dr. Althof is going to suggest replicating with a larger sample size.

This is a good opportunity to talk about the difference between *statistical significance* and *practical significance*. **Statistical significance** indicates that the observed difference between *sample* means is large enough to conclude that there is a difference between *population* means. It doesn't mean that the difference is a meaningful one.

Statistical significance is heavily influenced by sample size: the larger the sample size, the more likely it is that results will be statistically significant. If the sample size were increased in the depression follow-up study from 16 to 23, the results would be statistically significant and the confidence interval would range from 0.004 to 2.00, not capturing zero.

Think about the larger population. In the larger population, if the mean depression score were 14.00 at the end of treatment and 14.004 six months later, then the null hypothesis would be wrong and should be rejected. But a population difference of such a small amount, 0.004, would be so small as to be meaningless. It is of no practical significance. A result is of **practical significance** (also called **clinical significance**) if the size of the effect is large enough to make a real difference. Practical significance means that the explanatory variable has a meaningful impact on the outcome variable. A statistically significant result is no guarantee of a practically significant effect.

A statistically significant result is no guarantee of a practically significant effect.

Judging practical significance requires expertise in and familiarity with a specific area of research. Unless one has experience with a scale, it is hard to know how

meaningful a 2-point change on the Beck Depression Inventory. For now, the best option is to use Cohen's small, medium, and large effect sizes as an initial, and very rough, guide to practical and clinical significance.

A Common Question

Q Is it possible for a result to be practically significant but not statistically significant?

A No. When a result is not statistically significant, there is no evidence of a difference between the populations. The observed difference may be unique to the two samples in a study and wouldn't be found again. If an effect doesn't occur consistently, it can't be of practical use.

Putting it all together. Here's what Dr. Althof wrote for an interpretation. He followed the four-point plan for interpretations: (1) stating what the study was about, (2) giving the main results, (3) explaining them, and (4) making suggestions for future research.

A study was conducted that followed 16 people who had received psychodynamic therapy for severe to moderate depression for six months after treatment was complete. Their mean depression level was 14.00 at the end of treatment and 15.00 six months later. This 1-point increase in depression was not a statistically significant change [$t(15) = 1.72$, $p > .05$]. Thus, there is no evidence from this study to conclude that a relapse to depression occurs, for this population of patients, in the six months following treatment with psychodynamic therapy. The effects of psychodynamic therapy seem to be long-lasting.

To increase confidence in the robustness of these results, it would be a good idea to replicate this study. Additionally, it would be a good idea to add another form of therapy for depression to a study in order to see if lack of relapse after treatment ends is unique to psychodynamic therapy.

Application Demonstration

Here are some analyses based on a real data set to end our investigation of the paired-samples *t* test. A student, Kristin Brown, wondered whether there was a difference in cigarette smoking rates between men and women. Kristin found data reporting the rates of smoking for men and for women by state, and we'll use them to select 10 random states and determine if a difference exists in smoking rates between men and women. **Figure 9.7** shows the results for the 10 states.

Step 1 Pick a Test. The data are paired together by state—one rate for men and one rate for women. This seems sensible. Individual differences between states (such as whether the state produces tobacco or has active antismoking campaigns) might affect its smoking rate, but should affect both sexes in a state. The men and women are dependent samples, so this situation calls for a paired-samples *t* test.

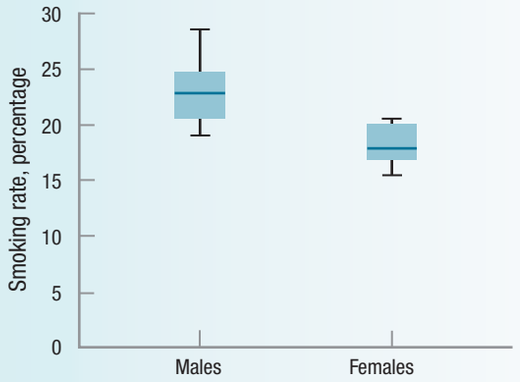

Figure 9.7 Smoking Rates for Males and Females in Sample of 10 States In this graph, it appears as if the smoking rate by state is higher for men than for women. However, to make sure the effect can't be explained by sampling error, hypothesis testing is needed.

Step 2 Check the Assumptions. Table 9.2 lists the assumptions.

- The random samples assumption is not violated. This is a random sample of 10 of the 50 states. The results can be generalized to all 50 states.
- The independence of observations assumption is not violated. The states were measured independently, so cases within a sample don't influence each other. Each state was only in each sample once.
- The normality assumption is not violated. It seems reasonable to assume that there is a normal distribution of difference scores in the larger population. In addition, if this assumption is wrong, it is robust to violations.

None of the assumptions (Table 9.2) was violated, so we can proceed with the paired-samples *t* test.

Step 3 List the Hypotheses. Hypotheses should be generated before any data are collected. The question being addressed is whether there is a difference in the smoking rates of men and women. The direction doesn't matter. This calls for a two-tailed test. The hypotheses are

$$H_0: \mu_{\text{Men}} = \mu_{\text{Women}}$$
$$H_1: \mu_{\text{Men}} \neq \mu_{\text{Women}}$$

Step 4 Set the Decision Rule. For a two-tailed test with alpha set at .05 and $10 - 1$ degrees of freedom, the critical value of t is ± 2.262. The decision rule is:

- If $t \leq -2.262$ *or* $t \geq 2.262$, reject the null hypothesis.
- If $-2.262 < t < 2.262$, fail to reject the null hypothesis.

Step 5 Calculate the Test Statistic. **Table 9.4** displays the means and standard deviations for men and women, and the difference scores for the 10 states. Using the standard deviation of the difference scores (1.98) and the number of pairs of cases (10), the standard error of the difference is calculated (Equation 9.3):

TABLE 9.4 Summary Data Comparing Smoking Rates of Men and Women

	Men	Women	Difference Score
Sample mean	23.00%	18.23%	4.77%
Sample standard deviation	3.05%	1.69%	1.98%

$$
\begin{aligned}
S_{M_D} &= \frac{s_D}{\sqrt{N}} \\
&= \frac{1.98}{\sqrt{10}} \\
&= \frac{1.98}{3.1623} \\
&= 0.6261 \\
&= 0.63
\end{aligned}
$$

The standard error of the mean difference is used to calculate the *t* value (Equation 9.4). Remember, it doesn't make a difference which mean is subtracted from the other mean, so subtract the mean for the women (18.23%) from the mean for the men (23.00%) to obtain a positive number:

$$
\begin{aligned}
t &= \frac{M_1 - M_2}{s_{M_D}} \\
&= \frac{23.00 - 18.23}{0.63} \\
&= \frac{4.77}{0.63} \\
&= 7.5714 \\
&= 7.57
\end{aligned}
$$

Step 6 Interpret the Results. *Was the null hypothesis rejected?*

- Is $7.57 \leq -2.262$ *or* $7.57 \geq 2.262$?
- Or, is $-2.262 < 7.57 < 2.262$?

7.57 is greater than or equal to 2.262, so the first statement is true. **Figure 9.8** shows that the observed value of *t,* 7.57, falls in the rare zone, so the null hypothesis is rejected.

The conclusion is that in the larger population of states, there is a difference in the smoking rate between men and women. Examining the sample means, 18.23% for the women and 23.00% for the men, leads to the conclusion that the smoking rate per state is higher for men than it is for women. In APA format, one would report

$$t(9) = 7.57,\ p < .05$$

How wide is the confidence interval? How big is the effect? To calculate the 95% confidence interval for the difference between population means,

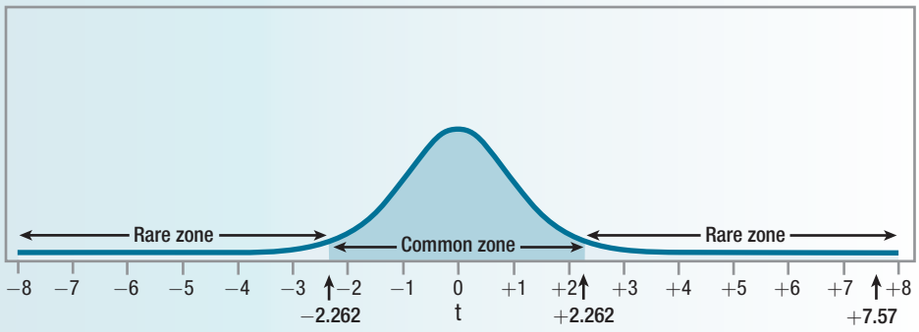

Figure 9.8 Observed Value of *t* from the Smoking Rate Study Note that the observed value of *t* falls in the rare zone. This means that the null hypothesis is rejected. It appears that the population mean for men differs from the population mean for women.

Equation 9.5 is used. Arrange the subtraction of one mean (the women at 18.23%) from the other (the men at 23.00%) to end up with a positive value. The other numbers in the equation, 2.262 and 0.63, are the critical value of *t* and the standard error of the difference, respectively:

$$\begin{aligned} 95\%\text{CI}\mu_{\text{Diff}} &= (M_1 - M_2) \pm (t_{cv} \times s_{M_D}) \\ &= (23.00 - 18.23) \pm (2.262 \times 0.63) \\ &= 4.7700 \pm 1.4251 \\ &= \text{from } 3.3449 \text{ to } 6.1951 \\ &= [3.34, 6.20] \end{aligned}$$

The confidence interval says that in the larger population of states, the mean male smoking rate is higher than the mean female smoking rate, by 3.34% to 6.20%.

- *Did the confidence interval capture zero?* As expected, because the null hypothesis was rejected, this confidence interval doesn't capture zero. It is unlikely that the difference in the smoking rates of men and women is zero.
- *How close does the confidence interval come to zero?* The lower end (3.34% difference in smoking rates) is not very close to zero and the upper end (6.20%) is a good distance from zero. It seems reasonable to conclude that the effect is of a moderate size.
- *How wide is the confidence interval?* At 2.86 percentage points, the confidence interval isn't too wide.

Putting it all together. Here's the four-point interpretation:

Data were analyzed from a study comparing the smoking rates of men to those of women for 10 randomly selected states. A mean of 23.00% of men in these states smoked, compared to 18.23% of the women. This 4.77 percentage points higher rate among the men was statistically significant [$t(9) = 7.57$, $p < .05$] and seems to show a reasonably strong effect of sex on the smoking rate. In the United States, men smoke at a higher rate. In future research, it would be interesting to see if the difference in rates between men and women has changed over time and to examine male/female differences in other countries.

Practice Problems 9.2

Apply Your Knowledge

9.04 Given $N = 46$, $M_1 = 23$, $M_2 = 32$, and $t = 3.67$, (a) write the results in APA format and (b) comment on the direction of the difference. Use $\alpha = .05$, two-tailed.

9.05 Given $N = 20$, $M_1 = 68$, $M_2 = 64$, and $t = 2.01$, (a) write the results in APA format and (b) comment on the direction of the difference. Use $\alpha = .05$, two-tailed.

9.06 Given $M_1 = 55$, $M_2 = 48$, $t_{cv} = 2.023$, and $s_{M_D} = 2.86$, calculate the 95%CIμ_{Diff}.

9.07 A tennis instructor compared two homework methods. She took beginning students at her tennis camp and matched them in pairs in terms of their tennis abilities. She then randomly assigned the players to two conditions. Those in the control condition had to practice for half an hour every day with another player. Those in the experimental condition had to practice against a wall for half an hour every day. After two weeks of practice, the instructor measured how often each player could hit targets in different locations on the tennis court. The higher the percentage, the better the player had become. Given the results below, interpret the study:

- $M_C = 48.00$, $M_E = 57.00$, $N = 12$, $s_D = 6.00$
- $s_{M_D} = 1.73$, $t = 5.20$, 95%CI$\mu_{\text{Diff}} = [5.19, 12.81]$

SUMMARY

Recognize the different types of dependent samples.

- In dependent samples, also called paired samples, cases are pairs of data points. Examples include (a) cases measured at two points in time (repeated measures, longitudinal, or pre-post designs); (b) cases measured in two different conditions (within-subjects designs); or (c) cases matched on some variable, so they are similar (matched-pairs designs).

Calculate and interpret a paired-samples *t* test.

- A paired-samples *t* test is used to compare the means of two dependent samples. Because paired-samples *t* tests control for the effects of individual differences, they have more statistical power, need a smaller sample size to reject H_0, and are useful for studying rare phenomena or those for which cases are hard to come by.
- To conduct a paired-samples *t* test, the assumptions must be met, hypotheses stated, and decision rule set before calculating the actual *t* value.
- In interpreting the results of a paired-samples *t* test, first apply the decision rule to determine if the null hypothesis was rejected, then write the results in APA format. If the null hypothesis was rejected, comment on the direction of the difference between the population means.
- Next, to find the effect size, calculate the 95% confidence interval for the difference between population means. Interpret the confidence interval, paying attention to (a) whether it captures zero, (b) how close it comes to zero, and (c) how wide it is. Do not calculate r^2 or *d*.
- Finally, in a four-point interpretation: (1) explain what the study was about, (2) present the main results, (3) interpret them, and (4) make suggestions for future research.

DIY

Is there a difference in the ages at which men and women get married? Has there been a change in this statistic over time?

Go online or get a newspaper from a recent Sunday and find the wedding section. Note the ages of each bride and groom. Are these paired data?

Check the assumptions and complete a paired-samples *t* test. Is there a difference? In what direction? What do the results tell us?

Want to do some extra work? Go to your school library and have the reference librarian help you find a Sunday newspaper from a decade or two ago. Collect and analyze the data for the ages of the brides and grooms back then. Has a change in the age difference occurred over the years?

KEY TERMS

clinical significance (or practical significance) – the size of the effect is large enough to say the independent variable has a meaningful impact on clinical outcome.

individual differences – attributes that vary from case to case.

longitudinal research (or repeated-measures design) – a study in which the same participants are measured at two or more points in time.

matched pairs – participants are grouped into sets of two based on their being similar on potential confounding variables.

paired-samples t test – hypothesis test used to compare the means of two dependent samples; also known as dependent-samples *t* test, correlated-samples *t* test, related-samples *t* test, matched-pairs *t* test, within-subjects *t* test, or repeated-measures *t* test.

practical significance (or clinical significance) – the size of the effect is large enough to say the independent variable has a meaningful impact on the dependent variable (or the clinical outcome).

pre-post design – participants are measured on the dependent variable before and after an intervention.

repeated-measures design (or longitudinal research) – a study in which the same participants are measured at two or more points in time.

standard error of the mean difference for difference scores – the standard deviation of the sampling distribution of difference scores, abbreviated s_{M_D}; used as the denominator in the paired-samples *t* test equation.

statistical significance – the observed difference between sample means is large enough to conclude that it represents a difference between population means.

within-subjects design – the same participants are measured in two different situations or under two different conditions.

CHAPTER EXERCISES

Review Your Knowledge

9.01 The independent-samples *t* test is used to compare means of two ____ samples; the ____ is used to compare means of two dependent samples.

9.02 The cases in the two samples in a paired-samples *t* test may be paired because they are the same cases measured at two different points in ____ or under two different conditions.

9.03 With matched samples, cases are paired together in order to control for ____ variables.

9.04 Two other names for a paired-samples *t* test are ____ and ____.

9.05 An advantage of a paired-samples *t* test is that it controls for ____.

9.06 An individual difference is an attribute, such as intelligence or weight, that varies from ____.

9.07 Paired-samples *t* tests are more powerful statistically. This means they have a ____ probability of being able to reject the null hypothesis if the null hypothesis was false.

9.08 Paired-samples *t* tests are good to use when ____ are hard to come by.

9.09 Paired-samples *t* tests can be used to determine if a difference between two sample means is statistically significant, or if it could be explained by ____.

9.10 If the random samples assumption for a paired-samples *t* test is not violated, the results can be ____ to the population.

9.11 The independence assumption for a paired-samples *t* test refers to independence ____ a sample, not ____ samples.

9.12 The normality assumption for a paired-samples *t* test refers to the ____ being normally distributed.

9.13 The hypotheses for a paired-samples *t* test are statements about the two ____, not the two ____.

9.14 The null hypothesis for a nondirectional paired-samples *t* test says there is ____ between population means.

9.15 The alternative hypothesis for a two-tailed, paired-samples *t* test says that the difference between population means could be positive or ____, large or ____.

9.16 The decision rule for a paired-samples *t* test compares the ____ value of *t* to the ____ value of *t*.

9.17 The degrees of freedom for a paired-samples *t* test are calculated by subtracting ____ from the number of pairs.

9.18 The default option for a paired-samples *t* test is to set alpha at ____ and to do a ____-tailed test.

9.19 The abbreviation for the standard error of the mean difference scores in a paired-samples *t* test is ____.

9.20 The value of the test statistic for a paired-samples *t* test is obtained by dividing ____ by the standard error of the mean difference of the difference scores.

9.21 It *does/does not* make a difference which sample mean is subtracted from the other mean when calculating a paired-samples *t* value.

9.22 If $t \geq t_{cv}$, one would ____ the null hypothesis.

9.23 If results are written in APA format as p ____ .05, then the researcher has failed to reject the null hypothesis.

9.24 If one fails to reject the null hypothesis, there is no need to comment on the ____ of the difference between the population means.

9.25 The direction of the difference between the population means is determined by comparing the two ____ means.

9.26 If the degrees of freedom for a paired-samples *t* test are 11, then there were ____ pairs of data.

9.27 Cohen's *d*, when calculated for a paired-samples *t* test, includes the effect of ____ as well as the effect of the explanatory variable on the ____.

9.28 r^2 *should/should not* be used as an effect size for a paired-samples *t* test.

9.29 To measure effect size for a paired-samples *t* test, use the ____.

9.30 Whether the 95% confidence interval captures zero for a two-tailed, paired-samples *t* test with alpha set at .05 provides information about whether the null hypothesis is ____.

9.31 If the confidence interval for a paired-samples *t* test does not capture zero, but both ends of it are close to zero, then the size of the effect is probably ____.

9.32 When the confidence interval for a paired-samples *t* test is ____, one has fairly precisely specified the size of the effect in the population.

Apply Your Knowledge

Picking a test

9.33 A sensory psychologist had participants rate the taste of two coffees, caffeinated and decaffeinated versions of the same brand. Each participant rated both types of coffee. What statistical test should the psychologist use to see if caffeinated and decaffeinated coffees differ in taste?

9.34 A nutritionist wanted to find out if coffee and tea, as served in restaurants, differed in caffeine content. She went to 30 restaurants. In 15 randomly selected restaurants, she ordered coffee; in the other 15 restaurants, she ordered tea. What statistical test should the nutritionist use to see if coffee and tea differ in mean caffeine content?

9.35 A researcher for a health magazine compared the mean caffeine content for a sample of coffees served at coffee houses to the USDA standard for the mean amount of caffeine in a cup of coffee. What statistical test should she use?

9.36 A developmental psychologist wondered if birth order had an impact on academic performance. She found families with two children and compared the mean high school GPA of first-born children to second-born children. What statistical test should she use?

Checking the assumptions

9.37 A group of high school students were in the same math, English, social studies, and science classes. A researcher monitored how much time, during a week, they spent online in school activities (doing homework) vs. how much time they spent online in social activities (on Facebook, chatting, playing multiplayer games). The researcher planned to use a paired-samples *t* test to analyze the data to see if there was a difference in time spent on the two activities. Based on the assumptions, is it OK to proceed with the test?

9.38 Elementary school teachers who did and did not have children of their own were matched in terms of the number of years of experience they had teaching. They were then asked how many minutes of homework a child should complete per night. The researcher planned to use a paired-samples *t* test to determine if having children of one's own was related to this response. Based on the assumptions for the test, is it OK to proceed?

Writing the hypotheses

9.39 Write the hypotheses for a paired-samples *t* test for Exercise 9.37.

9.40 Write the hypotheses for a paired-samples *t* test for Exercise 9.38.

Calculating difference scores

9.41 Calculate difference scores for the following pairs of scores: 5, 10; 7, 3; 6, 8; 4, 3; 7, 8.

9.42 Calculate difference scores for the following pairs of scores: 12, 13; 14, 12; 7, 4; 2, 4; 8, 6.

Calculating degrees of freedom

9.43 A researcher plans to use a paired-samples *t* test for two dependent samples, each with 10 cases. How many degrees of freedom are there?

9.44 If a consumer researcher compares the mean price of 25 items purchased at one store to the mean price for the same 25 items purchased at a second store, how many degrees of freedom are there?

Finding t_{cv} for a two-tailed test with $\alpha = .05$

9.45 If there are 19 cases in a paired-samples *t* test, what is t_{cv}?

9.46 If there are 33 pairs of data in a paired-samples *t* test, what is t_{cv}?

9.47 If $N = 118$ for a paired-samples *t* test, what is t_{cv}?

9.48 Given 2,012 cases in a paired-samples *t* test, what is t_{cv}?

Calculating standard error of the mean difference

9.49 Given $s_D = 2.57$ and $N = 17$, calculate s_{M_D}.

9.50 Given $s_D = 12.78$ and $N = 49$, calculate s_{M_D}.

Given s_{M_D}, finding t

9.51 If $M_1 = 18$, $M_2 = 14$, and $s_{M_D} = 5.34$, what is t?

9.52 If $M_1 = -12$, $M_2 = -15$, and $s_{M_D} = 2.81$, what is t?

Calculating t

9.53 Given $M_1 = 25$, $M_2 = 28$, $N = 5$, and $s_D = 7.00$, calculate t.

9.54 Given $M_1 = -7$, $M_2 = -9$, $N = 28$, and $s_D = 2.90$, calculate t.

Implementing the decision rule

9.55 If $t = 2.30$ and $t_{cv} = \pm 2.017$, (a) draw a sampling distribution of t, marking t and t_{cv}, and label the rare and common zones, then (b) report whether or not the null hypothesis was rejected.

9.56 If $t = 1.65$ and $t_{cv} = \pm 2.145$, (a) draw a sampling distribution of t, marking t and t_{cv}, and labeling the rare and common zones, then (b) report whether or not the null hypothesis was rejected.

Writing results in APA format (use $\alpha = .05$, two-tailed)

9.57 Given $N = 5$ and $t = 3.211$, write the results of this paired-samples t test in APA format.

9.58 Given $N = 27$ and $t = 2.033$, write the results of this paired-samples t test in APA format.

9.59 Given $N = 69$ and $t = 1.994$, write the results of this paired-samples t test in APA format.

9.60 Given $N = 181$ and $t = 1.981$, write the results of this paired-samples t test in APA format.

Determining the direction of the difference

9.61 Given these results, comment on the direction of the difference between the population means: $M_1 = 72$, $M_2 = 73$, $t(26) = 2.08$, $p < .05$.

9.62 Given these results, comment on the direction of the difference between the population means: $M_1 = 17$, $M_2 = 24$. $t(35) = 2.01$, $p > .05$.

9.63 Given these results, comment on the direction of the difference between the population means: $M_1 = 50$, $M_2 = 53$, $t(17) = 1.54$, $p > .05$.

9.64 Given these results, comment on the direction of the difference between the population means: $M_1 = 28$, $M_2 = 31$, $t(72) = 7.42$, $p < .05$.

Calculating a confidence interval

9.65 Given the following, calculate the 95% confidence interval for the difference between population means: $M_1 = 108$, $M_2 = 100$, $t_{cv} = 2.052$, and $s_{M_D} = 5.00$.

9.66 Given the following, calculate the 95% confidence interval for the difference between population means: $M_1 = 40$, $M_2 = 50$, $t_{cv} = 2.010$, and $s_{M_D} = 2.44$.

Interpreting confidence intervals (if necessary, assume the test was two-tailed with $\alpha = .05$)

9.67 Based on the confidence interval 18.00 to 27.00, decide if the null hypothesis was rejected for the paired-samples t test.

9.68 Based on the confidence interval −0.40 to 0.50, decide if the null hypothesis was rejected for the paired-samples t test.

Interpreting the results of a paired-samples t test

9.69 A sleep therapist wanted to see if an herbal tea advertised as a sleep aid really worked. He located 46 people with sleep problems and matched them into pairs on the basis of (a) how long they had suffered from insomnia, (b) how long it usually took them to go to sleep at night, (c) how much sleep onset anxiety they experienced, and (d) how suggestible they were. He then randomly assigned one person from each pair to drink the tea at bedtime (the experimental group), while the control group went to sleep as they normally did. He used an EEG to measure the minutes to sleep onset (the fewer the minutes to sleep onset, the better). He found $M_C = 21.20$, $M_E = 19.70$, $s_D = 5.47$, $s_{M_D} = 1.14$, $t = 1.32$, and 95%CIμ_{Diff} = from −3.86 to 0.86. Write a four-point interpretation.

9.70 A sportswriter was curious if football teams gained more yards rushing (the control condition) or passing (the experimental condition). She randomly selected nine teams and calculated the mean yards gained per game through rushing (101) and through passing (221). s_D was 27.99. She calculated $s_{M_D} = 9.33$, $t = 12.86$, and 95%CIμ_{Diff} = from 98.49 to 141.51. Write a four-point interpretation.

Completing all six steps of hypothesis testing

9.71 A dermatologist compared a new treatment for athlete's foot (the experimental condition) to the standard treatment (the control condition). He tracked down 30 people with athlete's foot on both feet and, for each participant, randomly assigned one foot to receive the new treatment and the other foot to receive the standard treatment. After three weeks of treatment, he measured the percentage of reduction in symptoms (the larger the number, the better the outcome). He found $M_E = 88$, $M_C = 72$, and $s_D = 8.65$. Analyze and interpret.

9.72 A college president wanted to know how 10-year-after-graduation salaries for academic majors (English, psychology, math, etc.) compared to salaries for career-oriented majors (business, engineering, computer science, etc.). She matched 84 academic majors at her college with 84 career-oriented majors on the basis of SAT scores and GPA. She found $M_{\text{Academic}} = \$59{,}250$, $M_{\text{Career}} = \$61{,}000$, $s_D = 9{,}500$. Analyze and interpret.

Expand Your Knowledge

9.73 If $N = 24$ and $s_{M_D} = 3.56$, for which situation is the 95% confidence interval for the difference between population means the widest?

a. $M_C = 0$ and $M_E = 2$
b. $M_C = 0$ and $M_E = 5$
c. $M_C = 0$ and $M_E = 10$
d. $M_C = 0$ and $M_E = -2$
e. All confidence intervals are equally wide.
f. Not enough information was presented to answer this question.

9.74 If $M_C = 5$ and $M_E = 10$, for which situation is the 95% confidence interval for the difference between population means the widest?

a. $N = 5$ and $s_{M_D} = 5.28$
b. $N = 10$ and $s_{M_D} = 5.28$
c. $N = 20$ and $s_{M_D} = 5.28$
d. All confidence intervals are equally wide.
e. Not enough information was presented to answer this question.

9.75 If $M_C = 51$ and $M_E = 43$, for which situation is the 95% confidence interval for the difference between population means the widest?

a. $N = 28$ and $s_{M_D} = 2.00$
b. $N = 28$ and $s_{M_D} = 4.00$
c. $N = 28$ and $s_{M_D} = 6.00$
d. $N = 28$ and $s_{M_D} = 9.00$
e. All confidence intervals are equally wide.
f. Not enough information was presented to answer this question.

9.76 If the 95% confidence interval for the difference between population means ranges from 1.00 to 9.00, what is $M_E - M_C$?

9.77 If the 95% confidence interval for the difference between population means ranges from −5.00 to 1.00 and $s_{M_D} = 4.00$, what is t?

9.78 A nurse at a health clinic wanted to see if its ear thermometers and oral thermometers registered the same body temperatures. She selected six healthy staff members and took their temperatures with both thermometers. Apply the six steps of hypothesis testing to the data collected. Is body temperature measured similarly by ear thermometers and oral thermometers?

Body Temperature, Measured in °F						
Case	1	2	3	4	5	6
Ear	97.8	98.6	98.9	97.9	99.0	98.2
Oral	97.4	97.9	98.3	97.4	98.1	97.7

SPSS

When entering data for a paired-samples *t* test into the data editor in SPSS, each pair of scores is on a row and the two data points are in separate columns. This means that the data for the humidity study would be entered as shown in **Figure 9.9**. Note that there is no column for a difference score since SPSS will calculate that internally and automatically.

The command for a paired-samples *t* test, as shown in **Figure 9.10**, is found under "Analyze" and then "Compare Means."

	low_humid	high_humid
1	76	81
2	80	90
3	78	85
4	72	82
5	76	82
6	68	75

Figure 9.9 SPSS Data Entry for a Paired-Samples *t* Test Note that each case has its own row and that the variables (high_humid and low_humid) are in the columns. (Source: SPSS)

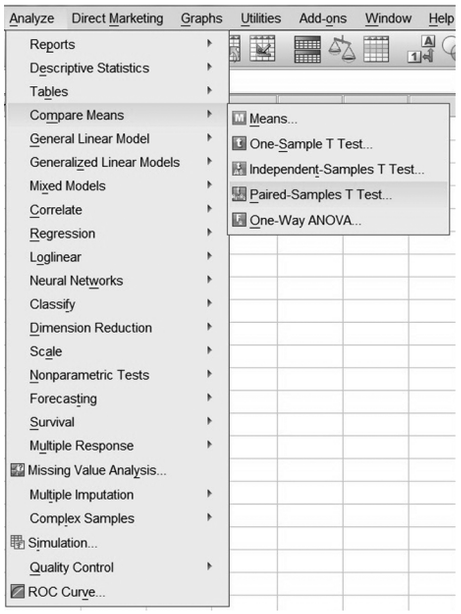

Figure 9.10 SPSS Paired-Samples *t* Test Command The paired-samples *t* test command in SPSS falls among the "Compare Means" commands. (Source: SPSS)

The commands for completing the paired-samples *t* test in SPSS are shown in **Figure 9.11**.

The SPSS output for the paired-samples *t* test is shown in **Figure 9.12**. The first of the three output tables provides descriptive statistics for the two samples. The second table reports the correlation between the two variables. Correlations will be covered in Chapter 13, so ignore this output for now.

The meat of the output appears after the first two tables. SPSS calculated the *t* value as −8.859, whereas in the chapter discussion it is 8.82. The sign is different because SPSS subtracted the two sample means in a different order. SPSS also carries more decimal places in its calculations, so its answer is more exact.

SPSS reports exact significance levels, seen in the last column of the final table. The value for this *t* test is .000. This means that if the null hypothesis is true, a result like the one found here, where $t = -8.859$, is a rare result; it has a probability of less than .001 of occurring. In percentage terms, a result such as this happens less than 0.1% of the time when the null hypothesis is true. The likelihood of a Type I error is very low.

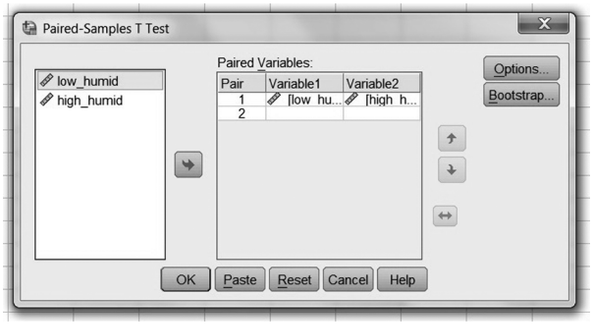

Figure 9.11 Picking the Variables to Analyze with a Paired-Samples *t* Test in SPSS The arrow button was used to move the two variables from the box on the left into the box labeled "Paired Variables." Notice how SPSS puts together the variables as a pair. (Source: SPSS)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	low_humid	75.00	6	4.336	1.770
	high_humid	82.50	6	4.930	2.012

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	low_humid & high_humid	6	.908	.012

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	low_humid - high_humid	-7.500	2.074	.847	-9.676	-5.324	-8.859	5	.000

Figure 9.12 SPSS Output for a Paired-Samples *t* Test The top table provides descriptive statistics for the two samples. The bottom table provides the *t* value, the significance level, and the 95% confidence interval for the difference between the population means. (Ignore the middle table for now.) (Source: SPSS)

For our purposes, if the exact significance value for a test as reported by SPSS is less than or equal to .05, then the null hypothesis is rejected. If the exact significance value for a test is greater than .05, then fail to reject the null hypothesis. Since .001 is less than or equal to .05, the null hypothesis is rejected.

APA format prefers reporting exact significance levels when possible. Here, the exact significance level is reported as .000. APA format would report this as $t(5) = -8.86$, $p \leq .001$.

The final thing to note is that SPSS calculates the 95% confidence interval for the difference between population means. Here, SPSS reports the difference between population means as ranging from −9.676 to −5.324. Again, the sign is different because of the order in which SPSS subtracted one mean from the other. Otherwise, the numbers calculated by SPSS and in the chapter (5.31 to 9.69) differ slightly because of the number of decimal places carried. Again, SPSS offers the more exact answer because it carries more decimal places.