CHAPTER

11

The *t* Test for Two Related Samples

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Introduction to the *t* statistic (Chapter 9)
  - Estimated standard error
  - Degrees of freedom
  - *t* Distribution
  - Hypothesis tests with the *t* statistic
- Independent-measures design (Chapter 10)

Preview

Swearing is a common, almost reflexive, response to pain. Whether you knock your shin into the edge of a coffee table or smash your thumb with a hammer, most of us respond with a streak of obscenities. One question, however, is whether swearing focuses attention on the pain and, thereby, increases its intensity, or serves as a distraction that reduces pain. To address this issue, Stephens, Atkins, and Kingston (2009) conducted an experiment comparing swearing with other responses to pain. In the study, participants were asked to place one hand in icy cold water for as long as they could bear the pain. Half of the participants were told to repeat their favorite swear word over and over for as long as their hands were in the water. The other half repeated a neutral word. The researchers recorded how long each participant was able to tolerate the ice water. After a brief rest, the two groups switched words and repeated the ice water plunge. Thus, all the participants experienced both conditions (swearing and neutral) with half swearing on their first plunge and half on their second. The results clearly showed that swearing significantly increased the average amount of time that participants could tolerate the pain.

The Problem: In the previous chapter, we introduced a statistical procedure for evaluating the mean difference between two sets of data (the independent-measures *t* statistic). However, the independent-measures *t* statistic is intended for research situations involving two separate and independent samples. You should realize that the two sets of scores in the swearing study are not independent samples. In fact, the same group individuals participated in both of the treatment conditions. What is needed is a new statistical analysis for comparing two means that are both obtained from the same group of participants.

The Solution: In this chapter, we introduce the *repeated-measures* t *statistic,* which is used for hypothesis tests evaluating the mean difference between two sets of scores obtained from the same group of individuals. As you will see, however, this new *t* statistic is very similar to the original *t* statistic that was introduced in Chapter 9.

Finally, we should note that researchers often have a choice when they are planning a research study that compares two different treatment conditions. Specifically, a researcher may choose to use two separate groups of participants, one for each of the treatments, or a researcher may choose to use one group and measure each individual in both of the treatment conditions. Later in this chapter, we take a closer look at the differences between these two research designs and discuss the advantages and disadvantages of each.

11.1 INTRODUCTION TO REPEATED-MEASURES DESIGNS

In the previous chapter, we introduced the independent-measures research design as one strategy for comparing two treatment conditions or two populations. The independent-measures design is characterized by the fact that two separate samples are used to obtain the two sets of scores that are to be compared. In this chapter, we examine an alternative strategy known as a *repeated-measures design,* or a *within-subjects design.* With a repeated-measures design, two separate scores are obtained for each individual in the sample. For example, a group of patients could be measured before therapy and then measured again after therapy. Or, response time could be measured in a driving simulation task for a group of individuals who are first tested when they are sober and then tested again after two alcoholic drinks. In each case, the same variable is being measured twice for the same set of individuals; that is, we are literally repeating measurements on the same sample.

DEFINITION

A **repeated-measures design,** or a **within-subject design,** is one in which the dependent variable is measured two or more times for each individual in a single sample. The same group of subjects is used in all of the treatment conditions.

The main advantage of a repeated-measures study is that it uses exactly the same individuals in all treatment conditions. Thus, there is no risk that the participants in one treatment are substantially different from the participants in another. With an independent-measures design, on the other hand, there is always a risk that the results are biased because the individuals in one sample are systematically different (smarter, faster, more extroverted, and so on) than the individuals in the other sample. At the end of this chapter, we present a more detailed comparison of repeated-measures studies and independent-measures studies, considering the advantages and disadvantages of both types of research.

THE MATCHED-SUBJECTS DESIGN

Occasionally, researchers try to approximate the advantages of a repeated-measures design by using a technique known as *matched subjects*. A matched-subjects design involves two separate samples, but each individual in one sample is matched one-to-one with an individual in the other sample. Typically, the individuals are matched on one or more variables that are considered to be especially important for the study. For example, a researcher studying verbal learning might want to be certain that the two samples are matched in terms of IQ and gender. In this case, a male participant with an IQ of 120 in one sample would be matched with another male with an IQ of 120 in the other sample. Although the participants in one sample are not *identical* to the participants in the other sample, the matched-subjects design at least ensures that the two samples are equivalent (or matched) with respect to some specific variables.

DEFINITION

In a **matched-subjects design,** each individual in one sample is matched with an individual in the other sample. The matching is done so that the two individuals are equivalent (or nearly equivalent) with respect to a specific variable that the researcher would like to control.

Of course, it is possible to match participants on more than one variable. For example, a researcher could match pairs of subjects on age, gender, race, and IQ. In this case, for example, a 22-year-old white female with an IQ of 115 who was in one sample would be matched with another 22-year-old white female with an IQ of 115 in the second sample. The more variables that are used, however, the more difficult it is to find matching pairs. The goal of the matching process is to simulate a repeated-measures design as closely as possible. In a repeated-measures design, the matching is perfect because the same individual is used in both conditions. In a matched-subjects design, however, the best you can get is a degree of match that is limited to the variable(s) that are used for the matching process.

In a repeated-measures design or a matched-subjects design comparing two treatment conditions, the data consist of two sets of scores, which are grouped into sets of two, corresponding to the two scores obtained for each individual or each matched pair of subjects (Table 11.1). Because the scores in one set are directly related, one-to-one, with the scores in the second set, the two research designs are statistically equivalent and share the common name *related-samples* designs (or correlated-samples designs). In this chapter, we focus our discussion on repeated-measures designs because they are overwhelmingly the more common example of related-samples designs. However, you should realize that the statistical techniques used for repeated-measures studies also can be applied directly to data from matched-subjects studies. We should also note that a matched-subjects study occasionally is called a *matched samples design*, but the subjects in the samples must be matched one-to-one before you can use the statistical techniques in this chapter.

TABLE 11.1

An example of the data from a repeated-measures or a matched-subjects study using $n = 5$ participants (or matched pairs).

Participant or Matched Pair	First Score	Second Score	
#1	12	15	←The 2 scores for
#2	10	14	one participant or
#3	15	17	one matched pair
#4	17	17	
#5	12	18	

Now we examine the statistical techniques that allow a researcher to use the sample data from a repeated-measures study to draw inferences about the general population.

11.2 THE *t* STATISTIC FOR A REPEATED-MEASURES RESEARCH DESIGN

The *t* statistic for a repeated-measures design is structurally similar to the other *t* statistics we have examined. As we shall see, it is essentially the same as the single-sample *t* statistic covered in Chapter 9. The major distinction of the related-samples *t* is that it is based on *difference scores* rather than raw scores (*X* values). In this section, we examine difference scores and develop the *t* statistic for related samples.

DIFFERENCE SCORES: THE DATA FOR A REPEATED-MEASURES STUDY

Many over-the-counter cold medications include the warning "may cause drowsiness." Table 11.2 shows an example of data from a study that examines this phenomenon. Note that there is one sample of $n = 4$ participants, and that each individual is measured twice. The first score for each person (X_1) is a measurement of reaction time before the medication was administered. The second score (X_2) measures reaction time 1 hour after taking the medication. Because we are interested in how the medication affects reaction time, we have computed the difference between the first score and the second score for each individual. The *difference scores*, or *D* values, are shown in the last column of the table. Notice that the difference scores measure the amount of change

TABLE 11.2

Reaction-time measurements taken before and after taking an over-the-counter cold medication.

Person	Before Medication (X_1)	After Medication (X_2)	Difference *D*
A	215	210	−5
B	221	242	21
C	196	219	23
D	203	228	25
			$\Sigma D = 64$

Note that M_D is the mean for the sample of *D* scores.

$$M_D = \frac{\Sigma D}{n} = \frac{64}{4} = 16$$

in reaction time for each person. Typically, the difference scores are obtained by subtracting the first score (before treatment) from the second score (after treatment) for each person:

$$\text{difference score} = D = X_2 - X_1 \tag{11.1}$$

Note that the sign of each D score tells you the direction of the change. Person A, for example, shows a decrease in reaction time after taking the medication (a negative change), but person B shows an increase (a positive change).

The sample of difference scores (D values) serves as the sample data for the hypothesis test and all calculations are done using the D scores. To compute the t statistic, for example, we use the number of D scores (n) as well as the sample mean (M_D) and the value of SS for the sample of D scores.

THE HYPOTHESES FOR A RELATED-SAMPLES STUDY

The researcher's goal is to use the sample of difference scores to answer questions about the general population. In particular, the researcher would like to know whether there is any difference between the two treatment conditions for the general population. Note that we are interested in a population of *difference scores*. That is, we would like to know what would happen if every individual in the population were measured in two treatment conditions (X_1 and X_2) and a difference score (D) were computed for everyone. Specifically, we are interested in the mean for the population of difference scores. We identify this population mean difference with the symbol μ_D (using the subscript letter D to indicate that we are dealing with D values rather than X scores).

As always, the null hypothesis states that, for the general population, there is no effect, no change, or no difference. For a repeated-measures study, the null hypothesis states that the mean difference for the general population is zero. In symbols,

$$H_0: \mu_D = 0$$

Again, this hypothesis refers to the mean for the entire population of difference scores. Figure 11.1(a) shows an example of a population of difference scores with a

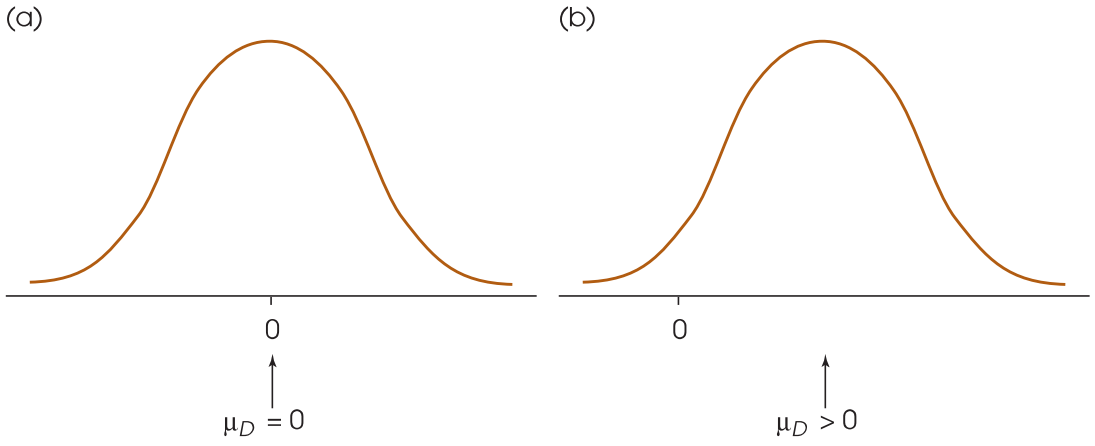


FIGURE 11.1

(a) A population of difference scores for which the mean is $\mu_D = 0$. Note that the typical difference score (D value) is not equal to zero. (b) A population of difference scores for which the mean is greater than zero. Note that most of the difference scores are also greater than zero.

mean of $\mu_D = 0$. Although the population mean is zero, the individual scores in the population are not all equal to zero. Thus, even when the null hypothesis is true, we still expect some individuals to have positive difference scores and some to have negative difference scores. However, the positives and negatives are unsystematic and in the long run balance out to $\mu_D = 0$. Also note that a sample selected from this population probably will not have a mean exactly equal to zero. As always, there will be some error between a sample mean and the population mean, so even if $\mu_D = 0$ (H_0 is true), we do not expect M_D to be exactly equal to zero.

The alternative hypothesis states that there is a treatment effect that causes the scores in one treatment condition to be systematically higher (or lower) than the scores in the other condition. In symbols,

$$H_1: \mu_D \neq 0$$

According to H_1, the difference scores for the individuals in the population tend to be systematically positive (or negative), indicating a consistent, predictable difference between the two treatments.

Figure 11.1(b) shows an example of a population of difference scores with a positive mean difference, $\mu_D > 0$. This time, most of the individuals in the population have difference scores that are greater than zero. A sample selected from this population will contain primarily positive difference scores and will probably have a mean difference that is greater than zero, $M_D > 0$. See Box 11.1 for further discussion of H_0 and H_1.

THE *t* STATISTIC FOR RELATED SAMPLES

Figure 11.2 shows the general situation that exists for a repeated-measures hypothesis test. You may recognize that we are facing essentially the same situation that we encountered in Chapter 9. In particular, we have a population for which the mean and the standard deviation are unknown, and we have a sample that will be used to test a hypothesis about the unknown population. In Chapter 9, we introduced the single-sample *t* statistic, which allowed us to use a sample mean as a basis for testing hypotheses about an unknown

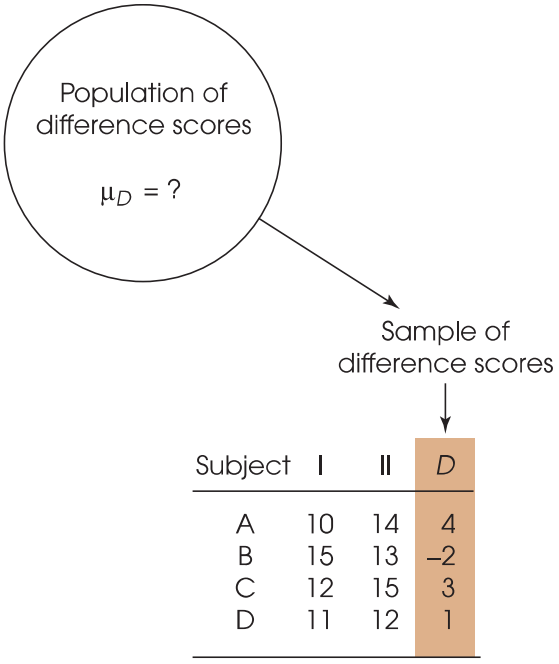
BOX 11.1 ANALOGIES FOR H_0 AND H_1 IN THE REPEATED-MEASURES TEST

An Analogy for H_0: Intelligence is a fairly stable characteristic; that is, you do not get noticeably smarter or dumber from one day to the next. However, if we gave you an IQ test every day for a week, we probably would get seven different numbers. The day-to-day changes in your IQ score are caused by random factors such as your health, your mood, and your luck at guessing answers you do not know. Some days your IQ score is slightly higher, and some days it is slightly lower. On average, the day-to-day changes in IQ should balance out to zero. This is the situation that is predicted by the null hypothesis for a repeated-measures test. According to H_0, any changes that occur either for an individual or for a sample are just due to chance, and in the long run, they will average out to zero.

An Analogy for H_1: On the other hand, suppose that we evaluate your performance on a new video game by measuring your score every day for a week. Again, we probably will find small differences in your scores from one day to the next, just as we did with the IQ scores. However, the day-to-day changes in your game score will not be random. Instead, there should be a general trend toward higher scores as you gain more experience with the new game. Thus, most of the day-to-day changes should show an increase. This is the situation that is predicted by the alternative hypothesis for the repeated-measures test. According to H_1, the changes that occur are systematic and predictable and will not average out to zero.

FIGURE 11.2

A sample of $n = 4$ people is selected from the population. Each individual is measured twice, once in treatment I and once in treatment II, and a difference score, *D*, is computed for each individual. This sample of difference scores is intended to represent the population. Note that we are using a sample of difference scores to represent a population of difference scores. Note that the mean for the population of difference scores is unknown. The null hypothesis states that for the general population there is no consistent or systematic difference between the two treatments, so the population mean difference is $\mu_D = 0$.



Subject	I	II	*D*
A	10	14	4
B	15	13	–2
C	12	15	3
D	11	12	1

population mean. This *t*-statistic formula is used again here to develop the repeated-measures *t* test. To refresh your memory, the single-sample *t* statistic (Chapter 9) is defined by the formula

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

In this formula, the sample mean, *M*, is calculated from the data, and the value for the population mean, μ, is obtained from the null hypothesis. The estimated standard error, s_M, is also calculated from the data and provides a measure of how much difference it is reasonable to expect between a sample mean and the population mean.

For the repeated-measures design, the sample data are difference scores and are identified by the letter *D*, rather than *X*. Therefore, we use *D*s in the formula to emphasize that we are dealing with difference scores instead of *X* values. Also, the population mean that is of interest to us is the population mean difference (the mean amount of change for the entire population), and we identify this parameter with the symbol μ_D. With these simple changes, the *t* formula for the repeated-measures design becomes

As noted earlier, the repeated-measures *t* formula is also used for matched-subjects designs.

$$t = \frac{M_D - \mu_D}{s_{M_D}} \tag{11.2}$$

In this formula, the *estimated standard error for* M_D, s_{M_D}, is computed in exactly the same way as it is computed for the single-sample *t* statistic. To calculate the estimated standard error, the first step is to compute the variance (or the standard deviation) for the sample of *D* scores.

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df} \quad \text{or} \quad s = \sqrt{\frac{SS}{df}}$$

The estimated standard error is then computed using the sample variance (or sample standard deviation) and the sample size, n.

$$s_{M_D} = \sqrt{\frac{s^2}{n}} \quad \text{or} \quad s_{M_D} = \frac{s}{\sqrt{n}} \tag{11.3}$$

Notice that all of the calculations are done using the difference scores (the D scores) and that there is only one D score for each subject. With a sample of n subjects, there are exactly n D scores, and the t statistic has $df = n - 1$. Remember that n refers to the number of D scores, not the number of X scores in the original data.

You should also note that the *repeated-measures* t *statistic* is conceptually similar to the t statistics that we have previously examined:

$$t = \frac{\text{sample statistic} - \text{population parameter}}{\text{estimated standard error}}$$

In this case, the sample data are represented by the sample mean of the difference scores (M_D), the population parameter is the value predicted by H_0 ($\mu_D = 0$), and the estimated standard error is computed from the sample data using Equation 11.3.

LEARNING CHECK

1. For a research study comparing two treatment conditions, what characteristic differentiates a repeated-measures design from an independent-measures design?
2. Describe the data used to compute the sample mean and the sample variance for the repeated-measures t statistic.
3. In words and in symbols, what is the null hypothesis for a repeated-measures t test?

ANSWERS

1. For a repeated-measures design, the same group of individuals is tested in both of the treatments. An independent-measures design uses a separate group for each treatment.
2. The two scores obtained for each individual are used to compute a difference score. The sample of difference scores is used to compute the mean and variance.
3. The null hypothesis states that, for the general population, the average difference between the two conditions is zero. In symbols, $\mu_D = 0$.

11.3 HYPOTHESIS TESTS AND EFFECT SIZE FOR THE REPEATED-MEASURES DESIGN

In a repeated-measures study, each individual is measured in two different treatment conditions and we are interested in whether there is a systematic difference between the scores in the first treatment condition and the scores in the second treatment condition. A difference score (D value) is computed for each person and the hypothesis test uses the difference scores from the sample to evaluate the overall mean difference, μ_D, for the entire population. The hypothesis test with the repeated-measures t statistic

follows the same four-step process that we have used for other tests. The complete hypothesis-testing procedure is demonstrated in Example 11.1.

EXAMPLE 11.1

Research indicates that the color red increases men's attraction to women (Elliot & Niesta, 2008). In the original study, men were shown women's photographs presented on either a white or a red background. Photographs presented on red were rated significantly more attractive than the same photographs mounted on white. In a similar study, a researcher prepares a set of 30 women's photographs, with 15 mounted on a white background and 15 mounted on red. One picture is identified as the test photograph, and appears twice in the set, once on white and once on red. Each male participant looks through the entire set of photographs and rates the attractiveness of each woman on a 12-point scale. Table 11.3 summarizes the ratings of the test photograph for a sample of $n = 9$ men. Are the ratings for the test photograph significantly different when it is presented on a red background compared to a white background?

STEP 1 State the hypotheses, and select the alpha level.

H_0: $\mu_D = 0$ (There is no difference between the two colors.)

H_1: $\mu_D \neq 0$ (There is a change.)

For this test, we use $\alpha = .01$.

STEP 2 Locate the critical region. For this example, $n = 9$, so the t statistic has $df = n - 1 = 8$. For $\alpha = .01$, the critical value listed in the t distribution table is ± 3.355. The critical region is shown in Figure 11.3.

TABLE 11.3

Attractiveness ratings for a woman shown in a photograph presented on a red and a white background.

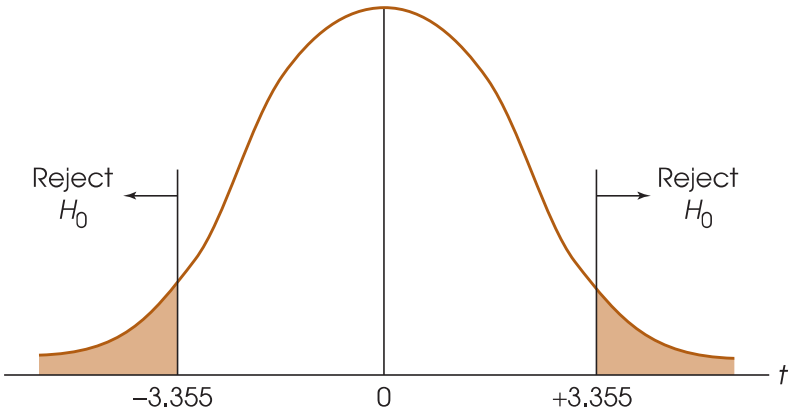
Participant	White Background	Red Background	D	D^2
A	6	9	+3	9
B	8	9	+1	1
C	7	10	+3	9
D	7	11	+4	16
E	8	11	+3	9
F	6	9	+3	9
G	5	11	+6	36
H	10	11	+1	1
I	8	11	+3	9
			$\Sigma D = 27$	$\Sigma D^2 = 99$

$$M_D = \frac{27}{9} = 3.00$$

$$SS = \Sigma D^2 - \frac{(\Sigma D)^2}{n} = 99 - \frac{(27)^2}{9} = 99 - 81 = 18$$

FIGURE 11.3

The critical region for the *t* distribution with $df = 8$ and $\alpha = .01$.



STEP 3 Calculate the *t* statistic. Table 11.3 shows the sample data and the calculations of $M_D = 3.00$ and $SS = 18$. Note that all calculations are done with the difference scores. As we have done with the other *t* statistics, we present the calculation of the *t* statistic as a three-step process.

First, compute the sample variance.

$$s^2 = \frac{SS}{n-1} = \frac{18}{8} = 2.25$$

Next, use the sample variance to compute the estimated standard error.

$$s_{M_D} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{2.25}{9}} = 0.50$$

Finally, use the sample mean (M_D) and the hypothesized population mean (μ_D) along with the estimated standard error to compute the value for the *t* statistic.

$$t = \frac{M_D - \mu_D}{s_{M_D}} = \frac{3.00 - 0}{0.50} = 6.00$$

STEP 4 Make a decision. The *t* value we obtained falls in the critical region (see Figure 11.3). The researcher rejects the null hypothesis and concludes that the background color has a significant effect on the judged attractiveness of the woman in the test photograph.

MEASURING EFFECT SIZE FOR THE REPEATED-MEASURES *t*

As we noted with other hypothesis tests, whenever a treatment effect is found to be statistically significant, it is recommended that you also report a measure of the absolute magnitude of the effect. The most commonly used measures of effect size are Cohen's *d* and r^2, the percentage of variance accounted for. The size of the treatment effect also can be described with a confidence interval estimating the population mean difference, μ_D. Using the data from Example 11.1, we demonstrate how these values are calculated to measure and describe effect size.

Cohen's *d* In Chapters 8 and 9, we introduced Cohen's *d* as a standardized measure of the mean difference between treatments. The standardization simply divides the population mean difference by the standard deviation. For a repeated-measures study, Cohen's *d* is defined as

$$d = \frac{\text{population mean difference}}{\text{standard deviation}} = \frac{\mu_D}{\sigma_D}$$

Because the population mean and standard deviation are unknown, we use the sample values instead. The sample mean, M_D, is the best estimate of the actual mean difference, and the sample standard deviation (square root of sample variance) provides the best estimate of the actual standard deviation. Thus, we are able to estimate the value of *d* as follows:

Because we are measuring the size of the effect and not the direction, it is customary to ignore the minus sign and report Cohen's *d* as a positive value.

$$\text{estimated } d = \frac{\text{sample mean difference}}{\text{sample standard deviation}} = \frac{M_D}{s} \tag{11.4}$$

For the repeated-measures study in Example 11.1, $M_D = 3$ and the sample variance is $s^2 = 2.25$, so the data produce

$$\text{estimated } d = \frac{M_D}{s} = \frac{3.00}{\sqrt{2.25}} = \frac{3.00}{1.5} = 2.00$$

Any value greater than 0.80 is considered to be a large effect, and these data are clearly in that category (see Table 8.2 on p. 264).

The percentage of variance accounted for, r^2 Percentage of variance is computed using the obtained *t* value and the *df* value from the hypothesis test, exactly as was done for the single-sample *t* (see p. 299) and for the independent-measures *t* (see p. 329). For the data in Example 11.1, we obtain

$$r^2 = \frac{t^2}{t^2 + df} = \frac{(6.00)^2}{(6.00)^2 + 8} = \frac{36}{44} = 0.818 \text{ or } 81.8\%$$

For these data, 81.8% of the variance in the scores is explained by the background color for the photograph. More specifically, the color red caused the difference scores to be consistently positive. Thus, the deviations from zero are largely explained by the treatment.

Confidence intervals for estimating μ_D As noted in the previous two chapters, it is possible to compute a confidence interval as an alternative method for measuring and describing the size of the treatment effect. For the repeated-measures *t*, we use a sample mean difference, M_D, to estimate the population mean difference, μ_D. In this case, the confidence interval literally estimates the size of the treatment effect by estimating the population mean difference between the two treatment conditions.

As with the other *t* statistics, the first step is to solve the *t* equation for the unknown parameter. For the repeated-measures *t* statistic, we obtain

$$\mu_D = M_D \pm t s_{M_D} \tag{11.5}$$

In the equation, the values for M_D and for s_{M_D} are obtained from the sample data. Although the value for the *t* statistic is unknown, we can use the degrees of freedom for the *t* statistic and the *t* distribution table to estimate the *t* value. Using the estimated *t* and the known values from the sample, we can then compute the value of μ_D. The following example demonstrates the process of constructing a confidence interval for a population mean difference.

EXAMPLE 11.2

In Example 11.1 we presented a research study demonstrating how men's attractiveness ratings for women are influenced by the color red. In the study, a sample of $n = 9$ men rated a woman shown in a photograph as significantly more attractive when the photo was presented on a red background than when it was on a white background. The mean difference between treatments was $M_D = 3$ points and the estimated standard error for the mean difference was $s_{M_D} = 0.50$. Now, we construct a 95% confidence interval to estimate the size of the population mean difference.

With a sample of $n = 9$ participants, the repeated-measures *t* statistic has $df = 8$. To have 95% confidence, we simply estimate that the *t* statistic for the sample mean difference is located somewhere in the middle 95% of all the possible *t* values. According to the *t* distribution table, with $df = 8$, 95% of the *t* values are located between $t = +2.306$ and $t = -2.306$. Using these values in the estimation equation, together with the values for the sample mean and the standard error, we obtain

$$\mu_D = M_D \pm ts_{M_D}$$

$$= 3 \pm 2.306(0.50)$$

$$= 3 \pm 1.153$$

This produces an interval of values ranging from 3 – 1.153 = 1.847 to 3 + 1.153 = 4.153. Our conclusion is that for general population of men, changing the background color from white to red increases the average attractiveness rating for the woman in the photograph between 1.847 and 4.153 points. We are 95% confident that the true mean difference is in this interval because the only value estimated during the calculations was the *t* statistic, and we are 95% confident that the *t* value is located in the middle 95% of the distribution. Finally note that the confidence interval is constructed around the sample mean difference. As a result, the sample mean difference, $M_D = 3$ points, is located exactly in the center of the interval.

As with the other confidence intervals presented in Chapters 9 and 10, the confidence interval for a repeated-measures *t* is influenced by a variety of factors other than the actual size of the treatment effect. In particular, the width of the interval depends on the percentage of confidence used, so that a larger percentage produces a wider interval. Also, the width of the interval depends on the sample size, so that a larger sample produces a narrower interval. Because the interval width is related to sample size, the confidence interval is not a pure measure of effect size like Cohen's *d* or r^2.

Finally, we should note that the 95% confidence interval computed in Example 11.2 does not include the value $\mu_D = 0$. In other words, we are 95% confident that the population mean difference is not $\mu_D = 0$. This is equivalent to concluding that a null hypothesis specifying that $\mu_D = 0$ would be rejected with a test using $\alpha = .05$. If $\mu_D = 0$ were included in the 95% confidence interval, it would indicate that a hypothesis test would fail to reject H_0 with $\alpha = .05$.

IN THE LITERATURE

REPORTING THE RESULTS OF A REPEATED-MEASURES *t* TEST

As we have seen in Chapters 9 and 10, the APA format for reporting the results of *t* tests consists of a concise statement that incorporates the *t* value, degrees of freedom, and alpha level. One typically includes values for means and standard deviations, either in a statement or a table (Chapter 4). For Example 11.1, we observed a mean difference of $M_D = 3.00$ with $s = 1.50$. Also, we obtained a *t* statistic of $t = 6.00$ with $df = 8$, and our decision was to reject the null hypothesis at the .01 level of significance. Finally, we measured effect size by computing the percentage of variance explained and obtained $r^2 = 0.818$. A published report of this study might summarize the results as follows:

> Changing the background color from white to red increased the attractiveness rating of the woman in the photograph by an average of $M = 3.00$ points with $SD = 1.50$. The treatment effect was statistically significant, $t(8) = 6.00$, $p < .01$, $r^2 = 0.818$.

When the hypothesis test is conducted with a computer program, the printout typically includes an exact probability for the level of significance. The *p*-value from the printout is then stated as the level of significance in the research report. However, the data from Example 11.1 produced a significance level of $p = .000$ in the computer printout. In this case, the probability was so small that the computer rounded it off to 3 decimal points and obtained a value of zero. In this situation you do not know the exact probability value and should report $p < .001$.

If the confidence interval from Example 11.2 is reported as a description of effect size together with the results from the hypothesis test, it would appear as follows:

> Changing the background color from white to red significantly increased the attractiveness rating, $t(8) = 6.00$, $p < .001$, 95% CI [1.817, 4.183].

DESCRIPTIVE STATISTICS AND THE HYPOTHESIS TEST

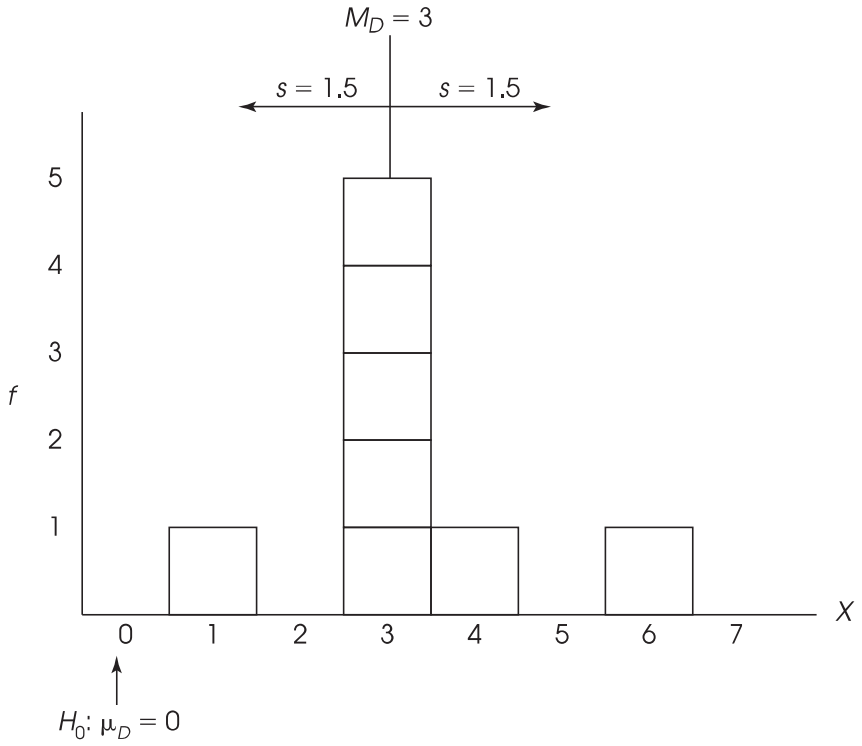
Often, a close look at the sample data from a research study makes it easier to see the size of the treatment effect and to understand the outcome of the hypothesis test. In Example 11.1, we obtained a sample of $n = 9$ men who produce a mean difference of $M_D = 3.00$ with a standard deviation of $s = 1.50$ points. The sample mean and standard deviation describe a set of scores centered at $M_D = 3.00$ with most of the scores located within 1.5 points of the mean. Figure 11.4 shows the actual set of difference scores that were obtained in Example 11.1. In addition to showing the scores in the sample, we have highlighted the position of $\mu_D = 0$; that is, the value specified in the null hypothesis. Notice that the scores in the sample are displaced away from zero. Specifically, the data are not consistent with a population mean of $\mu_D = 0$, which is why we rejected the null hypothesis. In addition, note that the sample mean is located 2 standard deviations above zero. This distance corresponds to the effect size measured by Cohen's $d = 2.00$. For these data, the picture of the sample distribution (see Figure 11.4) should help you to understand the measure of effect size and the outcome of the hypothesis test.

VARIABILITY AS A MEASURE OF CONSISTENCY FOR THE TREATMENT EFFECT

In a repeated-measures study, the variability of the difference scores becomes a relatively concrete and easy-to-understand concept. In particular, the sample variability describes the *consistency* of the treatment effect. For example, if a treatment consistently adds a few points to each individual's score, then the set of difference scores are clustered together with relatively small variability. This is the situation that we

FIGURE 11.4

The sample of difference scores from Example 11.1. The mean is $M_D = 3$ and the standard deviation is $s = 1.5$. The data show a consistent increase in scores (positive differences) and suggest that $\mu_D = 0$ is not a reasonable hypothesis.



observed in Example 11.1 (see Figure 11.4) in which all of the participants produced higher attractiveness ratings for the photograph on a red background. In this situation, with small variability, it is easy to see the treatment effect and it is likely to be significant.

Now consider what happens when the variability is large. Suppose that the red/white study in Example 11.1 produced a sample of $n = 9$ difference scores consisting of –4, –3, –2, +1, +1, +3, +8, +11, and +12. These difference scores also have a mean of $M_D = 3.00$, but now the variability is substantially increased so that $SS = 288$ and the standard deviation is $s = 6.00$. Figure 11.5 shows the new set of difference scores. Again, we have highlighted the position of $\mu_D = 0$, which is the value specified in the null hypothesis. Notice that the high variability means that there is no consistent treatment effect. Some participants rate the photograph as more attractive when it is on a red background (the positive differences) and some rate it higher on a white background (the negative differences). In the hypothesis test, the high variability increases the size of the estimated standard error and results in a hypothesis test that produces $t = 1.50$, which is not in the critical region. With these data, we would fail to reject the null hypothesis and conclude that the color has no effect on the perceived attractiveness of the woman in the photograph.

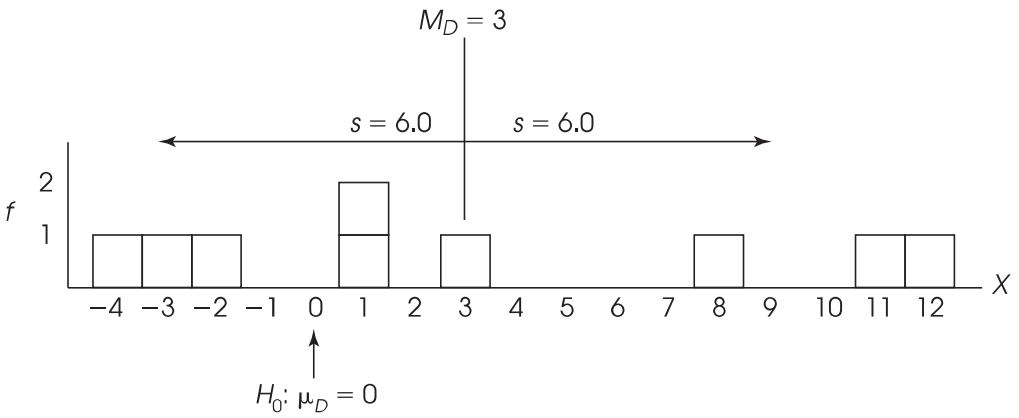
With small variability (see Figure 11.4), the 3-point treatment effect is easy to see and is statistically significant. With large variability (see Figure 11.5), the 3-point effect is not easy to see and is not significant. As we have noted several times in the past, large variability can obscure patterns in the data and reduces the likelihood of finding a significant treatment effect.

DIRECTIONAL HYPOTHESIS AND ONE-TAILED TESTS

In many repeated-measures and matched-subjects studies, the researcher has a specific prediction concerning the direction of the treatment effect. For example, in the study described in Example 11.1, the researcher expects the woman to be judged as more

FIGURE 11.5

A sample of difference scores with a mean difference of $M_D = 3$ and a standard deviation of $s = 6$. The data do not show a consistent increase or decrease in scores. Because there is no consistent treatment effect, $\alpha_D = 0$ is a reasonable hypothesis.



attractive when her photograph is presented on a red background. This kind of directional prediction can be incorporated into the statement of the hypotheses, resulting in a directional, or one-tailed, hypothesis test. The following example demonstrates how the hypotheses and critical region are determined for a directional test.

EXAMPLE 11.3

We reexamine the experiment presented in Example 11.1. The researcher is using a repeated-measures design to investigate the effect of the color red on the perceived attractiveness of a woman. The researcher predicts that the attractiveness ratings for the woman in a photograph will increase when the photograph is presented on a red background compared to a white background.

STEP 1 State the hypotheses and select the alpha level. For this example, the researcher predicts that attractiveness ratings will increase when the photograph is shown on the red background. The null hypothesis, on the other hand says that the attractiveness ratings will not increase but rather will be unchanged or even lowered with the red background. In symbols,

H_0: $\mu_D \leq 0$ (There is no increase with the color red.)

The alternative hypothesis says that the treatment does work. For this example, H_1 says that the color red will increase the attractiveness ratings.

H_1: $\mu_D > 0$ (The rating is increased.)

We use $\alpha = .01$.

STEP 2 Locate the critical region. As we demonstrated with the independent-measures t statistic (p. 305), the critical region for a one-tailed test can be located using a two-stage process. Rather than trying to determine which tail of the distribution contains the critical region, you first look at the sample mean difference to verify that it is in the predicted direction. If not, then the treatment clearly did not work as expected and you can stop the test. If the change is in the correct direction, then the question is whether it is large enough to be significant. For this example, change is in the predicted direction (the researcher predicted higher ratings and the sample mean shows an increase.) With $n = 9$, we obtain $df = 8$ and a critical value of $t = 2.896$ for a one-tailed test with $\alpha = .01$. Thus, any t statistic beyond 2.896 (positive or negative) is sufficient to reject the null hypothesis.

STEP 3 Compute the t statistic. We calculated the t statistic in Example 11.1, and obtained $t = 6.00$.

STEP 4 Make a decision. The obtained t statistic is well beyond the critical boundary. Therefore, we reject the null hypothesis and conclude that the color red significantly increased the attractiveness ratings for the woman in the photograph. In a research report, the use of a one-tailed test would be clearly noted as follows:

> Changing the background color from white to red significantly increased the attractiveness rating, $t(8) = 6.00$, $p < .01$, one tailed.

LEARNING CHECK

1. A researcher is investigating the effectiveness of acupuncture treatment for chronic back pain. A sample of $n = 4$ participants is obtained from a pain clinic. Each individual ranks the current level of pain and then begins a 6-week program of acupuncture treatment. At the end of the program, the pain level is rated again and the researcher records the amount of difference between the two ratings. For this sample, pain level decreased by an average of $M = 4.5$ points with $SS = 27$.
   a. Are the data sufficient to conclude that acupuncture has a significant effect on back pain? Use a two-tailed test with $\alpha = .05$.
   b. Can you conclude that acupuncture significantly reduces back pain? Use a one-tailed test with $\alpha = .05$.
2. Compute the effect size using both Cohen's d and r^2 acupuncture study in the previous question.
3. A computer printout for a repeated-measures t test reports a p value of $p = .021$.
   a. Can the researcher claim a significant effect with $\alpha = .01$?
   b. Is the effect significant with $\alpha = .05$?

ANSWERS

1. a. For these data, the sample variance is 9, the standard error is 1.50, and $t = 3.00$. With $df = 3$, the critical values are $t = \pm 3.182$. Fail to reject the null hypothesis.
   b. For a one-tailed test, the critical value is $t = 2.353$. Reject the null hypothesis and conclude that acupuncture treatment significantly reduces pain.
2. $d = 4.5/3 = 1.50$ and $r^2 = 9/12 = 0.75$.
3. a. The exact p value, $p = .021$, is not less than $\alpha = .01$. Therefore, the effect is not significant for $\alpha = .01$ ($p > .01$).
   b. The p value is less than .05, so the effect is significant with $\alpha = .05$.

11.4 USES AND ASSUMPTIONS FOR REPEATED-MEASURES *t* TESTS

REPEATED-MEASURES VERSUS INDEPENDENT-MEASURES DESIGNS

In many research situations, it is possible to use either a repeated-measures design or an independent-measures design to compare two treatment conditions. The independent-measures design would use two separate samples (one in each treatment condition) and the repeated-measures design would use only one sample with the same individuals participating in both treatments. The decision about which design to use is often

made by considering the advantages and disadvantages of the two designs. In general, the repeated-measures design has most of the advantages.

Number of subjects A repeated-measures design typically requires fewer subjects than an independent-measures design. The repeated-measures design uses the subjects more efficiently because each individual is measured in both of the treatment conditions. This can be especially important when there are relatively few subjects available (for example, when you are studying a rare species or individuals in a rare profession).

Study changes over time The repeated-measures design is especially well suited for studying learning, development, or other changes that take place over time. Remember that this design involves measuring individuals at one time and then returning to measure the same individuals at a later time. In this way, a researcher can observe behaviors that change or develop over time.

Individual differences The primary advantage of a repeated-measures design is that it reduces or eliminates problems caused by individual differences. *Individual differences* are characteristics such as age, IQ, gender, and personality that vary from one individual to another. These individual differences can influence the scores obtained in a research study, and they can affect the outcome of a hypothesis test. Consider the data in Table 11.4. The first set of data represents the results from a typical independent-measures study and the second set represents a repeated-measures study. Note that we have identified each participant by name to help demonstrate the effects of individual differences.

For the independent-measures data, note that every score represents a different person. For the repeated-measures study, on the other hand, the same participants are measured in both of the treatment conditions. This difference between the two designs has some important consequences.

1. We have constructed the data so that both research studies have exactly the same scores and they both show the same 5-point mean difference between treatments. In each case, the researcher would like to conclude that the 5-point difference was caused by the treatments. However, with the independent-measures design, there is always the possibility that the participants in treatment 1 have different characteristics than those in treatment 2. For example, the three participants in treatment 1 may be more intelligent than those in treatment 2 and their higher intelligence caused them to have higher scores. Note that this problem disappears with the repeated-measures design. Specifically, with repeated measures there is no possibility that the participants in one treatment are different from those in another treatment because the same participants are used in all of the treatments.

TABLE 11.4

Hypothetical data showing the results from an independent-measures study and a repeated-measures study. The two sets of data use exactly the same numerical scores and they both show the same 5-point mean difference between treatments.

Independent-Measures Study (2 Separate Samples)		Repeated-Measures Study (Same Sample in Both Treatments)		
Treatment 1	Treatment 2	Treatment 1	Treatment 2	*D*
(John) $X = 18$	(Sue) $X = 15$	(John) $X = 18$	(John) $X = 15$	-3
(Mary) $X = 27$	(Tom) $X = 20$	(Mary) $X = 27$	(Mary) $X = 20$	-7
(Bill) $X = 33$	(Dave) $X = 28$	(Bill) $X = 33$	(Bill) $X = 28$	-5
$M = 26$	$M = 21$			$M_D = -5$
$SS = 114$	$SS = 86$			$SS = 8$

2. Although the two sets of data contain exactly the same scores and have exactly the same 5-point mean difference, you should realize that they are very different in terms of the variance used to compute standard error. For the independent-measures study, you calculate the *SS* or variance for the scores in each of the two separate samples. Note that in each sample there are big differences between participants. In treatment 1, for example, Bill has a score of 33 and John's score is only 18. These individual differences produce a relatively large sample variance and a large standard error. For the independent-measures study, the standard error is 5.77, which produces a *t* statistic of $t = 0.87$. For these data, the hypothesis test concludes that there is no significant difference between treatments.

   In the repeated-measures study, the *SS* and variance are computed for the difference scores. If you examine the repeated-measures data in Table 11.4, you will see that the big differences between John and Bill that exist in treatment 1 and in treatment 2 are eliminated when you get to the difference scores. Because the individual differences are eliminated, the variance and standard error are dramatically reduced. For the repeated-measures study, the standard error is 1.15 and the *t* statistic is $t = -4.35$. With the repeated-measures *t,* the data show a significant difference between treatments. Thus, one big advantage of a repeated-measures study is that it reduces variance by removing individual differences, which increases the chances of finding a significant result.

TIME-RELATED FACTORS AND ORDER EFFECTS

The primary disadvantage of a repeated-measures design is that the structure of the design allows for factors other than the treatment effect to cause a participant's score to change from one treatment to the next. Specifically, in a repeated-measures design, each individual is measured in two different treatment conditions, usually *at two different times.* In this situation, outside factors that change over time may be responsible for changes in the participants' scores. For example, a participant's health or mood may change over time and cause a difference in the participant's scores. Outside factors such as the weather can also change and may have an influence on participants' scores. Because a repeated-measures study typically takes place over time, it is possible that time-related factors (other than the two treatments) are responsible for causing changes in the participants' scores.

Also, it is possible that participation in the first treatment influences the individual's score in the second treatment. If the researcher is measuring individual performance, for example, the participants may gain experience during the first treatment condition, and this extra practice may help their performance in the second condition. In this situation, the researcher would find a mean difference between the two conditions; however, the difference would not be caused by the treatments, instead it would caused be by practice effects. Changes in scores that are caused by participation in an earlier treatment are called *order effects* and can distort the mean differences found in repeated-measures research studies.

Counterbalancing One way to deal with time-related factors and order effects is to counterbalance the order of presentation of treatments. That is, the participants are randomly divided into two groups, with one group receiving treatment 1 followed by treatment 2, and the other group receiving treatment 2 followed by treatment 1. The goal of counterbalancing is to distribute any outside effects evenly over the two treatments. For example, if practice effects are a problem, then half of the participants gain experience in treatment 1, which then helps their performance in treatment 2. However, the other half gain experience in treatment 2, which helps their performance in treatment 1. Thus, prior experience helps the two treatments equally.

Finally, if there is reason to expect strong time-related effects or strong order effects, your best strategy is not to use a repeated-measures design. Instead, use independent-measures (or a matched-subjects design) so that each individual participates in only one treatment and is measured only one time.

ASSUMPTIONS OF THE RELATED-SAMPLES *t* TEST

The related-samples *t* statistic requires two basic assumptions:

1. The observations within each treatment condition must be independent (see p. 254). Notice that the assumption of independence refers to the scores *within* each treatment. Inside each treatment, the scores are obtained from different individuals and should be independent of one another.
2. The population distribution of difference scores (*D* values) must be normal.

   As before, the normality assumption is not a cause for concern unless the sample size is relatively small. In the case of severe departures from normality, the validity of the *t* test may be compromised with small samples. However, with relatively large samples ($n > 30$), this assumption can be ignored.

If there is reason to suspect that one of the assumptions for the repeated-measures *t* test has been violated, an alternative analysis known as the *Wilcoxon test* is presented in Appendix E. The Wilcoxon test requires that the original scores be transformed into ranks before evaluating the difference between the two treatment conditions.

LEARNING CHECK

1. What assumptions must be satisfied for repeated-measures *t* tests to be valid?
2. Describe some situations for which a repeated-measures design is well suited.
3. How is a matched-subjects design similar to a repeated-measures design? How do they differ?
4. The data from a research study consist of 10 scores in each of two different treatment conditions. How many individual subjects would be needed to produce these data
   a. For an independent-measures design?
   b. For a repeated-measures design?
   c. For a matched-subjects design?

ANSWERS

1. The observations within a treatment are independent. The population distribution of *D* scores is assumed to be normal.
2. The repeated-measures design is suited to situations in which a particular type of subject is not readily available for study. This design is helpful because it uses fewer subjects (only one sample is needed). Certain questions are addressed more adequately by a repeated-measures design—for example, any time one would like to study changes across time in the same individuals. Also, when individual differences are large, a repeated-measures design is helpful because it reduces the amount of this type of error in the statistical analysis.
3. They are similar in that the role of individual differences in the experiment is reduced. They differ in that there are two samples in a matched-subjects design and only one in a repeated-measures study.
4. a. The independent-measures design would require 20 subjects (two separate samples with $n = 10$ in each).
   b. The repeated-measures design would require 10 subjects (the same 10 individuals are measured in both treatments).
   c. The matched-subjects design would require 20 subjects (10 matched pairs).

SUMMARY

1. In a related-samples research study, the individuals in one treatment condition are directly related, one-to-one, with the individuals in the other treatment condition(s). The most common related-samples study is a repeated-measures design, in which the same sample of individuals is tested in all of the treatment conditions. This design literally repeats measurements on the same subjects. An alternative is a matched-subjects design, in which the individuals in one sample are matched one-to-one with individuals in another sample. The matching is based on a variable relevant to the study.

2. The repeated-measures *t* test begins by computing a difference between the first and second measurements for each subject (or the difference for each matched pair). The difference scores, or *D* scores, are obtained by

$$D = X_2 - X_1$$

   The sample mean, M_D, and sample variance, s^2, are used to summarize and describe the set of difference scores.

3. The formula for the repeated-measures *t* statistic is

$$t = \frac{M_D - \mu_D}{s_{M_D}}$$

   In the formula, the null hypothesis specifies $\mu_D = 0$, and the estimated standard error is computed by

$$s_{M_D} = \sqrt{\frac{s^2}{n}}$$

4. A repeated-measures design may be preferred to an independent-measures study when one wants to observe changes in behavior in the same subjects, as in learning or developmental studies. An important advantage of the repeated-measures design is that it removes or reduces individual differences, which, in turn lowers sample variability and tends to increase the chances for obtaining a significant result.

5. For a repeated-measures design, effect size can be measured using either r^2 (the percentage of variance accounted for) or Cohen's *d* (the standardized mean difference). The value of r^2 is computed the same way for both independent- and repeated-measures designs.

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

   Cohen's *d* is defined as the sample mean difference divided by standard deviation for both repeated- and independent-measures designs. For repeated-measures studies, Cohen's *d* is estimated as

$$\text{estimated } d = \frac{M_D}{s}$$

6. An alternative method for describing the size of the treatment effect is to construct a confidence interval for the population mean difference, μ_D. The confidence interval uses the repeated-measures *t* equation, solved for the unknown mean difference:

$$\mu_D = M_D \pm t s_{M_D}$$

   First, select a level of confidence and then look up the corresponding *t* values. For example, for 95% confidence, use the range of *t* values that determine the middle 95% of the distribution. The *t* values are then used in the equation along with the values for the sample mean difference and the standard error, which are computed from the sample data.

KEY TERMS

repeated-measures design (352)
within-subjects design (352)
matched-subjects design (353)
related-samples design (353)
difference scores (354)
estimated standard error for M_D (357)
repeated-measures *t* statistic (358)
individual differences (367)
order effects (368)
Wilcoxon test (369)

RESOURCES

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

You can find a tutorial quiz and other learning exercises for Chapter 11 on the book companion website. The website also provides access to a workshop entitled *Independent vs. Repeated t-tests* that compares the *t* test presented in this chapter with the independent-measures test that was presented in Chapter 10.

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

CENGAGE brain .com

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

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

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform **The Repeated-Measures *t* Test** presented in this chapter.

Data Entry

Enter the data into two columns (VAR0001 and VAR0002) in the data editor with the first score for each participant in the first column and the second score in the second column. The two scores for each participant must be in the same row.

Data Analysis

1. Click **Analyze** on the tool bar, select **Compare Means**, and click on **Paired-Samples T Test.**
2. One at a time, highlight the column labels for the two data columns and click the arrow to move them into the **Paired Variables** box.

3. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95%, but you can select **Options** and change the percentage.
4. Click **OK.**

SPSS Output

We used the SPSS program to analyze the data from the red/white photograph experiment in Example 11.1 and the program output is shown in Figure 11.6. The output includes a table of sample statistics with the mean and standard deviation for each treatment. A second table shows the correlation between the two sets of scores (correlations are presented in Chapter 15). The final table, which is split into two sections in Figure 11.6, shows the results of the hypothesis test, including the mean and standard deviation for the difference scores, the standard error for the mean, a 95% confidence interval for the mean difference, and the values for *t*, *df*, and the level of significance (the *p* value for the test).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	VAR00001	7.2222	9	1.48137	.49379
	VAR00002	10.2222	9	.97183	.32394

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	VAR00001 & VAR00002	9	.309	.419

Paired Samples Test

		Paired Differences		
		Mean	Std. Deviation	Std. Error Mean
Pair 1	VAR00001 - VAR00002	−3.00000	1.50000	.50000

Paired Samples Test

		Paired Differences				
		95% Confidence Interval of the Difference				
		Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	VAR00001 - VAR00002	−4.15300	−1.84700	−6.000	8	.000

FIGURE 11.6

The SPSS output for the repeated-measures hypothesis test in Example 11.1.

FOCUS ON PROBLEM SOLVING

1. Once data have been collected, we must then select the appropriate statistical analysis. How can you tell whether the data call for a repeated-measures *t* test? Look at the experiment carefully. Is there only one sample of subjects? Are the same subjects tested a second time? If your answers are yes to both of these questions, then a repeated-measures *t* test should be done. There is only one situation in which the repeated-measures *t* can be used for data from two samples, and that is for *matched-subjects* studies (p. 353).
2. The repeated-measures *t* test is based on difference scores. In finding difference scores, be sure that you are consistent with your method. That is, you may use either $X_2 - X_1$ or $X_1 - X_2$ to find D scores, but you must use the same method for all subjects.

DEMONSTRATION 11.1

A REPEATED-MEASURES *t* TEST

A major oil company would like to improve its tarnished image following a large oil spill. Its marketing department develops a short television commercial and tests it on a sample of $n = 7$ participants. People's attitudes about the company are measured with a short questionnaire, both before and after viewing the commercial. The data are as follows:

Person	X_1 (Before)	X_2 (After)	D (Difference)
A	15	15	0
B	11	13	+2
C	10	18	+8
D	11	12	+1
E	14	16	+2
F	10	10	0
G	11	19	+8

$\Sigma D = 21$

$M_D = \frac{21}{7} = 3.00$

$SS = 74$

Was there a significant change? Note that participants are being tested twice—once before and once after viewing the commercial. Therefore, we have a repeated-measures design.

STEP 1 **State the hypotheses, and select an alpha level.** The null hypothesis states that the commercial has no effect on people's attitude, or, in symbols,

$$H_0: \mu_D = 0 \text{ (The mean difference is zero.)}$$

The alternative hypothesis states that the commercial does alter attitudes about the company, or

$$H_1: \mu_D \neq 0 \text{ (There is a mean change in attitudes.)}$$

For this demonstration, we use an alpha level of .05 for a two-tailed test.

STEP 2 **Locate the critical region.** Degrees of freedom for the repeated-measures *t* test are obtained by the formula

$$df = n - 1$$

For these data, degrees of freedom equal

$$df = 7 - 1 = 6$$

The *t* distribution table is consulted for a two-tailed test with $\alpha = .05$ for $df = 6$. The critical *t* values for the critical region are $t = \pm 2.447$.

STEP 3 **Compute the test statistic.** Once again, we suggest that the calculation of the *t* statistic be divided into a three-part process.

Variance for the D scores: The variance for the sample of *D* scores is

$$s^2 = \frac{SS}{n-1} = \frac{74}{6} = 12.33$$

Estimated standard error for M_D: The estimated standard error for the sample mean difference is computed as follows:

$$s_{M_D} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{12.33}{7}} = \sqrt{1.76} = 1.33$$

The repeated-measures t statistic: Now we have the information required to calculate the *t* statistic.

$$t = \frac{M_D - \mu_D}{s_{M_D}} = \frac{3-0}{1.33} = 2.26$$

STEP 4 **Make a decision about H_0, and state the conclusion.** The obtained *t* value is not extreme enough to fall in the critical region. Therefore, we fail to reject the null hypothesis. We conclude that there is not enough evidence to conclude that the commercial changes people's attitudes, $t(6) = 2.26$, $p > .05$, two-tailed. (Note that we state that *p* is *greater than* .05 because we failed to reject H_0.)

DEMONSTRATION 11.2

EFFECT SIZE FOR THE REPEATED-MEASURES *t*

We estimate Cohen's *d* and calculate r^2 for the data in Demonstration 11.1. The data produced a sample mean difference of $M_D = 3.00$ with a sample variance of $s^2 = 12.33$. Based on these values, Cohen's *d* is

$$\text{estimated } d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{M_D}{s} = \frac{3.00}{\sqrt{12.33}} = \frac{3.00}{3.51} = 0.86$$

The hypothesis test produced $t = 2.26$ with $df = 6$. Based on these values,

$$r^2 = \frac{t^2}{t^2 + df} = \frac{(2.26)^2}{(2.26)^2 + 6} = \frac{5.11}{11.11} = 0.46 \quad (\text{or } 46\%)$$

PROBLEMS

1. For the following studies, indicate whether a repeated-measures *t* test is the appropriate analysis. Explain your answers.
- **a.** A researcher is comparing the amount of time spent playing video games each week for college males versus college females.
- **b.** A researcher is comparing two new designs for cell phones by having a group of high school students send a scripted text message on each model and measuring the difference in speed for each student.
- **c.** A researcher is evaluating the effects of fatigue by testing people in the morning when they are well rested and testing again at midnight when they have been awake for at least 14 hours.

2. Participants enter a research study with unique characteristics that produce different scores from one person to another. For an independent-measures study, these individual differences can cause problems. Briefly explain how these problems are eliminated or reduced with a repeated-measures study.

3. Explain the difference between a matched-subjects design and a repeated-measures design.

4. A researcher conducts an experiment comparing two treatment conditions and obtains data with 10 scores for each treatment condition.
- **a.** If the researcher used an independent-measures design, how many subjects participated in the experiment?
- **b.** If the researcher used a repeated-measures design, how many subjects participated in the experiment?
- **c.** If the researcher used a matched-subjects design, how many subjects participated in the experiment?

5. A sample of $n = 9$ individuals participates in a repeated-measures study that produces a sample mean difference of $M_D = 6.5$ with $SS = 200$ for the difference scores.
- **a.** Calculate the standard deviation for the sample of difference scores. Briefly explain what is measured by the standard deviation.
- **b.** Calculate the estimated standard error for the sample mean difference. Briefly explain what is measured by the estimated standard error.

6. a. A repeated-measures study with a sample of $n = 25$ participants produces a mean difference of $M_D = 3$ with a standard deviation of $s = 4$. Based on the mean and standard deviation, you should be able to visualize (or sketch) the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

b. Now assume that the sample standard deviation is $s = 12$, and once again visualize the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

Explain how the size of the sample standard deviation influences the likelihood of finding a significant mean difference.

7. a. A repeated-measures study with a sample of $n = 9$ participants produces a mean difference of $M_D = 3$ with a standard deviation of $s = 6$. Based on the mean and standard deviation, you should be able to visualize (or sketch) the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

b. Now assume that the sample mean difference is $M_D = 12$, and once again visualize the sample distribution. Use a two-tailed hypothesis test with $\alpha = .05$ to determine whether it is likely that this sample came from a population with $\mu_D = 0$.

c. Explain how the size of the sample mean difference influences the likelihood of finding a significant mean difference.

8. A sample of difference scores from a repeated-measures experiment has a mean of $M_D = 4$ with a standard deviation of $s = 6$.
- **a.** If $n = 4$, is this sample sufficient to reject the null hypothesis using a two-tailed test with $\alpha = .05$?
- **b.** Would you reject H_0 if $n = 16$? Again, assume a two-tailed test with $\alpha = .05$.
- **c.** Explain how the size of the sample influences the likelihood of finding a significant mean difference.

9. As mentioned in Chapters 2 and 3 (pp. 38 and 81), Steven Schmidt (1994) reported a series of studies examining the effect of humor on memory. In one part of the study, participants were presented with a list containing a mix of humorous and nonhumorous sentences, and were then asked to recall as many sentences as possible. Schmidt recorded the number of humorous and the number of nonhumorous sentences recalled by each individual. Notice that the data consist of two memory scores for each participant. Suppose that a difference score is computed for each individual in a sample of $n = 16$ and the resulting data show that participants recalled an average of $M_D = 3.25$ more humorous sentences than nonhumorous, with $SS = 135$. Are these results sufficient to conclude that humor has a significant effect on memory? Use a two-tailed test with $\alpha = .05$.

10. Research has shown that losing even one night's sleep can have a significant effect on performance of complex tasks such as problem solving (Linde & Bergstroem, 1992). To demonstrate this phenomenon, a sample of $n = 25$ college students was given a problem-solving task at noon on one day and again at noon on the following day. The students were not permitted any sleep between the two tests. For each student, the difference between the first and second score was recorded. For this sample, the students averaged $M_D = 4.7$ points better on the first test with a variance of $s^2 = 64$ for the difference scores.

a. Do the data indicate a significant change in problem-solving ability? Use a two-tailed test with $\alpha = .05$.

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

11. Strack, Martin, and Stepper (1988) reported that people rate cartoons as funnier when holding a pen in their teeth (which forced them to smile) than when holding a pen in their lips (which forced them to frown). A researcher attempted to replicate this result using a sample of $n = 25$ adults between the ages of 40 and 45. For each person, the researcher recorded the difference between the rating obtained while smiling and the rating obtained while frowning. On average the cartoons were rated as funnier when the participants were smiling, with an average difference of $M_D = 1.6$ with $SS = 150$.

a. Do the data indicate that the cartoons are rated significantly funnier when the participants are smiling? Use a one-tailed test with $\alpha = .01$.

b. Compute r^2 to measure the size of the treatment effect.

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

12. How would you react to doing much worse on an exam than you expected? There is some evidence to suggest that most individuals believe that they can cope with this kind of problem better than their fellow students (Igou, 2008). In the study, participants read a scenario of a negative event and were asked to use a 10-point scale to rate how it would affect their immediate well-being (−5 strongly worsen to +5 strongly improve). Then they were asked to imagine the event from the perspective of an ordinary fellow student and rate how it would affect that person. The difference between the two ratings was recorded. Suppose that a sample of $n = 25$ participants produced a mean difference of $M_D = 1.28$ points (self rated higher) with a standard deviation of $s = 1.50$ for the difference scores.

a. Is this result sufficient to conclude that there is a significant difference in the ratings for self versus others? Use a two-tailed test with $\alpha = .05$.

b. Compute r^2 and estimate Cohen's d to measure the size of the treatment effect.

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

13. Research results indicate that physically attractive people are also perceived as being more intelligent (Eagly, Ashmore, Makhijani, & Longo, 1991). As a demonstration of this phenomenon, a researcher obtained a set of 10 photographs, 5 showing men who were judged to be attractive and 5 showing men who were judged to be unattractive. The photographs were shown to a sample of $n = 25$ college students and the students were asked to rate the intelligence of the person in the photo on a scale from 1 to 10. For each student, the researcher determined the average rating for the 5 attractive photos and the average for the 5 unattractive photos, and then computed the difference between the two scores. For the entire sample, the average difference was $M_D = 2.7$ (attractive photos rated higher) with $s = 2.00$. Are the data sufficient to conclude that there was a significant difference in perceived intelligence for the two sets of photos? Use a two-tailed test at the .05 level of significance.

14. Researchers have noted a decline in cognitive functioning as people age (Bartus, 1990). However, the results from other research suggest that the antioxidants in foods such as blueberries may reduce and even reverse these age-related declines (Joseph et al., 1999). To examine this phenomenon, suppose that a researcher obtains a sample of $n = 16$ adults who are between the ages of 65 and 75. The researcher uses a standardized test to measure cognitive performance for each individual. The participants then begin a 2-month program in which they receive daily doses of a blueberry supplement. At the end of the 2-month period, the researcher again measures cognitive performance for each participant. The results show an average increase in performance of $M_D = 7.4$ with $SS = 1215$.

a. Does this result support the conclusion that the antioxidant supplement has a significant effect on cognitive performance? Use a two-tailed test with $\alpha = .05$.

b. Construct a 95% confidence interval to estimate the average cognitive performance improvement for the population of older adults.

15. The following data are from a repeated-measures study examining the effect of a treatment by measuring a group of $n = 4$ participants before and after they receive the treatment.

a. Calculate the difference scores and M_D.

b. Compute *SS*, sample variance, and estimated standard error.

c. Is there a significant treatment effect? Use $\alpha = .05$, two tails.

Participant	Before Treatment	After Treatment
A	7	10
B	6	13
C	9	12
D	5	8

16. A researcher for a cereal company wanted to demonstrate the health benefits of eating oatmeal. A sample of 9 volunteers was obtained and each participant ate a fixed diet without any oatmeal for 30 days. At the end of the 30-day period, cholesterol was measured for each individual. Then the participants began a second 30-day period in which they repeated exactly the same diet except that they added 2 cups of oatmeal each day. After the second 30-day period, cholesterol levels were measured again and the researcher recorded the difference between the two scores for each participant. For this sample, cholesterol scores averaged $M_D = 16$ points lower with the oatmeal diet with $SS = 538$ for the difference scores.

a. Are the data sufficient to indicate a significant change in cholesterol level? Use a two-tailed test with $\alpha = .01$.

b. Compute r^2, the percentage of variance accounted for by the treatment, to measure the size of the treatment effect.

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

17. A variety of research results suggest that visual images interfere with visual perception. In one study, Segal and Fusella (1970) had participants watch a screen, looking for brief presentations of a small blue arrow. On some trials, the participants were also asked to form a mental image (for example, imagine a volcano). The results for a sample of $n = 6$, show that participants made an average of $M_D = 4.3$ more errors while forming images than while not forming images. The difference scores had $SS = 63$. Do the data indicate a significant difference between the two conditions? Use a two-tailed test with $\alpha = .05$.

18. One of the primary advantages of a repeated-measures design, compared to independent-measures, is that it reduces the overall variability by removing variance caused by individual differences. The following data are from a research study comparing two treatment conditions.

a. Assume that the data are from an independent-measures study using two separate samples, each with $n = 6$ participants. Compute the pooled variance and the estimated standard error for the mean difference.

b. Now assume that the data are from a repeated-measures study using the same sample of $n = 6$ participants in both treatment conditions. Compute the variance for the sample of difference scores and the estimated standard error for the mean difference. (You should find that the repeated-measures design substantially reduces the variance and the standard error.)

Treatment 1	Treatment 2	Difference
10	13	3
12	12	0
8	10	2
6	10	4
5	6	1
7	9	2
$M = 8$	$M = 10$	$M_D = 2$
$SS = 34$	$SS = 30$	$SS = 10$

19. The previous problem demonstrates that removing individual differences can substantially reduce variance and lower the standard error. However, this benefit only occurs if the individual differences are consistent across treatment conditions. In problem 18, for example, the first two participants (top two rows) consistently had the highest scores in both treatment conditions. Similarly, the last two participants consistently had the lowest scores in both treatments. To construct the following data, we started with the scores in problem 18 and scrambled the scores in treatment 1 to eliminate the consistency of the individual differences.

a. Assume that the data are from an independent-measures study using two separate samples, each with $n = 6$ participants. Compute the pooled variance and the estimated standard error for the mean difference.

b. Now assume that the data are from a repeated-measures study using the same sample of $n = 6$ participants in both treatment conditions. Compute the variance for the sample of difference scores and the estimated standard error for the mean difference. (This time you should find that removing the individual differences does not reduce the variance or the standard error.)

Treatment 1	Treatment 2	Difference
6	13	7
7	12	5
8	10	2
10	10	0
5	6	0
12	9	−3
$M = 8$	$M = 10$	$M_D = 2$
$SS = 34$	$SS = 30$	$SS = 64$

20. A researcher uses a matched-subjects design to investigate whether single people who own pets are generally happier than singles without pets. A mood inventory questionnaire is administered to a group of 20- to 29-year-old non–pet owners and a similar age group of pet owners. The pet owners are matched one to one with the non–pet owners for income, number of close friendships, and general health. The data are as follows:

Matched Pair	Non–Pet Owner	Pet Owner
A	12	14
B	8	7
C	10	13
D	9	9
E	7	13
F	10	12

a. Is there a significant difference in the mood scores for non–pet owners versus pet owners? Test with $\alpha = .05$ for two tails.

b. Construct the 95% confidence interval to estimate the size of the mean difference in mood between the population of pet owners and the population of non–pet owners. (You should find that a mean difference of $\mu_D = 0$ is an acceptable value, which is consistent with the conclusion from the hypothesis test.)

21. There is some evidence suggesting that you are likely to improve your test score if you rethink and change answers on a multiple-choice exam (Johnston, 1975). To examine this phenomenon, a teacher gave the same final exam to two sections of a psychology course. The students in one section were told to turn in their exams immediately after finishing, without changing any of their answers. In the other section, students were encouraged to reconsider each question and to change answers whenever they felt it was appropriate. Before the final exam, the teacher had matched 9 students in the first section with 9 students in the second section based on their midterm grades. For example, a student in the no-change section with an 89 on the midterm exam was matched with student in the change section who also had an 89 on the midterm. The final exam grades for the 9 matched pairs of students are presented in the following table.

a. Do the data indicate a significant difference between the two conditions? Use a two-tailed test with $\alpha = .05$.

b. Construct a 95% confidence interval to estimate the size of the population mean difference.

c. Write a sentence demonstrating how the results of the hypothesis test and the confidence interval would appear in a research report.

Matched Pair	No-Change Section	Change Section
#1	71	86
#2	68	80
#3	91	88
#4	65	74
#5	73	82
#6	81	89
#7	85	85
#8	86	88
#9	65	76

22. The teacher from the previous problem also tried a different approach to answering the question of whether changing answers helps or hurts exam grades. In a separate class, students were encouraged to review their final exams and change any answers they wanted to before turning in their papers. However, the students had to indicate both the original answer and the changed answer for each question. The teacher then graded each exam twice, one using the set of original answers and once with the changes. In the class of $n = 22$ students, the average exam score improved by an average of $M_D = 2.5$ points with the changed answers. The standard deviation for the difference scores was $s = 3.1$. Are the data sufficient to conclude that rethinking and changing answers can significantly improve exam scores? Use a one-tailed test at the .01 level of significance.

23. At the Olympic level of competition, even the smallest factors can make the difference between winning and losing. For example, Pelton (1983) has shown that Olympic marksmen shoot much better if they fire between heartbeats, rather than squeezing the trigger during a heartbeat. The small vibration caused by a heartbeat seems to be sufficient to affect the marksman's aim. The following hypothetical data

demonstrate this phenomenon. A sample of $n = 8$ Olympic marksmen fires a series of rounds while a researcher records heartbeats. For each marksman, a score is recorded for shots fired during heartbeats and for shots fired between heartbeats. Do these data indicate a significant difference? Test with $\alpha = .05$.

Participant	During Heartbeats	Between Heartbeats
A	93	98
B	90	94
C	95	96
D	92	91
E	95	97
F	91	97
G	92	95
H	93	97

24. The Preview section of this chapter presented a repeated-measures research study demonstrating that swearing can help reduce pain (Stephens, Atkins, & Kingston, 2009). In the study, each participant was asked to plunge a hand into icy water and keep it there as long as the pain would allow. In one condition, the participants repeated their favorite curse words while their hands were in the water. In the other condition, the participants repeated a neutral word. Data similar to the results obtained in the study are shown in the following table.

a. Do these data indicate a significant difference in pain tolerance between the two conditions? Use a two-tailed test with $\alpha = .05$.

b. Compute r^2, the percentage of variance accounted for, to measure the size of the treatment effect.

c. Write a sentence demonstrating how the results of the hypothesis test and the measure of effect size would appear in a research report.

	Amount of Time (in Seconds)	
Participant	Swear Words	Neutral Words
1	94	59
2	70	61
3	52	47
4	83	60
5	46	35
6	117	92
7	69	53
8	39	30
9	51	56
10	73	61

PART III

REVIEW

After completing this part, you should be able to perform hypothesis tests and compute confidence intervals using *t* statistics. These include:

1. The single-sample *t* introduced in Chapter 9.
2. The independent-measures *t* introduced in Chapter 10.
3. The repeated-measures *t* introduced in Chapter 11.

In this part, we considered a set of three *t* statistics that are used to draw inferences about the means and mean differences for unknown populations. Because the populations are completely unknown, we rely on sample data to provide all of the necessary information. In particular, each inferential procedure begins by computing sample means and sample variances (or the corresponding *SS* values or standard deviations). Therefore, a good understanding of the definitions and formulas from Chapters 3 and 4 is a critical foundation for this section.

With three different *t* statistics available, the first problem is often deciding which one is appropriate for a specific research situation. Perhaps the best approach is to begin with a close look at the sample data.

1. For the single-sample *t* (Chapter 9), there is only one group of participants and only one score for each individual. With a single sample mean and a single sample variance, the *t* statistic can be used to test a hypothesis about a single unknown population mean or construct a confidence interval to estimate the population mean.
2. For the independent-measures *t*, there are two separate groups of participants who produce two groups of scores. The mean and variance are computed for each group, producing two sample means and two sample variances. After pooling the two variances, the *t* statistic uses the difference between the two sample means to test a hypothesis about the corresponding difference between the two unknown population means or estimate the population mean difference with a confidence interval. The null hypothesis always states that there is no difference between the two population means; $\mu_1 - \mu_2 = 0$.
3. For the repeated-measures *t*, there is only one group of participants but each individual is measured twice, at two different times and/or under two different treatment conditions. The two scores are then used to find a difference score for each person, and the mean and variance are computed for the sample of difference scores. The *t* statistic uses the sample mean difference to test a hypothesis about the corresponding population mean difference or estimate the population mean difference with a confidence interval. The null hypothesis always states that the mean for the population of difference scores is zero; $\mu_D = 0$.

REVIEW EXERCISES

1. People tend to evaluate the quality of their lives relative to others around them. In a demonstration of this phenomenon, Frieswijk, Buunk, Steverink, and Slaets (2004) conducted interviews with frail elderly people. In the interview, each person was compared with fictitious others who were worse off. After the interviews, the participants completed a life-satisfaction survey and reported more satisfaction with their own lives. Following are hypothetical data similar to those obtained in the research study, representing satisfaction scores for a sample of $n = 9$ older people who completed the interview. Assume that the average score on the life-satisfaction scale is $\mu = 20$. The scores for the sample are 18, 23, 24, 22, 19, 27, 23, 26, 25.

   a. Calculate the mean and standard deviation for the sample.
   b. Are the data sufficient to conclude that the people in this sample are significantly more satisfied than others in the general population? Use a one-tailed test with $\alpha = .05$.
   c. Compute Cohen's *d* to estimate the size of the effect.
   d. Compute the 90% confidence interval for the mean life-satisfaction score for people who participate in this type of interview.

2. In the problems at the end of Chapter 8, we presented a study indicating that people with visible tattoos are viewed more negatively than people without visible tattoos (Resenhoeft, Villa, & Wiseman, 2008). Suppose that a researcher intends to examine this phenomenon by asking participants to rate the attractiveness of women in a series of ten photographs. For one group of participants, none of the women has any visible tattoos. For a second group, however, the researcher modified one of the photographs by adding a tattoo of a butterfly on the woman's left arm. Using a 7-point rating scale, the $n = 15$ participants who viewed the photograph with no tattoo gave the woman an average rating of $M = 4.9$ with $SS = 15.0$. The $n = 15$ participants who saw the photograph with a tattoo gave the same woman an average rating of $M = 4.2$ with $SS = 18.6$.

   a. Does the existence of a tattoo have a significant effect on the attractiveness rating of the woman in the photograph? Use a two-tailed test with $\alpha = .05$.
   b. Compute r^2, the percentage of variance accounted for by the treatment, to measure the effect size.
   c. Write a sentence describing the outcome of the hypothesis test and the measure of effect size as it would appear in a research report.

3. The stimulant Ritalin has been shown to increase attention span and improve academic performance in children with ADHD (Evans, Pelham, Smith, et al., 2001). To demonstrate the effectiveness of the drug, a researcher selects a sample of $n = 20$ children diagnosed with the disorder and measures each child's attention span before and after taking the drug. The data show an average increase of attention span of $M_D = 4.8$ minutes with a variance of $s^2 = 125$ for the sample of difference scores.

a. Is this result sufficient to conclude that Ritalin significantly improves attention span? Use a one-tailed test with $\alpha = .05$.

b. Compute the 80% confidence interval for the mean change in attention span for the population.

This page intentionally left blank

PART

IV

Analysis of Variance: Tests for Differences Among Two or More Population Means

In Part III we presented a set of *t* statistics that use sample means and mean differences to draw inferences about the corresponding population means and mean differences. However, the *t* statistics are limited to situations that compare no more than two population means. Often, a research question involves the differences among more than two means and, in these situations, *t* tests are not appropriate. In this part we introduce a new hypothesis testing technique known as analysis of variance (ANOVA). ANOVA permits researchers to evaluate the mean differences among *two or more* populations using sample data. We present three different applications of ANOVA that apply to three distinct research situations:

1. Independent-measures designs: Using two or more separate samples to draw an inference about the mean differences between two or more unknown populations.
2. Repeated-measures designs: Using one sample, with each individual tested in two or more different treatment conditions, to draw an inference about the population mean differences among the conditions.
3. Two-factor designs: Allowing two independent variables to change simultaneously within one study to create combinations of treatment conditions involving both variables. The ANOVA then evaluates the mean differences attributed to each variable acting independently and to combinations of the two variables interacting together.

In the next three chapters we continue to examine statistical methods that use sample means as the foundation for drawing inferences about population means. The primary application of these