CHAPTER 8

t Tests

What Is a *t* Test?

Because there is a distinction between the common statistical description of *t* tests and the more technical definition, *t* tests can be a little confusing. The common-use definition or description of *t* tests is simply comparing two means to see if they are significantly different from each other. The more technical definition or description of a *t* test is any statistical test that uses the *t*, or Student's *t*, family of distributions. In this chapter, I will briefly describe the family of distributions known as the *t* distribution. Then I will discuss the three most commonly conducted *t* tests: the **one-sample *t* test**, the **independent samples *t* test**, and the **paired** or **dependent samples *t* test**.

t Distributions

In Chapters 4 and 5, I discussed the normal distribution and how to use the normal distribution to find *z* scores. The probabilities that are based on the normal distribution are accurate when (a) the population standard deviation is known, and/or (b) we have a large sample (i.e., $n > 120$). If neither of these is true, then we cannot assume that we have a nicely shaped bell curve and we cannot use the probabilities that are based on this normal distribution. Instead, we have to adjust our probability estimates by taking our sample size into account. As I discussed in Chapter 6, we are fortunate to have a set of distributions that have already been created for us that do this, and this is known as the family of *t* distributions. Which specific *t* distribution you use for a given problem depends on the size of your sample. There is a table of probabilities based on the different *t* distributions in Appendix B.

For a brief video describing how to read and interpret Appendix B, please refer to the website that accompanies this book.

The One-Sample *t* Test

We examined the one-sample *t* test in some detail in Chapters 6 and 7, so I will not spend too much time on it in this chapter. As the name implies, a one-sample *t* test is performed when you want to compare the mean from a single sample to a population mean. For example, suppose I wanted to compare the scores on a standardized mathematics test of the students of a particular teacher to the test scores of all the students in the school. The students in the school represent the population and I select a random sample of 25 students who all have the same particular teacher for mathematics. I calculate the mean of my sample and compare it to the mean of the population to see whether they differ. If the difference between the sample mean and the population mean is statistically significant, I would conclude that the sample represents a different population (i.e., the population of students who have this particular teacher for mathematics)

than the larger population of students in the school. However, if the one-sample *t* test produces a result that was not statistically significant, I would conclude that the sample does not represent a different population than the rest of the students in the school; they are all part of the same population in terms of their mathematics test scores. (I provide a worked example of how to calculate a one-sample *t* test later in this chapter.)

The Independent Samples *t* Test

One of the most commonly used *t* tests is the independent samples *t* test. You use this test when you want to compare the means of two *independent* samples on a given variable. For example, if you wanted to compare the average height of 50 randomly selected men to that of 50 randomly selected women, you would conduct an independent samples *t* test. Note that the sample of men is not related to the sample of women, and there is no overlap between these two samples (i.e., one cannot be a member of both groups). Therefore, these groups are *independent*, and an independent samples *t* test is appropriate. To conduct an independent samples *t* test, you need one **categorical** or **nominal independent variable** and one **continuous** or **intervally scaled dependent variable**. A dependent variable is a variable on which the scores may differ, or *depend* on the value of the independent variable. An independent variable is the variable that may cause, or simply be used to predict, the value of the dependent variable. The independent variable in a *t* test is simply a variable with two categories (e.g., men and women, fifth graders and ninth graders, etc.). In this type of *t* test, we want to know whether the average scores on the dependent variable differ according to which group one belongs to (i.e., the level of the independent variable). For example, we may want to know if the average height of people (height is the dependent, continuous variable) *depends* on whether the person is a man or a woman (the gender of the person is the independent, categorical variable).

In the real world of research, you can find many examples of independent *t* tests. Comparisons of genders, groups in experiments (i.e., treatment vs. control), and any number of other two-group comparisons can be found. One example that I found was a study examining the stress levels of Hispanic adolescents (Goldbach et al., 2015). The researchers compared a sample of Hispanic adolescents who drank alcohol in the previous 30 days with a sample that did not. Using independent *t* tests, the researchers found that the sample of Hispanic adolescents who used alcohol had significantly higher average levels of stress in several areas (e.g., violence in their community, economic stress in the family, discrimination) than did Hispanic adolescents who did not use alcohol. In this example, the group variable (i.e., Hispanic adolescents who used alcohol, Hispanic adolescents who did not use alcohol) was the categorical, independent variable, and the level of stress was the dependent, interval variable.

Dependent (Paired) Samples *t* Test

A dependent samples *t* test is also used to compare two means on a single dependent variable. Unlike the independent samples test, however, a dependent samples *t* test is used to compare the means of a single sample or of two **matched** or **paired samples**. For example, if a group of students took a math test in March and that same group of students took the same math test two months later in May, we could compare their average scores on the two test dates using a dependent samples *t* test. Or, suppose that we wanted to compare a sample of boys' Scholastic Aptitude Test (SAT) scores with their fathers' SAT scores. In this example, each boy in our study would be matched with his father. In both of these examples, each score is matched, or paired, with a second score. Because of this pairing, we say that the scores are *dependent* upon each other, and a dependent samples *t* test is warranted.

In the real world of research, dependent *t* tests are often used to examine whether the scores on some variable change significantly from one time to another. For example, a team of researchers at the University of California, San Diego conducted a study to examine whether wearing a backpack reduced blood flow to the shoulders, arms, and hands (Neuschwander et al., 2008). The researchers measured the blood flow of eight adults before wearing a backpack and again while wearing a 26-pound backpack for 10 minutes. They found that after wearing the backpack, blood flow to the arms and hands had decreased, which is known to cause fatigue and problems with fine motor control. The use of dependent *t* tests is common in this kind of before-and-after research design.

Independent Samples *t* Tests in Depth

To understand how *t* tests work, it may be most helpful to first try to understand the conceptual issues and then move on to the more mechanical issues involved in the formulas. Because the independent and dependent forms of the *t* tests are quite different, I discuss them separately. Let's begin with the independent samples *t* test.

Conceptual Issues with the Independent Samples *t* Test

The most complicated conceptual issue in the independent samples *t* test involves the standard error for the test. If you think about what this *t* test does, you can see that it is designed to answer a fairly straightforward question: Do two independent samples differ from each other *significantly* in their average scores on some variable? Using an example to clarify this question, we might want to know whether a random sample of 50 men differs *significantly* from a random sample of 50 women in their average enjoyment of a new television show. Suppose that I arranged to have each sample view my new television show and than rate, on a scale from 1 to 10, how much they enjoyed the show, with higher scores indicating greater enjoyment. In addition, suppose that my sample of men gave the show an average rating of 7.5 and my sample of women gave the show an average rating of 6.5.

In looking at these two means, I can clearly see that my sample of men had a higher mean enjoyment of the television show than did my sample of women. But if you look closely at my earlier question, you'll see that I did not ask simply whether my sample of men differed from my sample of women in their average enjoyment of the show. I asked whether they differed *significantly* in their average enjoyment of the show. The word *significantly* is critical in much of statistics, so I discuss it briefly here as it applies to independent *t* tests (for a more thorough discussion, see Chapter 7).

When I conduct an independent samples *t* test, I generally must collect data from two samples and compare the means of these two samples. But I am interested not only in whether these two samples differ on some variable, but also whether the differences in the two sample means are large enough to suggest that there are also differences in the two *populations* that these samples represent. So, returning to our previous example, I already know that the 50 men in my sample enjoyed the television show more, on average, than did the 50 women in my sample. So what? Who really cares about these 50 men and these 50 women, other than their friends and families? What I really want to know is whether the difference between these two samples of men and women is large enough to indicate that men *in general* (i.e., the population of men that this sample represents) will like the television show more than women *in general* (i.e., the population of women that this sample represents). In other words, is this difference of 1.0 between my two samples large enough to represent a real difference between the populations of men and women on this variable? The way of asking this question in statistical shorthand is to ask, "Is the difference between the means of these two samples statistically significant?" (or **significant** for short).

To answer this question, I must know how much difference I should *expect* to see between two samples of this size drawn from these two populations. On the one hand, my null hypothesis says that I am expecting no difference in my population means, and I am conducting this *t* test to determine whether to retain or reject my null hypothesis. But from a statistical sampling

perspective, we know that when we select random samples from populations, there is likely to be some difference between the sample means and the population means. (See the discussion of standard errors in Chapter 6.) If I were to randomly select a different sample of 50 men and a different sample of 50 women, I might get the opposite effect, where the women outscore the men. Or I might get an even larger difference, where men outscore the women by 3 points rather than 1. So the critical question here is this: If I were to repeatedly select random samples of this size ($n = 50$) from each of these populations, what would be the *average expected difference between the means*? In other words, *what is the* **standard error of the difference between the means**?

As I have said before, understanding the concept of standard error provides the key to understanding how inferential statistics work, so take your time and reread the preceding four paragraphs to make sure you get the gist. Regarding the specific case of independent samples *t* tests, we can conclude that the question we want to answer is whether the difference between our two sample means is large or small compared to the difference we would expect to see just by selecting two different random samples. Phrased another way, we want to know whether our *observed* difference between our two sample means is large relative to the standard error of the difference between the means. The general formula for this question is as follows:

$$t = \frac{\text{observed difference between sample means}}{\text{standard error of the difference between the means}}$$

or

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}1-\bar{x}2}}$$

where $\bar{X}_1$ is the mean for sample 1,
$\bar{X}_2$ is the mean for sample 2,
$s_{\bar{x}1-\bar{x}2}$ is the standard error of the difference between the means.

The Standard Error of the Difference between Independent Sample Means

The standard error of the difference between independent sample means is a little bit more complex than the standard error of the mean discussed in Chapter 6. That's because instead of dealing with a single sample, now we have to find a single standard error involving two samples. Generally speaking, this involves simply combining standard errors of the two samples. In fact, when the two samples are roughly the same size, the standard error for the difference between the means is similar to simply combining the two sample standard errors of the mean, as the formula presented in Table 8.1 indicates.

When the two samples are not roughly equal in size, there is a potential problem with using the formulas in Table 8.1 to calculate the standard error. Because these formulas essentially blend the standard errors of each sample together, they also essentially give each sample equal weight

TABLE 8.1 Formula for calculating the standard error of the difference between independent sample means when the sample sizes are roughly equal

$$s_{\bar{x}1-\bar{x}2} = \sqrt{s_{\bar{x}1}^2 + s_{\bar{x}2}^2}$$

$s_{\bar{x}1}$ is the standard error of the mean for the first sample.
$s_{\bar{x}2}$ is the standard error of the mean for the second sample.

and treat the two samples as one new, larger sample. But if the two samples are not of equal size, and especially if they do not have equal standard deviations, then we must adjust the formula for the standard error to take these differences into account. The only difference between this formula and the formula for the standard error when the sample sizes are equal is that the unequal sample size formula adjusts for the different sample sizes. This adjustment is necessary to give the proper weight to each sample's contribution to the overall standard error. Independent *t* tests assume that the size of the variance in each sample is about equal. If this assumption is violated, and one sample is considerably larger than the other, you could end up with a situation where a small sample with a large variance is creating a larger standard error than it should in the independent *t* test. To keep this from happening, when sample sizes are not equal, the formula for calculating the standard error of the independent *t* test needs to be adjusted to give each sample the proper weight. (If the variances of the two samples are grossly unequal, the sample sizes are very different, and/or the data are not normally distributed, a nonparametric alternative to the *t* test – the Mann–Whitney U test – should be considered.)

In practice, let us hope that you will never need to actually calculate any of these standard errors by hand. Because statistical computer programs compute these for us these days, it may be more important to understand the concepts involved than the components of the formulas themselves. In this spirit, try to understand what the standard error of the difference between independent sample means is and why it may differ if the sample sizes are unequal. Simply put, the standard error of the difference between two independent sample means is the average expected difference between any two samples of a given size randomly selected from two populations on a given variable. In our example comparing men's and women's enjoyment of the new television show, the standard error would be the average (i.e., *standard*) amount of difference (i.e., error) we would expect to find between any two samples of 50 men and 50 women selected randomly from the larger populations of men and women.

TIME OUT FOR TECHNICALITY: STANDARD ERROR FOR INDEPENDENT *t* TEST WITH UNEQUAL SAMPLE SIZES

In the formula for the standard error that was provided for the independent samples *t* test, you may have noticed that we simply combined the standard errors of the means for each sample together. This is known as the pooled variance method. This method works well when the two sample sizes are equal (or roughly equal). But if the sample sizes are not equal (and they often are not in social science research), just pooling the variances of the two samples is not accurate. There needs to be an adjustment to the standard error formula used in the independent *t* test to account for different sample sizes. This formula for calculating the standard error of the differences between the independent sample means includes an adjustment, or weighting, for unequal sample sizes:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

In this formula, SS_1 and SS_2 represent the sum of squares for each sample. The sum of squares is shorthand for the sum of the squared deviations from the mean. We discussed how to calculate the sum of squared deviations from the mean back in Chapter 3 when we learned about variance and standard deviation. The sum of squares is discussed again in Chapter 9 when we learn about ANOVA. For now, notice that the right side of this formula includes adjustments for sample sizes by building in the $1/n_1$ and $1/n_2$.

Large differences between the sample sizes are most problematic when the assumption of homogeneity of variance is violated. The results of an independent *t* test can be quite misleading when the variances between the two groups are unequal and the sample sizes are quite different. When the sample sizes are similar, unequal variances are not a serious problem. But when the variances are unequal and the sample sizes are unequal, it is wise to adjust the effective degrees of freedom using the Welch–Satterthwaite equation. (In an SPSS analysis, this adjustment to the degrees of freedom is made automatically when the program detects unequal variances between the groups.)

Determining the Significance of the *t* Value for an Independent Samples *t* Test

Once we calculate the standard error and plug it into our formula for calculating the *t* value, we are left with an *observed t* value. How do we know if this *t* value is statistically significant? In other words, how do we decide if this *t* value is large enough to indicate that the difference between my sample means probably represents a real difference between my population means? To answer this question, we must find the probability of getting a *t* value of that size by chance. In other words, what are the odds that the difference between my two sample means is just due to the luck of the draw when I selected these two samples at random, rather than some real difference between the two populations? Fortunately, statisticians have already calculated these odds for us, and a table with such odds is included in Appendix B. Even more fortunately, statistical software programs used on computers calculate these probabilities for us, so there will hopefully never be a need for you to use Appendix B. I provide it here because I think the experience of calculating a *t* value by hand and determining whether it is statistically significant can help you understand how *t* tests work.

In Chapter 5, we saw how statisticians generated probabilities based on the normal distribution. With *t* distributions, the exact same principles are involved, except that now we have to take into account the size of the samples we are using. This is because the shape of the *t* distribution changes as the sample size changes, and when the shape of the distribution changes, so do the probabilities associated with it. The way that we take the sample size into account in statistics is to calculate degrees of freedom (*df*). The explanation of exactly what a degree of freedom is may be a bit more complicated than is worth discussing here (although you can read about it in most statistics textbooks if you are interested). At this point, suffice it to say that in an independent samples *t* test, you find the degrees of freedom by adding the two sample sizes together and subtracting 2. So the formula is $df = n_1 + n_2 - 2$. Once you have your degrees of freedom and your *t* value, you can look in the table of *t* values in Appendix B to see if the difference between your two sample means is significant.

To illustrate this, let's return to our example comparing men's and women's enjoyment of the new television program. Let's just suppose that the standard error of the difference between the means is .40. When I plug this number into the *t* value formula, I get the following:

$$t = \frac{7.5 - 6.5}{.40}$$

$$t = \frac{1.0}{.40} = 2.50$$

$$df = 50 + 50 - 2 = 98$$

Now that we have a *t* value and our degrees of freedom, we can look in Appendix B to find the probability of getting a *t* value of this size ($t = 2.50$) or larger by chance when we have 98 degrees of freedom. Because 98 degrees of freedom is between 60 and 120, I will look in the $df = 60$ row to be on the safe side. Choosing the smaller degrees of freedom gives me a more

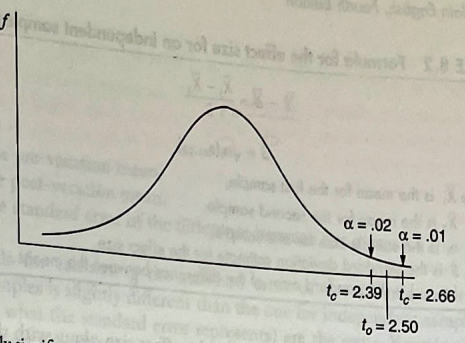

FIGURE 8.1 A statistically significant observed *t* value that falls between critical *t* values for alpha levels of .01 and .02.

conservative test (meaning that if my result is statistically significant at the $df = 60$ level, it will certainly be significant at the $df = 98$ level). Looking across the $df = 60$ row, and using the values for a two-tailed test, I can see that my observed *t* value ($t_o = 2.50$) is between the critical *t* values of 2.390 and 2.660. I can see that the alpha levels associated with these two critical *t* values in Appendix B are .02 and .01. Therefore, my table tells me that the probability of getting a *t* value this large by chance (i.e., due strictly to random sampling) is between 1 percent and 2 percent. In other words, when we randomly select two samples of 50 each from two different populations, we would expect to have a *t* value of this size less than 2 percent of the time *when there is no real difference between the population means* (for a more thorough discussion of this issue, see Chapter 7). Because this is such a small probability, I conclude that the difference between my sample of 50 men and my sample of 50 women that I observed in the average ratings of enjoyment of the television show probably represents a real difference between the larger populations of men and women, rather than some fluke difference that emerged simply because of who I happened to get in my samples (i.e., *random sampling error*; see Figure 8.1 for an illustration of where the observed *t* value falls in comparison to critical *t* values for alpha levels of .01 and .02).

It is important to remember that although this difference between the means was *statistically* significant (if we were using an alpha level of .05), that does not necessarily mean that it is *practically* significant (refer to the discussion about effect size in Chapter 7). Just as the standard error of the mean is influenced by the size of the sample, the standard error of the *difference* between the means is also affected by the sample size. The larger the samples, the smaller the standard error and the more likely it is that you will find a statistically significant result. To determine whether this difference between men and women is *practically* significant, we should consider the actual *raw score* difference. Men in our sample scored an average of 1 point higher on a 10-point scale than did women. Is that a big difference? Well, that is a judgment call. I would consider that a fairly inconsequential difference because we are talking about preferences for a television show. I don't consider a 1-point difference on a 10-point scale regarding television preferences to be important. But potential advertisers might consider this a meaningful difference. Those wanting to advertise female-oriented products may not select this show, which seems to appeal more to male viewers.

Another way to determine whether this difference in the means is practically significant is to calculate an effect size. The formula for the effect size for an independent samples *t* test is presented in Table 8.2. To calculate the effect size, you must first calculate the denominator. Using our example where the sample size for one group is 50 and the standard error of the difference between the means is .40, we get the following:

$$\hat{S} = \sqrt{50}(.40)$$
$$\hat{S} = 7.07(.40)$$
$$\hat{S} = 2.83$$

TABLE 8.2 Formula for the effect size for an independent samples *t* test

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{s}}$$

$$\hat{S} = \sqrt{n_1}(s_{\bar{x}1-\bar{x}2})$$

where $\bar{X}_1$ is the mean for the first sample,
$\bar{X}_2$ is the mean for the second sample,
n_1 is the sample size for one sample,
$\hat{s}$ is the standard deviation estimate for the effect size,
$(s_{\bar{x}1-\bar{x}2})$ is the standard error of the difference between the means.

We can then plug this into the formula for the effect size, along with the two sample means:

$$d = \frac{7.5 - 6.5}{2.83} \Rightarrow d = .35$$

So our effect size for this problem is .35, which would be considered a small- to medium-size effect.

Paired or Dependent Samples *t* Tests in Depth

Most of what I wrote before about the independent samples *t* test applies to the paired or dependent samples *t* test as well. We are still interested in determining whether the difference in the means that we observe in some sample(s) on some variable represents a true difference in the population(s) from which the sample(s) were selected. For example, suppose I wanted to know whether employees at my widget-making factory are more productive after they return from a two-week vacation. I randomly select 30 of my employees and calculate the average number of widgets made by each employee during the week before they go on vacation. I find that, on average, my employees made 250 widgets each during the week. During the week after they return from vacation, I keep track of how many widgets is made by *the same sample* of 30 employees and find that, on average, they made 300 widgets each during the week after returning from their vacations.

Just as with the independent samples *t* test, here I am concerned not only with whether this sample of 30 employees made more or fewer widgets after their vacation. I can look at the pre-vacation and post-vacation averages and see that these 30 employees, on average, made an average of 50 more widgets a week after their vacation. That is quite a lot. But I also want to know whether what I observed in this sample represents a likely difference in the productivity of the larger population of widget makers after a vacation. In other words, is this a statistically significant difference? The only real distinction between this dependent samples *t* test and the independent samples *t* test is that rather than comparing two samples on a single dependent variable, now I am comparing the average scores of a single sample (i.e., the same group of 30 employees) on two variables (i.e., pre-vacation widget-making average and post-vacation widget-making average). To make this comparison, I will again need to conduct a *t* test in which I find the difference between the two means and divide by the standard error of the difference between two *dependent sample* means. This equation looks like this:

$$t = \frac{\text{observed difference between post-vacation means}}{\text{standard error of the difference between the means}}$$

or

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}}$$

where $\bar{X}$ is the pre-vacation mean,
$\bar{Y}$ is the post-vacation mean,
$s_{\bar{D}}$ is the standard error of the difference between the means.

The formula for calculating the standard error of the difference between the means for dependent samples is slightly different than the one for independent samples, but the principles involved (i.e., what the standard error represents) are the same. Keep in mind that if I were to continually randomly select a samples of 30 widget makers and compare their pre-vacation and post-vacation productivity, I could generate a distribution of difference scores. For some samples, there would be no difference between pre-vacation and post-vacation productivity. For others, there would be increases in productivity and for still other samples there would be decreases in productivity. This distribution of difference scores (i.e., differences between pre-vacation and post-vacation averages) would have a mean and a standard deviation. The standard deviation of this distribution would be the standard error of the differences between dependent samples. The formula for this standard error is presented below in Table 8.3.

As you can see in Table 8.3, the easiest way to find the standard error is to follow a two-step process. First, we can find the standard deviation of the difference scores for my sample. Then we can divide this by the square root of the sample size to find the standard error. This formula is very similar to the formula for finding the standard error of the mean.

Another difference between dependent and independent samples *t* tests can be found in the calculation of the degrees of freedom. Whereas we had to add the two samples together and subtract 2 in the independent samples formula, for dependent samples we find the number of pairs of scores and subtract 1. In our example of widget makers, we have 30 pairs of scores because we have two scores for each person in the sample (one pre-vacation score and one post-vacation score). In the case of a paired *t* test where we have two paired samples (e.g., fathers and their sons), we use the same formula for calculating the standard error and the degrees of freedom. We must simply remember to match each score in one sample with a corresponding score in the second sample (e.g., comparing each father's score with only his son's score).

Once we've found our *t* value and degrees of freedom, the process for determining the probability of finding a *t* value of a given size with a given number of degrees of freedom is exactly the same as it was for the independent samples *t* test.

TABLE 8.3 Formula for the standard error of the difference between dependent sample means

$$\text{Step 1: } s_{\bar{D}} = \frac{s_D}{\sqrt{N}}$$

$$\text{Step 2: } s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{N}}{N-1}}$$

where $s_{\bar{D}}$ is the standard error of the difference between dependent sample means,
s_D is the standard deviation of the difference between dependent sample means,
D is the difference between each pair of X and Y scores (i.e., $X - Y$),
N is the number of pairs of scores.

Example 1: Comparing Boys' and Girls' Grade Point Averages

To illustrate how *t* tests work in practice, I provide one example of an independent samples *t* test and one of a dependent samples *t* test using data from a longitudinal study conducted by Carol Midgley and her colleagues. In this study, a sample of students was given surveys each year for several years beginning when the students were in the fifth grade. In the examples that follow, I present two comparisons of students' grade point averages (GPAs). The GPA is an average of students' grades in the four core academic areas: math, science, English, and social studies. Grades were measured using a 14-point scale with 13 = "A+" and 0 = "F".

In the first analysis, an independent samples *t* test was conducted to compare the average grades of sixth-grade boys and girls. This analysis was conducted using SPSS computer software. Thankfully, this program computes the means, standard error, *t* value, and probability of obtaining the *t* value by chance. Because the computer does all of this work, there is nothing to compute by hand, and I can focus all of my energy on interpreting the results. I present the actual results from the *t* test conducted with SPSS in Table 8.4.

SPSS presents the sample sizes for boys (n = 361) and girls (n = 349) first, followed by the mean, standard deviation ("SD"), and standard error of the mean ("SE of mean") for each group. Next, SPSS reports the actual difference between the two sample means ("Mean Difference = −1.5604"). This mean difference is negative because boys are the X_1 group and girls are the X_2 group. Because girls have the higher mean, when we subtract the girls' mean from the boys' mean (i.e., $\bar{X}_1 - \bar{X}_2$) we get a negative number. Below the mean difference we see "Levene's Test for Equality of Variances."[1] We can see that the *F* value for this test equals .639 and p = .424. This test tells us that there is not a significant difference between the variances of the two groups on the dependent variable (GPA) because our p value is larger than .05. Below the test for equality of variances, SPSS prints two lines with the actual *t* value (−7.45), the degrees of freedom ("*df*" = 708), the p value ("Sig. 2-Tailed" = .000), and the standard error of the difference between the means ("SE of Diff." = .210 and .209). These two lines of statistics are presented separately depending on whether we have equal or unequal variances. (See the "Time Out for Technicality" box for more information about the effects of unequal variances and sample sizes in independent *t* tests.) Because we had equal variances (as determined by Levene's test), we should interpret the top line, which is identified by the "Equal" name in the left column. Notice that these two lines of statistics are almost identical. That is because the variances are not significantly different between the two groups, and because the sample sizes were almost equal. If the variances and sample sizes were more different, the statistics presented in these two lines would differ more dramatically.

TABLE 8.4 SPSS results of independent samples *t* test

Variable	Number of Cases	Mean	SD	SE of Mean
Sixth-Grade GPA				
Male	361	6.5783	2.837	.149
Female	349	8.1387	2.744	.147

Mean Difference = −1.5604
Levene's Test for Equality of Variances: F = .639, p = .424

	t Test for Equality of Means			
Variances	*t* Value	*df*	2-Tailed Sig.	SE of Diff.
Equal	−7.45	708	.000	.210
Unequal	−7.45	708.00	.000	.209

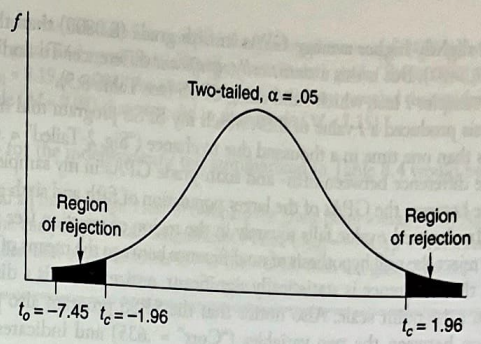

FIGURE 8.2 Results of a statistically significant *t* value for an independent *t* test comparing the average GPAs of boys and girls. Regions of rejection, critical *t* values, and observed *t* value for a two-tailed, independent samples *t* test are shown.

If we take the difference between the means and divide by the standard error of the difference between the independent sample means, we get the following equation for *t*:

$$t = -1.5604 \div .210$$

$$t = -7.45$$

The probability of getting a *t* value of −7.45 with 708 degrees of freedom is very small, as our p value ("Sig. Two-Tailed") of .000 reveals. Because *t* distributions are symmetrical (as are normal distributions), there is the exact same probability of obtaining a given negative *t* value by chance as there is of obtaining the same positive *t* value. For our purposes, then, we can treat negative *t* values as absolute numbers. (If you were testing a one-tailed alternative hypothesis, you would need to take into account whether the *t* value is negative or positive. See Chapter 7 for a discussion of one-tailed and two-tailed tests.)

The results of the *t* test presented in Table 8.4 indicate that our sample of girls had higher average GPAs than did our sample of boys, and that this difference was statistically significant. In other words, if we kept randomly selecting samples of these sizes from the larger populations of sixth-grade boys and girls and comparing their average GPAs, the odds of finding a difference between the means that is this large *if there is no real difference between the means of the two populations* is .000. This does not mean there is absolutely no chance. It just means that SPSS does not print probabilities smaller than .001 (e.g., .00001). Because this is such a small probability, we conclude that the difference between the two sample means probably represents a genuine difference between the larger populations of boys and girls that these samples represent. Notice in Figure 8.2 that this observed *t* value falls in the region of rejection, further indication that we should reject the null hypothesis of no difference between the means of boys and girls. Girls have *significantly* higher GPAs than boys (see Figure 8.2). Reminder: Statistical significance is influenced by sample size. Our sample size was quite large, so a difference of about 1.56 points on a 14-point scale was statistically significant. But is it practically significant? You can compute an effect size to help you decide.

Example 2: Comparing Fifth- and Sixth-Grade GPAs

Our second example involves a comparison of students' GPAs in fifth grade (when children are usually 10 or 11 years old) with the same sample's GPAs a year later, at the end of sixth grade. For each student in the sample (n = 689), there are two scores: one GPA for fifth grade and one GPA for sixth grade. This provides a total of 689 pairs of scores, and leaves us with 688 degrees of freedom (*df* = number of pairs − 1). A quick glance at the means reveals that, in this sample,

students had slightly higher average GPAs in fifth grade (8.0800) than they did a year later in sixth grade (7.3487). But is this a *statistically significant* difference? To find out, we must conduct a dependent samples *t* test, which I did using SPSS (see Table 8.5).

This analysis produced a *t* value of 8.19, which my SPSS program told me had a probability of occurring less than one time in a thousand due to chance ("Sig. 2-Tailed" = .000). Therefore, I conclude that the difference between fifth- and sixth-grade GPAs in my sample probably represents a real difference between the GPAs of the larger population of fifth and sixth graders that my sample represents. My observed *t* value falls squarely in the region of rejection (see Figure 8.3), indicating that I should reject the null hypothesis of no difference between the means of fifth and sixth graders.

Although this difference is statistically significant, notice that it is a difference of only about .73 points on a 14-point scale. Also notice that the SPSS program also provides a measure of the correlation between the two variables ("Corr" = .635) and indicates that this correlation coefficient is statistically significant. This tells you that students' fifth-grade GPAs are strongly related to their sixth-grade GPAs, as you might expect. Finally, notice that at the bottom left of Table 8.5, the difference between the means ("Paired Differences Mean"), the standard deviation of the difference between the means ("SD"), and the standard error of the difference between the means ("SE of Mean") are presented. The differences between the means divided by the standard error of the difference between the means produces the *t* value.

Writing it Up

Writing up *t* test results for publication is generally similar for independent, dependent, and single-sample *t* tests. Usually, what gets reported are the means for the groups being compared, the *t* value, and the degrees of freedom (*df*). The write-up for the results of the paired *t* test described in Table 8.5 and Figure 8.3 would be as follows:

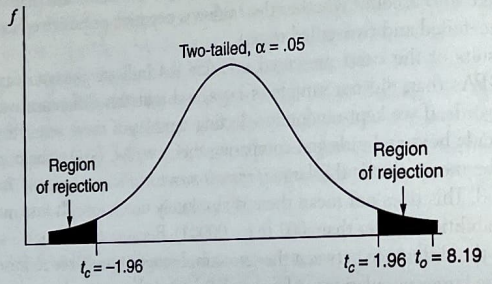

FIGURE 8.3 Region of rejection, critical *t* value, and observed *t* value for a two-tailed, dependent samples *t* test comparing fifth-grade and sixth-grade GPAs of one sample of students.

TABLE 8.5 SPSS results for dependent samples *t* test comparing fifth- and sixth-grade GPA

Variable	Number of Pairs	Corr	2-Tail Sig.	Mean	SD	SE of Mean
Fifth-Grade GPA	689	.635	.000	8.0800	2.509	.096
Sixth-Grade GPA				7.3487	2.911	.111

Paired Differences					
Mean	SD	SE of Mean	*t* Value	*df*	2-Tailed Sig.
.7312	2.343	.089	8.19	688	.000

"A paired *t* test was calculated to compare the grade point averages (GPAs) of students when they were in fifth grade and a year later when they were in sixth grade. The analysis produced a significant *t* value ($t_{(688)} = 8.19, p < .001$). An examination of the means revealed that students had higher GPAs in fifth grade ($M = 8.08$) than they did in sixth grade ($M = 7.35$)."

The write-up for the independent *t* test summarized in Table 8.4 would be very similar:

"I performed an independent *t* test to compare the grade point averages (GPAs) of sixth-grade boys and girls. The analysis produced a significant *t* value ($t_{(708)} = -7.45, p < .001$). An examination of the means revealed that boys had lower GPAs ($M = 6.58$) than did girls ($M = 8.14$)."

Worked Examples

In this section I present three worked examples of *t* tests: a one-sample *t* test, an independent samples *t* test, and a paired or dependent samples *t* test.

One-Sample *t* Test

For the one-sample *t* test, suppose I want to know whether students at my university differ from other students in the country in the amount of time they spend on school work outside of class. I happen to know that in the U.S., university students spend an average of 15 hours per week working on school work outside of class. I select a random sample of 36 students from my university and find that they spend an average of 14 hours per week on school work, with a standard deviation of 3 hours. Is the difference between the sample mean and the population mean significant, either practically or statistically?

The first step in the process of answering this question is to realize that I am actually trying to determine whether the *population* that my sample represents differs from the population of students at universities in the U.S. in terms of hours spent on school work outside of class. So my null hypothesis is that the population mean of students at my university will equal the population mean of students at universities in the U.S. Because this is a two-tailed test (i.e., my research question is whether the means *differ*), my alternative hypothesis is that the population mean of students at my university will differ (either be larger or smaller) than the population mean of university students in the U.S.

Next, I calculate a standard error of the mean using the sample size (36) and the sample standard deviation (3):

$$s_{\bar{x}} = \frac{3}{\sqrt{36}}$$

$$s_{\bar{x}} = \frac{3}{6}$$

$$s_{\bar{x}} = .5$$

Now that I have my standard error of the mean, I can calculate my observed *t* value:

$$t = \frac{14 - 15}{.5} = -2.0$$

To determine whether my results are statistically significant, I compare the *t* value that I calculated to a critical *t value* that I will look up in Appendix B. Remember, with a two-tailed test, we can use the absolute value of the observed *t* value, so I will compare the critical *t* value that I find in Appendix B with an observed *t* value of 2.0. With a sample size of 36 and one sample,

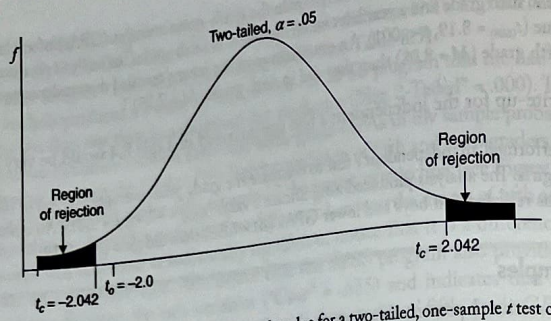

FIGURE 8.4 Regions of rejection, critical *t* values, and observed *t* value for a two-tailed, one-sample *t* test comparing a sample and population of college students on the number of hours spent studying.

my degrees of freedom will be 36 – 1 = 35. Appendix B lists *t* values for 30 *df* or 40 *df*, so I will use those two and find the average. With an alpha level of .05, I find a critical *t* value of 2.042 for *df* = 30 and 2.021 for a *df* = 40. The average of those two values is about 2.031 for a *df* of 35, and this critical *t* value is larger than my observed *t* value of 2.0. Therefore, using a two-tailed alpha level of .05, the difference between my sample mean and the population mean is *not* statistically significant. There is no difference between the average number of hours spent on school work outside of class between the students in my university and the university students in the U.S. population. (See Figure 8.4 for an illustration of the two-tailed regions of rejection, critical *t* values, and my observed *t* value.)

I also want to report an effect size for this *t* test, so I will calculate a Cohen's *d* value:

$$d = \frac{114 - 115}{3} = -.33$$

The Cohen's *d* is expressed in standard deviation units (Cohen, 1988). This effect size of .33 is considered a weak-to-moderate effect size, meaning that the difference between the means has little practical significance. Combined with the lack of statistical significance, we would conclude that the population of students at my university does not really differ from the larger population of university students in the U.S. on this variable.

Independent Samples *t* Test

Suppose I work for a marketing firm and I want to know whether men and women differ in their enjoyment of a new movie that my client has produced. I select random samples of 100 men and 100 women, show them the movie, and ask them to rate, on a 10-point scale, how much they enjoyed the movie. The average movie-enjoyment rating for my sample of men was 7 with a standard deviation of 3, and the average for women was 6 with a standard deviation of 4. Is this a meaningful difference between the means?

The first step in solving this problem is to calculate the standard error of the difference between the means. This involves finding the standard errors of the means for each sample:

$$\text{For men: } s_{\bar{x}} = \frac{3}{\sqrt{25}} = .60$$

$$\text{For women: } s_{\bar{x}} = \frac{4}{\sqrt{25}} = .80$$

Now that I have calculated the standard errors of the mean for each sample, I need to square each one and add them together:

$$.60^2 = .36$$

$$.80^2 = .64$$

$$.36 + .64 = 1$$

The final step in the process of calculating the standard error of the difference between the means is to calculate the square root of this sum:

$$s_{\bar{x}1-\bar{x}2} = \sqrt{1} = 1$$

Once I've calculated the standard error of the difference between the sample means, I can plug that into my formula for calculating the observed *t* value:

$$t = \frac{7 - 6}{1} = 1.0$$

Notice that in this *t* value formula I placed the mean for men first, creating a positive *t* value. Had I placed the mean for women first and produced a negative *t* value, it would not have mattered as the research question created a two-tailed test (i.e., testing whether there is a *difference* between the means without any statement about the direction of that difference).

To determine whether my observed *t* value is statistically significant, I will compare it to a critical *t* value that I find in Appendix B. The degrees of freedom for an independent samples *t* test is $n + n - 2$, so in this situation we have *df* = 25 + 25 – 2 = 48. In Appendix B, the closest I can get to a *df* of 48 is a *df* of 40, so I will look in that row for my critical *t* value. Using an alpha level of .05 and a two-tailed test, the critical *t* value is 2.021. Because my $t_c > t_o$, I conclude that the difference between the sample means of men and women is *not* statistically significant. Therefore, I conclude that there is no difference between the population means of men and women in their enjoyment of this movie. I will market the movie to both genders. (See Figure 8.5 for an illustration of the regions of rejection, the critical *t* values, and the observed *t* value for this analysis.)

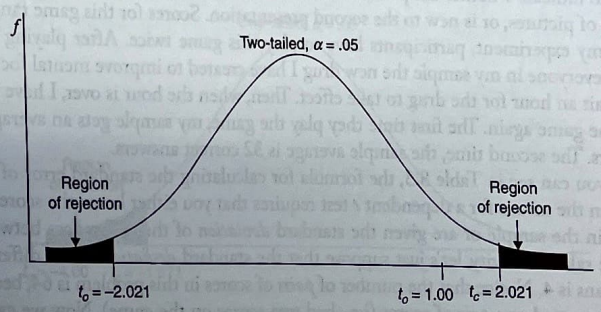

FIGURE 8.5 Regions of rejection, critical *t* values, and observed *t* value for a two-tailed, independent samples *t* test comparing men and women on average enjoyment of a movie.

To provide more information about the differences between these sample means, I will calculate an effect size (d) and a confidence interval. First, I need to calculate the standard deviation of the estimate of the effect size:

$$\hat{S} = \sqrt{25}(1) = 5$$

Then I plug that into my effect size formula to find d:

$$d = \frac{7-6}{5} = .20$$

This is quite a small effect size, supporting our previous finding of no statistically significant difference between the means of men and women. Finally, let's calculate a 95 percent confidence interval for the difference between the means:

$$CI_{95} = (7 - 6) \pm (1)(2.021)$$

$$CI_{95} = 1 \pm 2.021$$

$$CI_{95} = -1.021, 3.021$$

This confidence interval tells us that we can be 95 percent confident that the actual difference between the population means of men and women in their enjoyment of this movie is between women rating their enjoyment of the movie about 1 point higher than men up to men rating their enjoyment of the movie about 3 points higher than women. This is a fairly large range (about 4 points on a 10-point scale) that includes the value of 0 (i.e., no difference between the population means), again indicating that there are no reliable differences between the population means on this variable.

Dependent/Paired *t* Test

Suppose that I have developed a drug that is supposed to improve mental focus and memory. To test whether it works, I select a random sample of 64 adults and ask them to play a game where they are presented with a set of 50 pictures, one at a time, on a computer screen. After a 5-minute break, they are presented with another set of 50 pictures, some of which appeared in the first set that was presented. During this second presentation of pictures, my participants have to indicate whether the picture shown on the screen was also presented in the first presentation of pictures, or is new to the second presentation. Scores for this game range from 0 to 50.

In my experiment, participants have to play this game twice. After playing the game once, I give everyone in my sample the new drug I have created to improve mental focus and memory, and wait an hour for the drug to take effect. Then, when the hour is over, I have my participants play the game again. The first time they play the game, my sample gets an average of 30 correct answers. The second time, the sample average is 32 correct answers.

As you can see in Table 8.3, the formula for calculating the standard error of the differences between the means for a dependent *t* test requires that you either know the scores for each individual in the sample or are given the standard *deviation* of the differences between the means. For the sake of brevity, let's just suppose that the standard *deviation* of the differences between the means is 4. Notice that the number of *pairs* of scores in this problem is 64, because each participant created one *pair* of scores (i.e., had two scores on the game). Now we can calculate the standard *error* of the difference between the mean:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{N}}$$

$$s_{\bar{D}} = \frac{4}{\sqrt{64}} = .5$$

Now I can plug my sample means and the standard error into the *t* value formula to find my observed *t* value. Note that scores from the first test are my *X* variable and scores from the second test are my *Y* variable:

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}}$$

$$t = \frac{30 - 32}{.5} = -4.0$$

The next step is to determine whether this is a statistically significant result. In this situation, it is reasonable to argue that this is a one-tailed test. I created the drug assuming it would enhance focus and memory. If this assumption is true, I would expect the test scores to be higher after taking the drug (i.e., the *Y* variable) than before taking the drug (the *X* variable). So in this situation, my null and alternative hypotheses would be as follows:

Null hypothesis (H_0): $\mu_X = \mu_Y$
Alternative hypothesis (H_A): $\mu_X < \mu_Y$

Because the formula for calculating the *t* value for the dependent *t* test involves subtracting the mean of *Y* from the mean of *X*, my alternative hypothesis would produce a negative *t* value. For a one-tailed test, I am only going to be looking at the negative side of the distribution and will only consider my results to be statistically significant if my observed *t* value is *less than* my critical *t* value. The critical *t* value, from Appendix B, with 60 *df*, an alpha level of .05, and a one-tailed test is −1.671. This critical value is *larger* than my observed *t* value of −4.0. In other words, my observed *t* value is *beyond* (i.e., further out in the tail than) my critical *t* value. Therefore, my results *are* statistically significant, and I conclude that, in the population of adults, test scores are higher, on average, after taking my new drug than before taking it. (See Figure 8.6

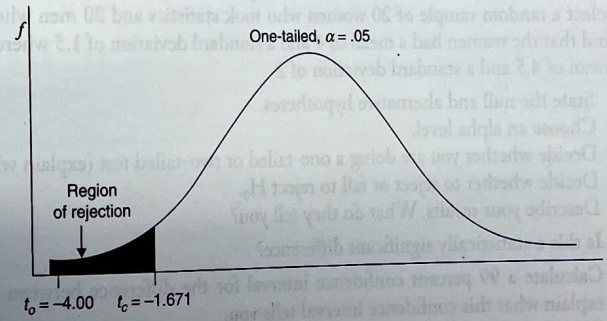

FIGURE 8.6 Region of rejection, critical *t* value, and observed *t* value for a one-tailed, dependent samples *t* test comparing test scores of adults before and after taking my new memory drug.

for an illustration of the one-tailed region of rejection, the critical t value, and the observed t value for this analysis.)

For brief videos demonstrating how to calculate one-sample, independent, and dependent t tests, see the website that accompanies this book.

Wrapping Up and Looking Forward

The two types of t tests described in this chapter share two things in common. First, they both test the equality of means. Second, they both rely on the t distribution to produce the probabilities used to test statistical significance. Beyond that, these two types of t tests are really quite different. The independent samples t test is used to examine the equality of means from two independent groups. Such a test has much in common with one-way ANOVA (Chapter 9) and factorial ANOVA (Chapter 10). In contrast, the dependent samples t test is used to examine whether the means of *related* groups, or of two variables examined within the same group, are equal. This test is more directly related to repeated-measures ANOVA as discussed in Chapter 11.

Work Problems

1. I want to know whether Californians, who tend to be pretty health conscious, eat fewer hot dogs than the general population of Americans. Suppose that in the population of adults, the average number of hot dogs eaten per year is 15. I select a random sample of 25 adults from California and find that they eat an average of 17 hot dogs with a standard deviation of 3.
   a. State the null and alternative hypotheses.
   b. Choose an alpha level.
   c. Decide whether you are doing a one-tailed or two-tailed test (explain why).
   d. State your degrees of freedom.
   e. Find and report your critical value for t.
   f. Compute your observed t value.
   g. Decide whether to reject or fail to reject H_0.
   h. Describe your results. Is this a statistically significant difference?
2. Suppose you want to know whether men and women who take a college statistics class differ in their enjoyment of statistics. Enjoyment of statistics is measured by asking people to rate "How much do you enjoy statistics?" on a 10-point scale (1 = *"I hate it"* and 10 = *"I love it"*). I select a random sample of 20 women who took statistics and 20 men who took statistics. I find that the women had a mean of 4 and a standard deviation of 1.5 whereas the men had a mean of 4.5 and a standard deviation of 2.
   a. State the null and alternative hypotheses.
   b. Choose an alpha level.
   c. Decide whether you are doing a one-tailed or two-tailed test (explain why).
   d. Decide whether to reject or fail to reject H_0.
   e. Describe your results. What do they tell you?
   i. Is this a statistically significant difference?
   f. Calculate a 99 percent confidence interval for the difference between the means and explain what this confidence interval tells you.
   g. Calculate an effect size (d) and describe what it means.

3. I think people are getting taller. I select a random sample of 15 fathers and their adult sons to take part in a study. I find that the fathers have an average height of 71 in. and their sons have an average height of 73 in. The standard *deviation* of the difference between the means is 1.3.
   a. Write the null and alternative hypotheses.
   b. Select an alpha level.
   c. Is this a statistically significant difference? (Do the calculations and explain what your results mean.)

For answers to these work problems, and for additional work problems, see the website that accompanies this book.

Note

1 As noted in the "Time Out for Technicality" box, unequal variances between the two samples on the dependent variable in the independent t test can create unreliable probability estimates. When unequal variances are detected, the degrees of freedom must be adjusted. SPSS uses Levene's test to determine whether the sample variances are equal and adjusts the degrees of freedom automatically, as the example presented in Table 8.4 indicates.

CHAPTER 9

One-Way Analysis of Variance

The purpose of a one-way analysis of variance (**one-way ANOVA**) is to compare the means of two or more groups (the independent variable) on one dependent variable to see if the group means are significantly different from each other. In fact, if you want to compare the means of two independent groups on a single variable, you can use either an independent samples *t* test or a one-way ANOVA. The results will be identical, except that instead of producing a *t* value, the ANOVA will produce an ***F*** ratio, which is simply the *t* value squared (more about this in the next section of this chapter). Because the *t* test and the one-way ANOVA produce identical results when there are only two groups being compared, most researchers use the one-way ANOVA only when they are comparing three or more groups. To conduct a one-way ANOVA, you need to have a categorical (or nominal) variable that has at least two independent groups (e.g., a race variable with the categories African-American, Latino, and Euro-American) as the independent variable and a continuous variable (e.g., achievement test scores) as the dependent variable. It is assumed in ANOVA that the variance in the dependent variable is equal in each of the groups being compared.

In many fields of research, including the social sciences, one-way ANOVA is an extremely popular statistical technique. Experiments that involve comparisons among more than two groups often employ one-way ANOVA. For example, a researcher may conduct an experiment to examine whether a new treatment for diabetes is more effective than an existing treatment, or than a placebo. These three groups would be compared using one-way ANOVA. Researchers conducting correlational research (i.e., not manipulating the independent variable) also often use one-way ANOVA. For example, a researcher who wants to compare the academic performance of students from different countries might use one-way ANOVA. Because so much research involves comparing the means of multiple groups, there are thousands of studies across many disciplines that have used one-way ANOVA. In this chapter, we consider the most basic form of ANOVA, one-way ANOVA, involving one independent variable and one dependent variable. In Chapters 10 and 11, we consider more complex forms of ANOVA that involve multiple independent variables.

ANOVA vs. Independent *t* Tests

Because the independent *t* test and the one-way ANOVA are so similar, people often wonder, Why don't we just use *t* tests instead of one-way ANOVAs? Perhaps the best way to answer this question is by using an example. Suppose that I want to go into the potato chip business. I've got three different recipes, but because I'm new to the business and don't have a lot of money, I can produce only one flavor. I want to see which flavor people like best and produce that one. I randomly select 90 adults and randomly divide them into three groups. One group tries my BBQ-flavored chips, the second group tries my ranch-flavored chips, and the third group tastes my cheese-flavored chips. All participants in each group fill out a

rating form after tasting the chips to indicate how much they liked the taste of the chips. The rating scale goes from a score of 1 (*"Hated it"*) to 7 (*"Loved it"*). I then compare the average ratings of the three groups to see which group liked the taste of their chips the most. In this example, the chip flavor (BBQ, Ranch, Cheese) is my categorical, independent variable and the rating of the taste of the chips is my continuous, dependent variable.

To see which flavor received the highest average rating, I could run three separate independent *t* tests comparing (a) BBQ with Ranch, (b) BBQ with Cheese, and (c) Ranch with Cheese. The problem with running three separate *t* tests is that each time we run a *t* test, we must make a decision about whether the difference between the two means is meaningful, or statistically significant. This decision is based on probability, and every time we make such a decision, there is a slight chance we might be wrong (see Chapter 7 on statistical significance). The more times we make decisions about the significance of *t* tests, the greater the chances are that we will be wrong. In other words, the more *t* tests we run, the greater the chances become of us deciding that a *t* test is significant (i.e., that the means being compared are really different) when it really is not. In still other words, running multiple *t* tests increases the likelihood of making a Type I error (i.e., rejecting the null hypothesis when in fact it is true). A one-way ANOVA fixes this problem by adjusting for the number of groups being compared. To see how it does this, let's take a look at one-way ANOVA in more detail.

One-Way ANOVA in Depth

The purpose of a one-way ANOVA is to divide up the variance in some dependent variable into two components: the variance attributable to **between-group** differences, and the variance attributable to **within-group** differences, also known as *error*. When we select a sample from a population and calculate the mean for that sample on some variable, that sample mean is our best predictor of the population mean. In other words, if we do not know the mean of the population, our best guess about what the population mean is would have to come from the mean of a sample drawn randomly from that population. Any scores in the sample that differ from the sample mean are believed to include what statisticians call error. For example, suppose I have a sample of 20 randomly selected fifth graders. I give them a test of basic skills in math and find out that, in my sample, the average number of questions answered correctly on my test is 12. If I were to select one student in my sample and find that she had a score of 10 on the test, the difference between her score and the sample mean would be considered error, as indicated in Figure 9.1.

The variation that we find among the scores in a sample is not just considered error. In fact, it is thought to represent a specific kind of error: **random error**. When we select a sample at random from a population, we expect that the members of that sample will not all have identical

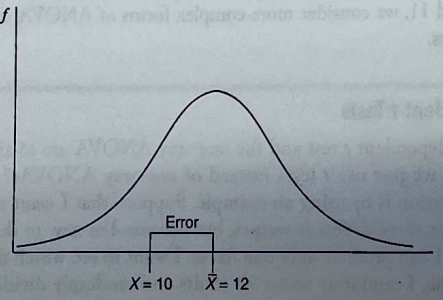

FIGURE 9.1 An example of within-group error.

TABLE 9.1 Formula for the *F* value

$$F = \frac{\text{mean square between}}{\text{mean square error}}$$

or

$$F = \frac{MS_b}{MS_e}$$

where *F* is the *F* value,
MS_b is the mean square between groups,
MS_e is the mean square error, or within group error.

scores on our variable of interest (e.g., test scores). That is, we expect that there will be some variability in the scores of the members of the sample. That's just what happens when you select members of a sample randomly from a population. Therefore, the variation in scores that we see among the members of our sample is just considered random error.

The question that we can address using ANOVA is this: Is the average amount of difference, or variation, between the scores of members of *different* samples large or small compared to the average amount of variation *within* each sample, otherwise known as random error (a.k.a. error)? To answer this question, we have to determine three things. First, we have to calculate the average amount of variation within each of our samples. This is called the **mean square error** (MS_e), also referred to as the **mean square within** (MS_w) in some textbooks and websites. Second, we have to find the average amount of variation *between* the groups. This is called the **mean square between** (MS_b). Once we've found these two statistics, we must find their ratio by dividing the mean square between by the mean square error. This ratio provides our ***F* value**, and when we have our *F* value we can look at our family of *F* distributions to see if the differences between the groups are statistically significant (see Table 9.1).

Note that, although it may sound like analysis of variance is a whole new concept, in fact it is virtually identical to the independent *t* test discussed in Chapter 8. Recall that the formula for calculating an independent *t* test also involves finding a ratio. The top portion of the fraction is the difference between two sample means, which is analogous to the mean square between (MS_b) just presented. The only differences between the two are (a) rather than finding a simple difference between two means as in a *t* test, in ANOVA we are finding the *average* difference between means, because we are often comparing more than two means; and (b) we are using the squared value of the difference between the means. (Notice that because we are squaring the values we use to calculate an *F* value, the *F* value will always be positive.) The bottom portion of the fraction for the *t* test is the standard *error* of the difference between two sample means. This is exactly the same as the *average*, or standard, error within groups. In the formula used to calculate the *F* value in ANOVA, we must square this average within-group error. So, just as in the *t* test, in ANOVA we are trying to find the average difference *between* group means relative to the average amount of variation *within* each group.

To find the MS_e and MS_b, we must begin by finding the **sum of squares error** (SS_e) and the **sum of squares between** (SS_b). This sum of squares idea is not new. It is the same sum of squares as introduced in Chapter 3 in the discussion about variance and standard deviation. Sum of squares is actually short for *sum of squared deviations*. In the case of ANOVA, we have two types of deviations. The first is the deviation between each score in a sample and the mean for that sample (i.e., error). The second type of deviation is between each sample mean and the mean for all of the groups combined, called the **grand mean** (i.e., between groups). These two types of deviations are presented in Figure 9.2. Notice that the deviation between the individual score (X_{3i}) is considered random error and is part of the SS_e, and the deviation between the group mean and the grand mean is part of the between-groups SS_b.

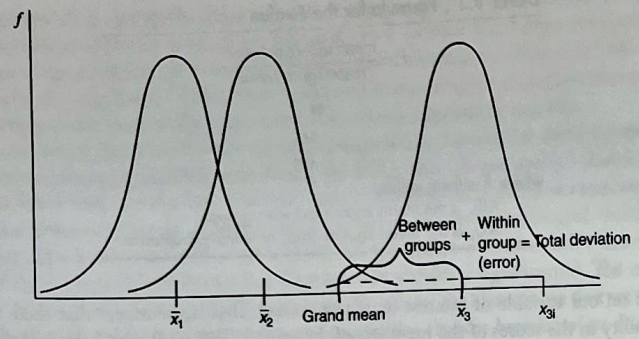

FIGURE 9.2 Illustrations of between-group and within-group deviations.

For a video explanation of how variance is partitioned in a one-way ANOVA, please refer to the website that accompanies this book.

To find the sum of squares error (SS_e):

1. Calculate the mean for each group: $\bar{X} = \frac{\Sigma X}{n}$
2. Subtract the group mean from each individual score in each group: $X - \bar{X}$
3. Square each of these deviation scores: $(X - \bar{X})^2$
4. Add them all up for each group: $\Sigma(X - \bar{X})^2$
5. Then add up all of the sums of squares for all of the groups:

$$\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2 + \ldots + \Sigma(X_k - \bar{X}_k)^2$$

Note: The subscripts indicate the individual groups, through to the last group, which is indicated by the subscript *k*.

The method used to calculate the sum of squares between groups (SS_b) is just slightly more complicated than the SS_e formula. To find the SS_b:

1. Subtract the grand mean from the group mean: $(\bar{X} - \bar{X}_T)$; *T* indicates total, or the mean for all of the cases across all of the groups
2. Square each of these deviation scores: $(\bar{X} - \bar{X}_T)^2$
3. Multiply each squared deviation by the number of cases in the group: $[n(\bar{X} - \bar{X}_T)^2]$
4. Add these squared deviations from each group together: $\Sigma\left[n(\bar{X} - \bar{X}_T)^2\right]$

The only real differences between the formulas for calculating the SS_e and the SS_b are:

1. In the SS_e we subtract the group mean from the individual scores in each group, whereas in the SS_b we subtract the grand mean from each group mean.
2. In the SS_b we multiply each squared deviation by the number of cases in each group. We must do this to get an approximate deviation between the group mean and the grand mean *for each case in every group.*

If we were to add the SS_e to the SS_b, the resulting sum would be called the **sum of squares total** (SS_T). A brief word about the SS_T is in order. Suppose that we have three randomly selected

samples of students. One is a sample of fifth graders, the second is a sample of eighth graders, and the third is a sample of eleventh graders. If we were to give each student in each sample a spelling test, we could add up the scores for all of the students in the three samples combined and divide by the total number of scores to produce one average score. Because we have combined the scores from all three samples, this overall average score would be called the grand mean, or total mean, which would have the symbol $\bar{X}_T$. Using this grand mean, we could calculate a squared deviation score for each student in all three of our samples combined using the familiar formula $(X - \bar{X})^2$. The interesting thing about these squared deviations is that, for each student, the difference between each student's score and the grand mean is the sum of that student's deviation from the mean of his or her own group and the deviation of that group mean from the grand mean. So, suppose Jimmy is in the fifth-grade sample. Jimmy gets a score of 25 on the spelling test. The average score for the fifth-grade sample is 30, and the average score for all of the samples combined (i.e., the grand mean) is 35. The difference between Jimmy's score (25) and the grand mean (35) is just the difference between Jimmy's score and the mean for his group (25 – 30 = –5) plus the difference between his group's mean and the grand mean (30 – 35 = –5). Jimmy's deviation from the grand mean is –10 (see Figure 9.3). If we square that deviation score, we end up with a squared deviation of 100 for Jimmy.

Now, if we calculated a deviation score for each student in all three samples and added up all of these deviation scores using the old $\Sigma(X - \bar{X}_T)^2$ formula, the result would be the sum of squares total, or the SS_T (Notice that this formula is the same one that we used way back in Chapter 3! It is the numerator for the variance formula!) The interesting thing about this SS_T is that it is really just the sum of the SS_b and the SS_e. $SS_T = SS_b + SS_e$. This makes sense, because, as we saw with Jimmy, the difference between any individual score and the grand mean is just the sum of the difference between the individual score and the mean of the group that the individual is from plus the difference between that group mean and the grand mean. This is the crux of ANOVA.

Deciding if the Group Means Are Significantly Different

Once we have calculated the SS_b and the SS_e, we have to convert them to average squared deviation scores, or MS_b and MS_e. This is necessary because there are far more deviation scores in the SS_e than there are in the SS_b, so the sums of squares can be a bit misleading. What we want to know in an ANOVA is whether the *average* difference between the group means is large or small relative to the *average* difference between the individual scores and their respective group means,

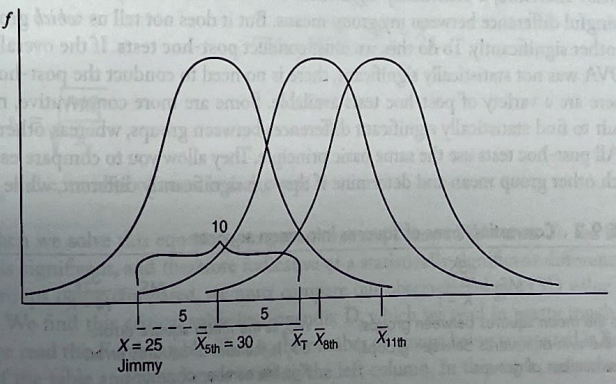

FIGURE 9.3 Within- and between-group deviations for a particular individual.

or the average amount of error within each group. To convert these sums of squares into mean squares, we must divide the sums of squares by their appropriate degrees of freedom.

For the SS_b, remember that we are only making comparisons between each of the groups. The number of degrees of freedom for the SS_b is always the number of groups minus 1. If we use K to represent the number of groups, and *df* to represent the degrees of freedom, then the formula for the between-group degrees of freedom is $df = K - 1$. So, to convert an SS_b to an MS_b, we divide the SS_b by $K - 1$. The number of degrees of freedom for the SS_e is found by taking the number of scores in each group and subtracting 1 from each group. So, if we have three groups, our *df* for the SS_e will be $(n_1 - 1) + (n_2 - 1) + (n_3 - 1)$. Notice that this is the same formula for the degrees of freedom as was used for the independent samples *t* test in Chapter 8. The only difference is that we have one more group here. A simpler way to write this *df* formula is $N - K$, where N is the total number of cases for *all* groups combined and K is the number of groups. Once we have this *df*, we can convert the SS_e into an MS_e by simply dividing the SS_e by $N - K$. Table 9.2 contains a summary of the formulas for converting the sums of squares into mean squares.

Once we have found our MS_b and our MS_e, all we have to do is divide the MS_b by the MS_e to find our *F* value. Once we've found our *F* value, we need to look in our table of *F* values (Appendix C) to see whether it is statistically significant. This table of *F* values is similar to the table of *t* values we used in Chapter 8, with one important difference. Unlike *t* values, the significance of *F* values depends on both the number of cases in the samples (i.e., the *df* for the MS_e) *and* the number of groups being compared (i.e., the *df* for the MS_b). This second *df* is critical, because it is what is used to control for the fact that we are comparing more than two groups. Without it, we might as well conduct multiple *t* tests, and this is problematic for the reasons discussed at the beginning of the chapter. In Appendix C, we can find critical values for *F* associated with different alpha levels. If our observed value of *F* (F_o) is larger than our critical value of *F* (F_c), we must conclude that there are statistically significant differences between the group means. In Figure 9.4, you can see that the observed *F* value is in the shaded area beyond the critical value of *F* that indicates the region of rejection, therefore indicating that we should reject the null hypothesis of no difference between the population means.

Post-Hoc Tests

Our work is not done once we have found a statistically significant difference between the group means. Remember that when we calculated the MS_b, we ended up with an *average* difference between the group means. If we are comparing three group means, we might find a relatively large average difference between these group means even if two of the three group means are identical. Therefore, a statistically significant *F* value tells us only that somewhere there is a meaningful difference between my group means. But it does not tell us *which* groups differ from each other significantly. To do this, we must conduct **post-hoc tests**. If the overall *F* value of the ANOVA was not statistically significant, there is no need to conduct the post-hoc tests.

There are a variety of post-hoc tests available. Some are more conservative, making it more difficult to find statistically significant differences between groups, whereas others are more liberal. All post-hoc tests use the same basic principle. They allow you to compare each group mean to each other group mean and determine if they are significantly different, while controlling for

TABLE 9.2 Converting sums of squares into mean squares

$MS_b = \frac{SS_b}{K-1}$	$MS_e = \frac{SS_e}{N-K}$
MS_b is the mean squares between groups.	MS_e is the mean square error.
SS_b is the sum of squares between groups.	SS_e is the sum of squares error.
K is the number of groups.	K is the number of groups.
	N is the number of cases combined across all groups.

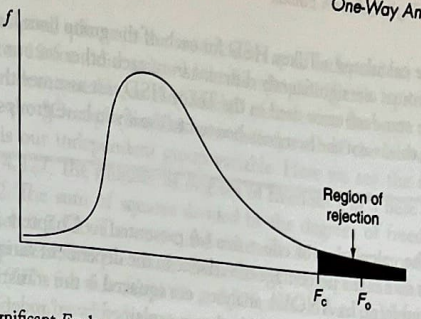

FIGURE 9.4 A statistically significant *F* value.

the number of group comparisons being made. Conceptually, post-hoc tests in ANOVA are like a series of independent *t* tests where each group mean is compared to each other group mean. Unlike *t* tests, however, post-hoc tests keep the Type I error rate constant by taking into consideration how many groups are being compared. As we saw in Chapters 7 and 8, to determine if the difference between two group means is statistically significant, we subtract one group mean from the other and divide by a standard error. The difference between the various types of post-hoc tests is what each test uses for the standard error. Some formulas are designed to take different sample sizes for different groups into account (e.g., Tukey–Kramer) and others are designed to handle situations where the variances are not equal across all of the groups in the independent variable (e.g., Dunnett's test, the Games–Howell test).

In this book, for the purposes of demonstration, we will consider the **Tukey HSD** (HSD stands for Honestly Significantly Different) post-hoc test. I chose the Tukey HSD test because it is one of the simpler formulas and it allows me to demonstrate how to calculate a post-hoc test by hand more simply than another, more complicated formula would. But the Tukey HSD is not better than other post-hoc tests and is only appropriate to use when the sample sizes and variances of all the groups in the ANOVA are equal. This is a fairly liberal test, meaning that it is more likely to produce statistically significant differences than some other tests (e.g., the Scheffé test). When tests are more liberal, they are more likely to produce statistically significant results, even when they should not (i.e., Type I errors). In general, more conservative tests allow researchers to be more confident that when they proclaim a result is statistically significant, they are not making a Type I error.

The Tukey test compares each group mean to each other group mean by using the familiar formula described for *t* tests in Chapter 8. Specifically, it is the mean of one group minus the mean of a second group divided by the standard error:

$$\text{Tukey HSD} = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}}}$$

where $s_{\bar{x}} = \sqrt{\frac{MS_e}{n_g}}$

and n_g = the number of cases *in each group*.

When we solve this equation, we get an *observed* Tukey HSD value. To see if this observed value is significant, and therefore indicative of a statistically significant difference between the two groups being compared, we must compare our observed Tukey HSD value with a critical value. We find this critical value in Appendix D, which we read in pretty much the same way that we read the *F*-value table. That is, the number of groups being compared is listed on the top row of the table and the *df* error is along the left column. In this table, only the critical values for an alpha level of .05 are presented.

Once we have calculated a Tukey HSD for each of the group comparisons we need to make, we can say which groups are significantly different from each other on our dependent variable. Notice that, because the standard error used in the Tukey HSD test assumes that each group has an equal number of cases, this is not the best post-hoc test to use if you have groups with unequal sample sizes.

Effect Size

In addition to the calculation of effect size (*d*) presented in Chapters 7 and 8, another common measure of effect size is the percentage of variance in the dependent variable that is explained by the independent variable(s). In ANOVA analyses, **eta squared** is the statistic that reveals the percentage of variance in the dependent variable that is explained by an independent variable, and is the most common measure of effect size. Remember that in a one-way ANOVA, the independent variable is always categorical. Therefore, the eta squared statistic tells us how much of the variance in the dependent variable can be explained by differences between groups in the independent variable. If I am comparing the average temperature in January (dependent variable) in Maine, California, and Sydney, Australia, I would expect a fairly large eta squared value because the average temperatures vary widely by location (independent variable). In other words, a large amount of the variation in January temperatures is explained by the location. (For a more detailed discussion of explained variance, see Chapters 12 and 13.) To illustrate how this works, I present the results of an analysis using the SPSS computer software program to analyze a set of fictional data that I made up.

Suppose that I want to test a drug that I developed to increase students' interest in their school work. I randomly select 75 third-grade students and randomly assign them to one of three groups: a "High Dose" group, a "Low Dose" group, and a "Placebo" group. After dividing the students into their respective groups, I give them the appropriate dosage of my new drug (or a placebo) and then give them all the exact same school work assignment. I measure their interest in the school work by asking them to rate how interesting they thought the work was on a scale from 1 (*Not interesting*) to 5 (*Very interesting*). Then I use SPSS to conduct an ANOVA on my data, and I get the output from the program presented in Table 9.3.[1]

The results produced by SPSS include descriptive statistics such as the means, standard deviations, and sample sizes for each of the three groups, as well as the overall mean ("Total") for the entire sample of 75 students. In the descriptive statistics, we can see that the "Low Dose" group has a somewhat higher average mean on the dependent variable (i.e., interest in the school work) than do the other two groups. Turning now to the ANOVA results below the descriptive statistics in Table 9.3, there are identical statistics for the "Corrected Model" row and the "Group" row. The

TABLE 9.3 SPSS output for ANOVA examining interest by drug treatment group

Descriptive Statistics

Independent Variable	Mean	Std. Deviation	N
High Dose	2.7600	1.2675	25
Low Dose	3.6000	1.2583	25
Placebo	2.6000	.9129	25
Total	2.9867	1.2247	75

ANOVA Results

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Eta Squared
Corrected Model	14.427	2	7.213	5.379	.007	.130
Intercept	669.013	1	669.013	498.850	.000	.874
Group	14.427	2	7.213	5.379	.007	.130
Error	96.560	72	1.341			

"Model" row includes all effects in the model, such as all independent variables and interaction effects (see Chapter 10 for a discussion of these multiple effects). In the present example, there is only one independent variable, so the "Model" statistics are the same as the "Group" statistics.

Let's focus on the "Group" row. This row includes all of the between-group information, because "Group" is our independent group variable. Here we see the sum of squares between (SS_b),[2] which is 14.427. The number of degrees of freedom ("*df*") here is 2, because with three groups, $K - 1 = 2$. The sum of squares divided by the degrees of freedom produces the mean square (MS_b), which is 7.213. The statistics for the sum of squares error (SS_e), degrees of freedom for the error component, and the mean square error (MS_e) are all in the row below the "Group" row. The *F* value ("*F*") for this ANOVA is 5.379, which was produced by dividing the mean square from the "Group" row by the mean square from the "Error" row. This *F* value is statistically significant ("Sig." = .007). The "Sig." is the same thing as the *p* value (described in Chapter 7). Finally, in the "Eta Squared" column, we can see that we have a value of .130 in the "Group" row. Eta squared is a measure of the association between the independent variable ("Group") and the dependent variable (the interest variable). It indicates that 13 percent of the variance in the scores on the interest variable can be explained by the group variable. In other words, I can account for 13 percent of the variance in the interest scores simply by knowing whether students were in the "High Dose," "Low Dose," or "Placebo" group. Eta squared is similar to the coefficient of determination (r^2) discussed in Chapters 12 and 13.

Now that we know that there is a statistically significant difference between the three groups in their level of interest, and that group membership accounts for 13 percent of the variance in interest scores, it is time to look at our Tukey post-hoc analysis to determine which groups significantly differ from each other. The SPSS results of this analysis are presented in Table 9.4. The far left column of this table contains the reference group (*I*), and the column to the right of this shows the comparison groups (*J*). So, in the first comparison, the mean for the "High Dose" group is compared to the mean for the "Low Dose" group. We can see that the "Mean Difference" between these two groups is −.8400, indicating that the "High Dose" group had a mean that was .84 points lower than the mean of the "Low Dose" group on the interest variable. In the last column, we can see that this difference is statistically significant ("Sig." = .033). So we can conclude that students in the "Low Dose" group, on average, were more interested in their work than were students in the "High Dose" group. In the next comparison, between "High Dose" and "Placebo," we find a mean difference of .16, which was not significant ("Sig." = .877). Looking at the next set of comparisons, we see that the "Low Dose" group is significantly different from both the "High Dose" group (we already knew this) and the "Placebo" group. At this point, all of our comparisons have been made and we can conclude that, on average, students in the "Low Dose" group were significantly more interested in their work than were students in the "High Dose" and "Placebo" groups, but there was no significant difference between the interest of students in the "High Dose" and "Placebo" groups.

For a video demonstration of how to interpret the SPSS output for a one-way ANOVA, please refer to the website that accompanies this book.

TABLE 9.4 SPSS results of Tukey HSD post-hoc tests comparing three drug-treatment groups

(I) Treatment 1, Treatment 2, Control	(J) Treatment 1, Treatment 2, Control	Mean Difference (I − J)	Std. Error	Sig.
High Dose	Low Dose	−.8400	.328	.033
	Placebo	.1600	.328	.877
Low Dose	High Dose	.8400	.328	.033
	Placebo	1.0000	.328	.009
Placebo	High Dose	−.1600	.328	.877
	Low Dose	−1.0000	.328	.009

TABLE 9.5 Data for 5-, 8-, and 12-year-olds' hours slept per day

5-Year-Olds	8-Year-Olds	12-Year-Olds
12	12	10
11	10	9
11	10	8
10	9	8
9	9	7
$Mean_1 = 10.6$	$Mean_2 = 10.0$	$Mean_3 = 8.4$

Example: Comparing the Sleep of 5-, 8-, and 12-Year-Olds

Suppose that I've got three groups: 5-year-olds, 8-year-olds, and 12-year-olds. I want to see whether children at these different age levels differ in the amount of sleep they get per day, on average. I get the data presented in Table 9.5. From the individual scores presented for each group, all of the additional data can be calculated. Let's walk through these steps.

Step 1: Find the mean for each group.
To find the mean for each group, add the scores together within the group and divide by the number of cases in the group. These group means have been calculated and are presented in Table 9.5.

Step 2: Calculate the grand mean.
This can be done either by adding up all of the 15 scores across the groups and dividing by 15 or, because each group has the same number of cases in this example, by adding up the three group means and dividing by 3: 10.6 + 10.0 + 8.4 = 29 / 3 = 9.67.

Step 3: Calculate the sum of squares error (SS_e).
First, we must find the squared deviation between each individual score and the group mean. These calculations are presented in Table 9.6. When we sum the three sums of squares, we get $SS_e = 16.40$.

Step 4: Calculate the sum of squares between groups (SS_b).
Recall that to find the SS_b we need to subtract the grand mean from the group mean, square it, and multiply by the number of cases in the group. Then we add each of these numbers together. So for our three groups we get:

$$\text{Group 1: } 5(10.6 - 9.67)^2 = 5(.86) = 4.30$$

$$\text{Group 2: } 5(10.0 - 9.67)^2 = 5(.11) = .55$$

$$\text{Group 3: } 5(8.4 - 9.67)^2 = 5(1.61) = 8.05$$

$$\textit{Sum}: 4.30 + .55 + 8.05 = 12.90$$

TABLE 9.6 Squared deviations for the ANOVA example

5-Year-Olds	8-Year-Olds	12-Year-Olds
$(12 - 10.6)^2 = 1.96$	$(12 - 10.0)^2 = 4.0$	$(10 - 8.4)^2 = 2.56$
$(11 - 10.6)^2 = .16$	$(10 - 10.0)^2 = 0$	$(9 - 8.4)^2 = .36$
$(11 - 10.6)^2 = .16$	$(10 - 10.0)^2 = 0$	$(8 - 8.4)^2 = .16$
$(10 - 10.6)^2 = .36$	$(9 - 10.0)^2 = 1.0$	$(8 - 8.4)^2 = .16$
$(9 - 10.6)^2 = 2.56$	$(9 - 10.0)^2 = 1.0$	$(7 - 8.4)^2 = 1.96$
$SS_1 = 5.20$	$SS_2 = 6.00$	$SS_3 = 5.20$

Step 5: Calculate the mean square error (MS_e).
To find the MS_e we divide the SS_e by the degrees of freedom for the error (df_e). The df_e is $N - K$. In this example we have 15 cases across three groups, so the degrees of freedom are 15 – 3 = 12. When we divide the SS_e by 12 we get

$$MS_e = 16.40/12 = 1.37$$

Step 6: Calculate the mean square between groups (MS_b).
To find the MS_b we divide the SS_b by the degrees of freedom between groups (df_b). The df_b is $K - 1$. In this example we have three groups, so the degrees of freedom are 3 – 1 = 2. When we divide the SS_b by 2 we get

$$MS_b = 12.90/2 = 6.45$$

Step 7: Calculate the F ratio.
The F ratio can be found by dividing the MS_b by the MS_e:

$$F = \frac{6.45}{1.37} = 4.71$$

Step 8: Find the critical value for F from Appendix C.
Looking in Appendix C, with 2 degrees of freedom in the numerator and 12 degrees of freedom in the denominator, we find a critical value of F of 3.88 (with $\alpha = .05$).

For a brief video explaining how to read and interpret Appendix C, please refer to the website that accompanies this book.

Step 9: Decide whether the F value is statistically significant.
By comparing our observed F value of 4.71 with the critical F value of 3.88, we can see that $F_o > F_c$. Therefore, we conclude that our results are statistically significant (see Figure 9.5).

Assuming we selected an alpha level of .05, we now know that we have a statistically significant F value. This tells us that there is a statistically significant difference between the means of our three groups in the average amount of sleep they get per day. But I do not yet know which of my three groups differ. To figure this out, I need to conduct post-hoc tests. So, I conduct Tukey

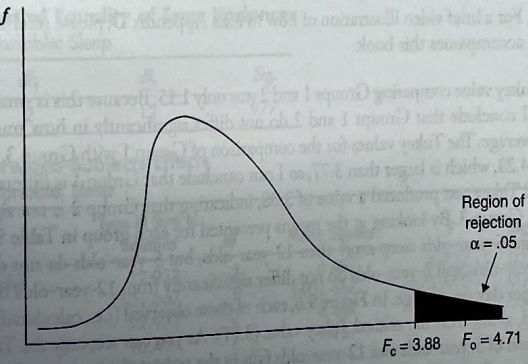

FIGURE 9.5 Critical and observed F values for the ANOVA example.

tests to compare my three groups. Recall that the formula for the Tukey test is the mean of one group minus the mean of another group divided by the standard error. When all of our groups have equal numbers of cases, then the standard error for the Tukey test is the same for each comparison of groups. In our example, we have equal numbers of cases in each group, so we only need to calculate the standard error once:

$$s_{\bar{x}} = \sqrt{\frac{MS_e}{n_g}}$$

$$s_{\bar{x}} = \sqrt{\frac{1.37}{5}}$$

$$s_{\bar{x}} = \sqrt{.27} \Rightarrow .52$$

With our standard error for the Tukey tests in place, we can compare the means for each of the three groups:

$$\text{Tukey}_{1-2} = \frac{10.6 - 10.0}{.52} \Rightarrow \frac{.6}{.52} \Rightarrow 1.15$$

$$\text{Tukey}_{1-3} = \frac{10.6 - 8.4}{.52} \Rightarrow \frac{2.2}{.52} \Rightarrow 4.23$$

$$\text{Tukey}_{2-3} = \frac{10.0 - 8.4}{.52} \Rightarrow \frac{1.6}{.52} \Rightarrow 3.08$$

The final step in our analysis is to determine whether each of these Tukey HSD values is statistically significant. To do this, we must look at the table of critical values for the **studentized range statistic** in Appendix D. The values in this table are organized in a similar way to those presented in the table of *F* values in Appendix C. However, instead of using the degrees of freedom between groups to find the appropriate column, we use the number of groups. In this example, we have three groups, so we find the column labeled "3." To find the appropriate row, we use the degrees of freedom for the error. In this example our df_e was 12. So, with an alpha level of .05, our Tukey value must be larger than 3.77 before we consider it statistically significant. I know this because the critical Tukey value in Appendix D for 3 groups and 12 degrees of freedom is 3.77.

For a brief video illustration of how to read Appendix D, please refer to the website that accompanies this book.

My Tukey value comparing Groups 1 and 2 was only 1.15. Because this is smaller than the value of 3.77, I conclude that Groups 1 and 2 do not differ significantly in how much they sleep per day, on average. The Tukey values for the comparison of Group 1 with Group 3 produced a Tukey value of 4.23, which is larger than 3.77, so I can conclude that Group 1 is different from Group 3. My third Tukey test produced a value of 3.08, indicating that Group 2 is not significantly different from Group 3. By looking at the means presented for each group in Table 9.5, I can see that, on average, 5-year-olds sleep more than 12-year-olds, but 5-year-olds do not differ significantly from 8-year-olds and 8-year-olds do not differ significantly from 12-year-olds in how much sleep they get per day, on average. In Figure 9.6, each of these observed (i.e., calculated) Tukey values are presented, along with the critical Tukey value (3.77). As you can see, only the Tukey value for the comparison of 5-year-olds and 12-year-olds falls in the region of rejection, indicating that it is the only statistically significant difference between the three groups being compared.

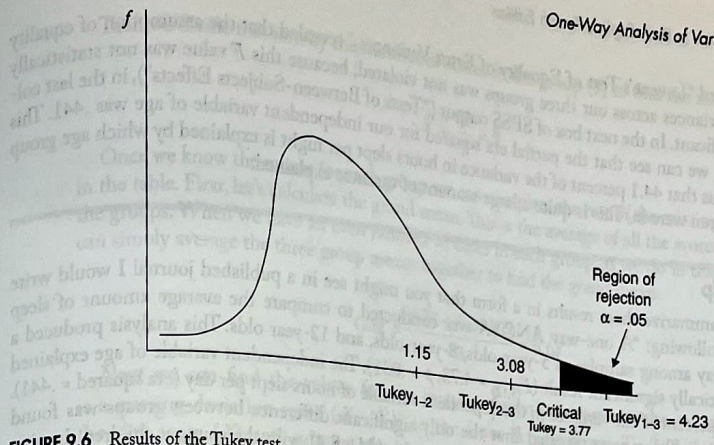

FIGURE 9.6 Results of the Tukey test.

For the sake of comparison, I analyzed the same data using the SPSS computer program. The results of this analysis are presented in Table 9.7. You will notice some minor differences in the values produced by SPSS and those that I calculated above. For example, we calculated SS_b = 12.90, but SPSS reported an SS_b = 12.933 (under the "Type III Sum of Squares" for the age variable). This difference is due simply to rounding error. This rounding error also produced a small difference between the MS_b values and the observed *F* values. Nonetheless, both my analysis and the SPSS analysis found a statistically significant *F* value. The SPSS output, in the box

TABLE 9.7 SPSS output for one-way ANOVA comparing different age groups' average hours of sleep per day

Descriptive Statistics
Dependent Variable: Sleep

Age	Mean	Std. Deviation	N
5.00	10.6000	1.14018	5
8.00	10.0000	1.22474	5
12.00	8.4000	1.14018	5
Total	9.6667	1.44749	15

Levene's Test of Equality of Error Variances
Dependent Variable: Sleep

F	df_1	df_2	Sig.
.023	2	12	.977

Tests of Between-Subjects Effects
Dependent Variable: Sleep

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	12.933[1]	2	6.467	4.732	.031	.441
Intercept	1401.667	1	1401.667	1025.610	.000	.988
Age	12.933	2	6.467	4.732	.031	.441
Error	16.400	12	1.367			
Total	1431.000	15				
Corrected Total	29.333	14				

labeled "Levene's Test of Equality of Error Variances," revealed that the assumption of equality of variances across our three groups was not violated, because this F value was not statistically significant. In the next box of SPSS output ("Tests of Between-Subjects Effects"), in the last column, we can see that the partial eta squared for our independent variable of age was .441. This tells us that 44.1 percent of the variance in hours slept per night is explained by which age group children were in. This is quite a large amount of variance explained.

Writing it Up

To summarize these results in a form that you might see in a published journal I would write the following: "A one-way ANOVA was conducted to compare the average amount of sleep per day among samples of 5-year-olds, 8-year-olds, and 12-year olds. This analysis produced a statistically significant result ($F_{(5,12)} = 4.73, p = .031$). The independent variable of age explained 44.1 percent of the variance in the dependent variable of hours slept per day (eta squared = .441). Post-hoc Tukey tests revealed that the only significant difference between groups was found between 5-year-olds ($M = 10.6$) and 12-year-olds ($M = 8.4$), with the younger children sleeping significantly more than the older children."

Worked Example

Suppose I work for a drug company and we have developed a new weight-loss drug. The new drug, Lose-A-Lot, is given to a random sample of 7 adult men. The old weight-loss drug, Meltaway, is given to a different random sample of 7 adult men. A third random sample of 7 adult men take a placebo (i.e., a water-filled capsule that they believe is a weight-loss drug). After one month of taking the pills, the men are weighed to see how much weight each man has lost (or gained) during the month. The data are summarized in Table 9.8.

As you can see, the table includes some blank boxes that will allow me to demonstrate how to do each of the necessary calculations in a one-way ANOVA. The first step in any ANOVA calculation is to find the mean for each group. In this example, the means for the Lose-A-Lot and Meltaway groups are provided in the table, but the mean for the Placebo group needs to be calculated. The raw scores for the Placebo group are provided in the table. To find the mean, or average, of these values we first need to add them together to find their sum:

$$5 + 4 + 7 + 9 + 11 + 15 + 12 = 63$$

TABLE 9.8 Comparison of three weight-loss groups for a one-way ANOVA example

	Lose-A-Lot		Meltaway		Placebo	
	X	$(X - \bar{X})^2$	X	$(X - \bar{X})^2$	X	$(X - \bar{X})^2$
Man 1	2	7.34	6	0.08	5	16
Man 2	5	0.08	4	[ ]	4	25
Man 3	4	0.50	8	2.92	7	4
Man 4	6	1.66	7	0.50	9	0
Man 5	8	10.82	9	7.34	11	4
Man 6	5	[ ]	6	0.08	15	36
Man 7	3	2.92	4	5.24	12	9
Group means	$\bar{x} = 4.71$		$\bar{x} = 6.29$		[ ]	
SS (each group)	[ ]	23.40		[ ]		94
Grand mean						

Then we divide this sum by the number of scores in the group (7) to find the mean:

$$\bar{X} = \frac{63}{7} = 9$$

Once we know the means for each group, we can complete the rest of the missing information in the table. First, let's calculate the grand mean. This is the average of all the scores, across all of the groups. When we have an even number of cases in each group, as we do in this example, we can simply average the three group means together to find the grand mean:

$$GM = \frac{4.71 + 6.29 + 9}{3} = 6.67$$

Next we can find the values for the missing deviation scores in the table. First, there is a missing deviation score $(X - \bar{X})^2$ in the Lose-A-Lot group. To find this score I take the raw score (X) and subtract the group mean from it: $(5 - 4.71)^2 = (.29)^2 = .08$. Next, we do the same process to find the missing deviation score for the Meltaway group: $(4 - 6.29)^2 = (-2.29)^2 = 5.24$.

Now that we have all of the deviation scores for each group, we can calculate the sum of the squared deviations for each group. For the Lose-A-Lot group, the sum of the squared deviations is 23.4. For the Placebo group, the sum of the squared deviation is 94. Both of these sums of squared deviations are provided in the table. For the Meltaway group, we need to add together the squared deviations to find the sum for that group:

$$.08 + 5.24 + 2.92 + .50 + 7.34 + .08 + 5.24 = 21.40$$

We have now found all of the missing values from Table 9.8, and the completed table can be found in Table 9.9. With all of the pieces of the table in place, we can move on to calculating the necessary statistics for our F value. First, with the sums of squared deviations for each group known, we can calculate the sum of squares within (SS_w), also known as the sum of squares error (SS_e). To find this value we simply add together the sums of squared deviations from each of the three groups in our analysis:

$$SS_e = 23.40 + 21.40 + 94$$

$$SS_e = 138.80$$

TABLE 9.9 Completed table for weight-loss comparison example with missing values added

	Lose-A-Lot		Meltaway		Placebo	
	X	$(X - \bar{X})^2$	X	$(X - \bar{X})^2$	X	$(X - \bar{X})^2$
Man 1	2	7.34	6	0.08	5	16
Man 2	5	0.08	4	5.24	4	25
Man 3	4	0.50	8	[2.92]	7	4
Man 4	6	1.66	7	0.50	9	0
Man 5	8	10.82	9	7.34	11	4
Man 6	5	.08	6	0.08	15	36
Man 7	3	[2.92]	4	5.24	12	9
Group mean	$\bar{x} = 4.71$		$\bar{x} = 6.29$		[9]	
SS (each group)		23.40		[21.40]		94
Grand mean	[6.67]					

Next, we use the group means and the grand mean to calculate the sum of squares between (SS_b) groups. Remember from earlier in the chapter that we need to multiply each deviation between the group mean and the grand mean by the number of cases in the group:

Lose-A-Lot: $7(4.71 - 6.67)^2 = 7(-1.96)^2 = 7(3.84) = 26.89$

Meltaway: $7(6.29 - 6.67)^2 = 7(-.38)^2 = 7(.14) = .98$

Placebo: $7(9 - 6.67)^2 = 7(2.33)^2 = 7(5.43) = 38.01$

To find the SS_b we add these values together:

$$SS_b = 26.89 + .98 + 38.01 = 65.88$$

The next step is to convert our SS_e and SS_b into the values that make up the F value: the MS_e and the MS_b. We accomplish this by dividing the two SS values by their respective degrees of freedom, or df. For the SS_e, the degrees of freedom are the total number of cases, across all of the groups, minus the number of groups. In this example there are 21 cases across 3 groups, so $df_e = 21 - 3 = 18$. The degrees of freedom for the SS_b is the number of groups minus 1: $df_b = 3 - 1 = 2$.

To find the MS_e we divide the SS_e by the df_e: $MS_e = \frac{138.8}{18} = 7.71$

To find the MS_b we divide the SS_b by the df_b: $MS_b = \frac{65.88}{2} = 32.94$

Now we are ready to calculate the observed F value by dividing the MS_b by the MS_e:

$$F = \frac{32.94}{7.71} = 4.27$$

So now we know that our observed F value for this problem is 4.27. But is that statistically significant? To make that determination we need to find the *critical* F value and compare it to our observed F value. Remember that the two degrees of freedom values that we found are 2 and 18. Using Appendix C, the number of degrees of freedom for the numerator of the F value is 2, and the number of degrees of freedom for the denominator is 18. Using an alpha level of .05, we find a critical F value of 3.55. When we compare this to our observed F value of 4.27, we can see that $4.27 > 3.55$, so our observed F value > our critical F value. Therefore, we conclude that our results are statistically significant. The three populations represented by the samples in the Lose-A-Lot, Meltaway, and Placebo groups differ in their average weight lost, but we do not yet know which groups differ from which other groups. For that, we need to perform Tukey post-hoc tests.

The first step in performing a Tukey test is to calculate the standard error that we will use as the denominator in our Tukey comparisons. To do this, we need to know the MS_e, which you may recall from our F value is 7.71, and the number of cases in each group, which is 7. Then we plug those numbers into our standard error formula for the Tukey test:

$$s_{\bar{x}} = \sqrt{\frac{MS_e}{n_g}}$$

$$s_{\bar{x}} = \sqrt{\frac{7.71}{7}}$$

$$s_{\bar{x}} = \sqrt{1.10} = 1.05$$

Now that we have our standard error, we can use the Tukey HSD formula to compare the means of each group to each other group:

$$\text{Tukey HSD} = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}}}$$

Lose-A-Lot–Meltaway: $\text{Tukey HSD} = \frac{4.71 - 6.29}{1.05} = -1.50$

Lose-A-Lot–Placebo: $\text{Tukey HSD} = \frac{4.71 - 9.0}{1.05} = -4.09$

Placebo–Meltaway: $\text{Tukey HSD} = \frac{9.0 - 6.29}{1.05} = 2.58$

We now have three, calculated, observed Tukey HSD values. We can compare each one of them to the critical Tukey value that we will look up in Appendix D, using 3 for the number of levels of the independent variable (because we are comparing three different groups) and 18 for our df error. This gives us a critical Tukey value of 3.61 if we are using an alpha level of .05. Comparing our observed Tukey values to the critical Tukey value, we can conclude that only the difference between the Lose-A-Lot group and the Placebo group is statistically significant. Therefore, we would conclude that, in the populations of these three groups, the average amount of weight loss does not differ for the Lose-A-Lot and the Meltaway groups, or for the Meltaway and the Placebo groups, but the Lose-A-Lot group lost significantly more weight than did those in the Placebo group, on average.

For a video demonstration of how to calculate and interpret a one-way ANOVA and post-hoc tests, please refer to the website that accompanies this book.

Wrapping Up and Looking Forward

One-way ANOVA, when combined with post-hoc tests, is a powerful technique for discovering whether group means differ on some dependent variable. The F value from a one-way ANOVA tells us whether, overall, there are significant differences between our group means. But we cannot stop with the F value. To get the maximum information from a one-way ANOVA, we must conduct the *post-hoc* tests to determine *which* groups differ. ANOVA incorporates several of the concepts that I have discussed in previous chapters. The sum of squares used in ANOVA is based on the squared deviations first introduced in Chapter 3 in the discussion of variance. The comparisons of group means is similar to the information about independent samples t tests presented in Chapter 8. And the eta squared statistic, which is a measure of association between the independent and dependent variables, is related to the concepts of shared variance and variance explained discussed in Chapters 12 and 13, as well as the notion of effect size discussed in Chapter 7.

In this chapter, a brief introduction to the most basic ANOVA model and post-hoc tests was provided. It is important to remember that many models are not this simple. In the real world of social science research, it is often difficult to find groups with equal numbers of cases. When groups have different numbers of cases, the ANOVA model becomes a bit more complicated. I encourage you to read more about one-way ANOVA models, and I offer some references to help you learn more. In the next two chapters, I examine two more advanced types of ANOVA techniques: factorial ANOVA and repeated-measures ANOVA.

In this chapter and those that preceded it, I examined several of the most basic, and most commonly used, statistics in the social sciences. These statistics form the building blocks for most of the more advanced techniques used by researchers. For example, t tests and one-way

ANOVA represent the basic techniques for examining the relations between nominal or categorical independent variables and continuous dependent variables. More advanced methods of examining such relations, such as factorial ANOVA and repeated-measures ANOVA, are merely elaborations of the more basic methods I have already discussed. Techniques for examining the associations among two or more continuous variables are all based on the statistical technique discussed in Chapter 12, correlations. More advanced techniques, such as factor analysis and regression, are based on correlations.

Work Problems

1. What type of research question would you use ANOVA to answer?
2. Why would we use ANOVA rather than multiple *t* tests?
3. Describe what an *F* ratio is. What does it tell you?
4. How should we interpret a significant *F* ratio? That is, what does a statistically significant *F* ratio tell us? What does it not tell us?
5. Suppose I want to know whether drivers in Ohio, Texas, and California differ in the average number of miles they commute to work each day. So I select random samples of 5 drivers from each state and ask them how far they drive to work. I get the data that is summarized in Table 9.10. Please answer the following questions based on these data.
   a. Calculate the missing values in the blank cells in the table.
   b. Perform all of the necessary calculations to determine whether the *F* value is statistically significant with an alpha level of .05. DO NOT perform the Tukey tests here.
   c. Interpret your results. What can you say now about the differences between the population means?
   d. Conduct the Tukey post-hoc tests to determine which means differ significantly using an alpha level of .05, then interpret your results. Now what can you say about the differences between the population means?

TABLE 9.10 Data for numbers of hours Ohio, Texas, and California drivers spend commuting per day

	Ohio		Texas		California	
	X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$
Driver 1	11	.36	11	23.04	13	49
Driver 2	5	29.16	13	7.84	16	16
Driver 3	12	[ ]	15	.64	19	1
Driver 4	16	31.36	19	[ ]	24	16
Driver 5	8	5.76	21	27.04	28	64
Group means	10.40		[ ]		20.00	
SS (each group)	[ ]	69.20		[ ]		146
Grand mean						

6. I recently conducted an experiment with college students to examine whether telling them about the expectations for how well they would perform on a test influenced how well they actually performed. Students were randomly assigned to one of three groups: Neutral, in which students were simply told they would be given a test and to try to do well; Encouraged, in which students were told they were expected to do well because they were good students attending a good school; and Disrespected, in which they were told the test was created for students at a more prestigious school so the test creator did not expect them to do very well. Then the students in my study completed a set of test questions. I conducted a one-way ANOVA to compare the means of the participants in the three groups, and the SPSS results of this analysis are presented in Table 9.11. In this analysis, test performance is the total number of test questions that were correctly answered by the participants. The independent variable is labeled "Group" in the SPSS output. Take a look at the results presented in Table 9.11 and answer the following questions.
   a. Report the means for each group.
   b. Report whether the assumption of equal variances is violated and explain how you know.
   c. Report the *F* value for the one-way ANOVA examining differences between the groups.
      i. Is this a statistically significant *F* value? How do you know?
   d. Report the effect size for the independent group variable and explain what it means.
   e. Write a sentence or two to summarize what this ANOVA reveals about the differences in test performance between the groups.

TABLE 9.11 SPSS output for one-way ANOVA comparing test performances of three groups (Neutral, Encouraged, and Disrespected)

Descriptive Statistics
Dependent Variable: Test performance

Group	Mean	Std. Deviation	N
1 = Neutral	10.9423	2.60778	52
2 = Encouraged	10.5577	2.93333	52
3 = Disrespected	9.8519	2.92940	54
Total	10.4430	2.84749	158

Levene's Test of Equality of Error Variances
Dependent Variable: Test performance

F	df_1	df_2	Sig.
.887	2	155	.414

Tests of Between-Subjects Effects
Dependent Variable: Test performance

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	32.519[1]	2	16.259	2.032	.135	.026
Intercept	17250.573	1	17250.573	2155.507	.000	.933
Group	32.519	2	16.259	2.032	.135	.026
Error	1240.469	155	8.003			
Total	18504.000	158				
Corrected Total	1272.987	157				

Note: 1 R Squared = .026 (Adjusted R Squared = .013)

You can find the answers to these work problems, as well as additional work problems, on the website that accompanies this book.

Notes

1 There are actually two different ways to conduct a one-way ANOVA in SPSS. In this chapter I presented the results from the "General Linear Model → Univariate" option. This method allowed me to present the eta squared statistic. The other approach in SPSS is to choose the "Compare Means → One-way ANOVA" option. This method does not produce an eta squared statistic, but it does produce a Levene's test of homogeneity of variance to examine whether this assumption of ANOVA has been violated.

2 SPSS generally reports this as the Type III sum of squares. This sum of squares is known as the "residual" sum of squares because it is calculated after taking the effects of other independent variables, covariates, and interaction effects into account.