This Chapter’s Terra Incognita

You are reading a research article. It describes an experiment in which subjects in the experimental group consumed a large amount of caffeine, whereas those in the control group consumed none. The dependent variables were four different measures of subject irritability.

The Results section reports the following:

Because the study included four dependent variables, the Bonferroni adjustment was used to maintain a familywise Type I error rate of .05. The resulting Bonferroni-adjusted alpha level was $\alpha_{ADJ} = .0125$.

For self-rated irritability, the mean score for the high-caffeine condition ($M = 575.00, SD = 35.49$) was higher than the mean for the no-caffeine condition ($M = 544.73, SD = 36.13$). The observed difference between means was 30.27, 95% CI [18.71, 41.82], $t(148) = 5.18, p < .001$. The point-biserial correlation between the independent variable and the dependent variable was $r_{pb} = .39$, and this fell just short of a large effect according to Cohen’s (1988) criteria.

Have courage. All shall be revealed.
**Introduction to the Big-Three Results**

Research articles provide three types of statistical results that are used to make sense of the study’s findings: significance tests, confidence intervals, and indices of effect size. This chapter explains the meaning of these results and shows how they are usually presented in the tables and text of an article.

The three results are discussed separately because each provides a somewhat different perspective regarding the investigation’s findings. Specifically:

- The significance test communicates *probability*: It tells us how likely the current results would be if the study’s null hypothesis were true.
- The confidence interval communicates *precision*: It provides a range of plausible values for the population parameter being estimated. Confidence intervals help us evaluate whether this range is relatively precise or relatively imprecise.
- The index of effect size communicates *strength*: It tells us about the strength of the relationship between the predictor variable and the criterion variable.

**Illustrative Study: The Effects of Caffeine**

Assume that a fictitious researcher named Dr. O’Day wants to determine whether consuming caffeine causes people to be more irritable (i.e., to be more easily upset or angered). To determine this, she conducts an experiment in which she manipulates the amount of caffeine consumed and observes how this manipulation affects scores on four measures of irritability.

**Manipulating the Independent (Predictor) Variable.** Dr. O’Day began with a pool of 150 potential participants. She randomly assigned half \( (n = 75) \) to the experimental group, and the other half \( (n = 75) \) to the control group. Regardless of condition, all participants were instructed to drink five cups of coffee each day. The coffee was provided by Dr. O’Day, and she saw to it that:

- Participants in the experimental group consumed regular coffee that contained a total of 1,000 mg of caffeine per day (we will call this the *high-caffeine* condition).
- Participants in the control group consumed decaffeinated coffee that contained a total of 0 mg of caffeine per day (this is the *no-caffeine* condition).

**Measuring the Dependent (Criterion) Variables.** Dr. O’Day collected scores on the following four dependent variables:

- **Self-rated irritability.** This variable indicated the extent to which the subject reported feeling irritable, upset, and out-of-sorts. Each day, the subject completed a brief multiple-item summated rating scale. Scores on this scale could range from 200 to 800 with higher scores indicating greater
feelings of irritability. For a given subject, Dr. O’Day computed the mean of that subject’s scores over the course of the study, and that mean served as the self-reported irritability score for that participant.

- **OTHER-RATED IRRITABILITY.** This variable indicated the extent to which the subject felt irritable, upset, and out-of-sorts, as rated by someone who saw the subject on a daily basis (such as a friend or relative). It was measured and scored in the same way that the self-rated irritability scale was scored.

- **SELF-REPORTED OUTBURSTS.** This variable indicated the number of anger-related outbursts (e.g., temper tantrums) that the participant displayed over the four-week course of the study. Scores on this variable were reported by the participant.

- **OTHER-REPORTED OUTBURSTS.** This variable was similar to the previous one, except that the reports were again provided by friend or relative who saw the participant on a regular basis.

---

**Table 5.1**

Scores on Four Measures of Irritability as a Function of Amount of Caffeine Consumed (High-Caffeine Versus No-Caffeine): Significance Tests, Confidence Intervals, and Point-Biserial Correlations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M1 (SD1)</th>
<th>M2 (SD2)</th>
<th>M1 − M2 [95% CI]</th>
<th>t(148)</th>
<th>p</th>
<th>rpb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-rated irritability</td>
<td>575.00 (35.49)</td>
<td>544.73 (36.13)</td>
<td>30.27 [18.71, 41.82]</td>
<td>5.18</td>
<td>&lt;.001</td>
<td>.39</td>
</tr>
<tr>
<td>Other-rated irritability</td>
<td>556.46 (35.86)</td>
<td>546.27 (36.10)</td>
<td>10.19 [-1.42, 21.80]</td>
<td>1.73</td>
<td>.085</td>
<td>.14</td>
</tr>
<tr>
<td>Self-reported outbursts</td>
<td>6.24 (2.67)</td>
<td>5.03 (2.69)</td>
<td>1.21 [0.35, 2.08]</td>
<td>2.77</td>
<td>.006</td>
<td>.22</td>
</tr>
<tr>
<td>Other-reported outbursts</td>
<td>5.24 (2.67)</td>
<td>4.99 (2.65)</td>
<td>0.25 [-0.61, 1.11]</td>
<td>0.58</td>
<td>.561</td>
<td>.05</td>
</tr>
</tbody>
</table>

Note. N = 150. CI = Confidence interval; rpb = Point-biserial correlation between the independent variable and dependent variable.

Results from the Caffeine Study

Dr. O’Day used an independent-samples *t* test to analyze data from the experiment. She performed the analysis four times: once for each criterion variable. She published her findings in a research journal, and part of the *Results* section from the article is reproduced below.
For each dependent variable, an independent-samples \( t \) test was performed to determine whether the mean score for the high-caffeine condition was significantly different from the mean score for the no-caffeine condition. The results from these analyses are summarized in Table 5.1.

The first row of Table 5.1 presents results obtained for the \( t \) test performed on self-rated irritability. The table shows that the mean score for the high-caffeine condition (\( M = 575.00, SD = 35.49 \)) was higher than the mean for the no-caffeine condition (\( M = 544.73, SD = 36.13 \)). The observed difference between the two means was 30.27, 95% CI [18.71, 41.82], and this difference was statistically significant, \( t(148) = 5.18, p < .001 \). As a measure of effect size, the point-biserial correlation was computed between the independent variable (amount of caffeine) and the dependent variable (self-rated irritability). For this dependent variable, \( r_{pb} = .39 \), and this fell just short of a large effect according to Cohen’s (1988) criteria.

The bar chart in Figure 5.1 presents mean scores on self-rated irritability for the two treatment conditions. Error bars represent 95% confidence intervals.

![Bar chart showing mean self-rated irritability scores](image)

*Figure 5.1. Mean self-rated irritability scores as a function of the amount of caffeine consumed (error bars represent 95% CI).*

The remainder of this chapter will explore the results presented in this excerpt in some detail. We will see that the significance tests reported in the excerpt provide one perspective on the findings, the confidence intervals provide a different perspective, and the indices of effect size provide still another perspective.

This exploration begins where most researchers typically begin: with the null-hypothesis significance test.
Big-Three Results, Part 1: Statistical Significance

When you think statistically significant you should think probability or likelihood. When researchers say that results are statistically significant, they are saying the following:

If the null hypothesis were true, it is unlikely that I would have obtained sample results such as these.

In general, a null-hypothesis significance test (abbreviation: NHST) is a procedure for determining whether the current results are statistically significant. It is a procedure for determining the probability that you would have obtained a sample statistic as large as the statistic you obtained (or an even larger statistic) if the null hypothesis were true.

Statistical Significance: Basic Concepts

To understand the meaning of statistical significance, you must first understand the meaning of the various questions and hypotheses that are investigated in a study. First, the research question is a concise summary of what the investigator hopes to learn from the study. For example, Dr. O’Day may have stated her initial research question as:

Does caffeine cause people to be more irritable?

Next, the research hypothesis is an educated guess regarding the likely answer to the research question. It is a description of the relationships that are likely to exist between the variables of interest, typically stated in either the present tense or the future tense. Dr. O’Day might state her initial research hypothesis like this:

Caffeine causes people to be more irritable.

The preceding paragraph described a research hypothesis as being an educated guess about the likely answer to the research question. To be a truly educated guess, the researcher should do her homework before stating it—she should thoroughly review theory and previous research that is relevant to the research question, and the research hypothesis that she states should be informed by this theory and research.

Statistical null hypothesis. Eventually, the researcher will conduct an empirical investigation and will analyze data to determine whether the results are consistent with her research hypothesis. The statistical procedures that determine whether her results are “statistically significant” are actually tests of a specific type of hypothesis called the statistical null hypothesis. The statistical null hypothesis (symbol: $H_0$) may be defined in this way:

- It is a statement about some characteristic of the population(s) being studied.
- It is usually a statement of “no effect,” “no difference,” or “no relationship.”
• It usually makes a prediction that is the opposite of the prediction made by the research hypothesis.

Here is the most important thing to remember about the statistical null hypothesis:

• The statistical null hypothesis is the hypothesis that the researcher typically hopes to reject.

The exact way that a researcher states a statistical null hypothesis is determined by the nature of the statistical test being performed, what outcome is predicted by the research hypothesis, and other factors. For the current caffeine study, Dr. O’Day might have stated the statistical null hypothesis in this manner:

\[ H_0: \mu_E - \mu_C = 0. \]

In the population, the difference between the mean irritability score for the experimental condition versus the mean irritability score for the control condition is equal to zero.

Notice that the preceding used the Greek letter \( \mu \) as a symbol for “mean score in the population.” More specifically, it used the symbol \( \mu_E \) to represent the mean irritability score for the experimental condition (i.e., the high-caffeine condition) in the population, and it used the symbol \( \mu_C \) to represent the mean irritability score for the control condition (i.e., the no-caffeine condition) in the population. Symbolically, the null hypothesis indicated that the difference between these two population means is equal to zero: \( \mu_E - \mu_C = 0 \).

Some researchers state exactly the same null hypothesis in a different way: by stating that the two population means are equal. Here is this equivalent version:

\[ H_0: \mu_E = \mu_C. \]

In the population, the mean irritability score for the experimental condition is equal to the mean irritability score for the control condition.

In almost all cases, the statistical null hypothesis is the hypothesis that the researcher hopes to reject at the conclusion of the statistical analysis. Think of it as a “straw man” that the researcher hopes to knock down. In the present case, the null hypothesis says that (with respect to irritability scores in the population), the mean score for the experimental condition is equal to the mean score for the control condition. Dr. O’Day very much hopes that this statement is not true—she hopes that the results that she obtains in her investigation will allow her to reject this idea.

**STATISTICAL ALTERNATIVE HYPOTHESIS.** The counterpart to the statistical null hypothesis is the statistical alternative hypothesis. The **statistical alternative hypothesis** (symbol: \( H_1 \) or \( H_A \)) may be defined in this way:

• It is a statement about some characteristic of the population(s) being studied.

• It is usually a statement that there is an effect, that there is a difference, or that there is a relationship.
It usually makes a prediction that is consistent with the prediction made by
the research hypothesis.

For the current study, Dr. O’Day might state the statistical alternative hypothesis
in this manner:

\[ H_1 : \mu_E - \mu_C \neq 0 \]

In the population, the difference between the mean irritability
score for the experimental condition versus the mean irritability score for the
control condition is not equal to zero.

Or, she might state an equivalent alternative hypothesis in this way:

\[ H_1 : \mu_E \neq \mu_C \]

In the population, the mean irritability score for the experimental
condition is not equal to the mean irritability score for the control condition.

In most cases, the statistical alternative hypothesis is the hypothesis that
researchers hope will be supported by their study’s results. In this case, Dr. O’Day
hopes to obtain evidence that the mean irritability score for the high-caffeine
condition is not equal to the mean irritability score for the no-caffeine condition.

The Sampling Distribution of the Statistic

Dr. O’Day will determine whether the results from her study are statistically
significant by:

- analyzing the sample data in order to compute an obtained \( t \) statistic
  (symbol: \( t_{\text{obt}} \)), and then
- determining where that \( t_{\text{obt}} \) statistic is located within a null-hypothesis
  sampling distribution.

This means that understanding the concept of a sampling distribution is essential
if we are to understand what is meant by the words statistically significant. The
sampling distribution for a given statistic (such as the \( t \) statistic) is:

- the distribution of all possible values of that statistic that would be
  obtained...
- if an infinite number of samples of the same size were drawn...
- from the populations described by the null hypothesis.

**EXAMPLE.** To make this concept a bit more concrete, imagine how we might
theoretically create the sampling distribution for Dr. O’Day’s study:

- First (and most important), assume that the null hypothesis is true (i.e.,
  assume that caffeine has no effect on irritability in the population).

- Imagine that Dr. O’Day conducts the caffeine study described above.
  Specifically, she draws a sample of 150 participants, randomly assigns each
  subject to one of the two treatment conditions, and manipulates the
  independent variable. She computes the mean irritability score displayed
  by the 75 subjects in the high-caffeine condition (symbol: \( X_E \)). She also
computes the mean irritability score displayed by the 75 subjects in the no-caffeine condition (symbol: $\bar{X}_C$). She subtracts $\bar{X}_C$ from $\bar{X}_E$, producing a *mean-difference* score. This mean difference is a single number which represents the size of the difference between two sample means. It is computed as follows:

$$Mean\ difference = \bar{X}_E - \bar{X}_C$$

- She adds this mean difference to the sampling distribution (at this point, the distribution contains only one mean-difference score).
- She repeats this entire process an infinite number of times. In other words, she conducts the experiment an infinite number of times, each time computing a new difference score and adding it to the sampling distribution).

The resulting distribution is called a *sampling distribution*. Figure 5.2 provides an illustration of the shape of this distribution. This sampling distribution contains the differences between means that we would expect to see if (a) caffeine has no effect on irritability, and (b) the experiment (with total $N = 150$) were conducted an infinite number of times.

![Sampling distribution of differences between means for the caffeine study.](image)

The sampling distribution in Figure 5.2 looks very much like the standard normal distribution, or $z$ distribution, which was covered in *Chapter 4*. Technically, it is actually a distinct distribution called the $t$ distribution. With large samples, (such as the combined sample of $N = 150$ in the current investigation), however, the shape of the $t$ distribution is almost identical to the shape of the $z$ distribution, so the things that you learned about the $z$ distribution in *Chapter 4* will also apply here.
The Critical Value and the Region of Rejection

Notice that the sampling distribution in the figure has been divided into two regions: (a) the middle 95% (labeled the “Region of Nonsignificance”) and (b) most extreme 5% of the distribution, out in the two tails (labeled the “Region of Rejection”). The two vertical dashed lines that divide the distribution into these regions are labeled with the values “−1.96” and “+1.96.” These values are called the critical value of the statistic.

**The critical value of the statistic.** First, the critical value of the $t$ statistic ($t_{crit}$) is the value that separates the region of rejection from the region of nonsignificance. When Dr. O’Day analyzes the data from her study, she will compute an obtained $t$ statistic ($t_{obt}$), and will compare the size of this obtained statistic against the critical value of the statistic: $t_{crit}$, which in this case is equal to ±1.96. If $t_{obt}$ is larger than ±1.96 in absolute value (such as $t_{obt} = 5.18$), this would indicate that the obtained statistic is in the region of rejection. In this case, Dr. O’Day will:

- Reject the null hypothesis.
- Tell the world that there is a statistically significant difference between the means of the two conditions.

However, Dr. O’Day may not be so fortunate. If the value of $t_{obt}$ is smaller than ±1.96 in absolute value (such as $t_{obt} = 0.20$), this indicates that the obtained statistic is in the region of nonsignificance. In this case, Dr. O’Day will:

- Fail to reject the null hypothesis.
- Tell the world that there is a statistically nonsignificant difference between the means of the two conditions.

How did Dr. O’Day know that the critical value of the statistic was equal to ±1.96 for this analysis? She consulted a table of critical values of the $t$ statistic. Such tables appear in the back of most statistics textbooks. One caveat: In this analysis, the $t_{crit}$ is equal to ±1.96, but $t_{crit}$ is likely to be a different value in a different investigation, especially if the number of subjects is different.

**The region of nonsignificance.** The middle 95% of the sampling distribution in Figure 5.2 is labeled the “Region of Nonsignificance.” When the null hypothesis is true, there is a 95% probability that the $t_{obt}$ statistic produced by a given investigation will fall in this section of the sampling distribution. Therefore, when researchers obtain a relatively small statistic that fall in the region of nonsignificance, they fail to reject the null hypothesis—they tell the world that their results are statistically nonsignificant.

**The region of rejection.** The most extreme 5% of the sampling distribution in Figure 5.2 is labeled the “Region of Rejection.” When the null hypothesis is true, there is only a 5% probability that the $t_{obt}$ statistic produced by a given investigation will fall in this section of the sampling distribution. Therefore, when the obtained
statistic is large enough to be in the region of rejection, researchers reject the null hypothesis and tell the world that their results are statistically significant.

**Sampling Error and the Standard Error**

In general, *sampling error* is the difference between a population parameter versus a sample statistic attributable to the fact that samples are usually not perfectly representative of the populations from which they were drawn. Think about it this way: The null hypothesis said that there is no difference between the two population means (\(\mu_E\) and \(\mu_C\)). If this null hypothesis were true, and if all of the samples used in creating the sampling distribution had been perfectly representative of the populations from which they were drawn, then all of the differences between the pairs of sample means would have been exactly equal to zero: \(\bar{X}_E - \bar{X}_C = 0\). If this had happened, then there would have been no variability in the sampling distribution—all of its scores would have been exactly equal to zero.

There was a problem, however. There were only 150 people in Dr. O’Day’s study, and, even if they had been randomly sampled from their populations, we would not expect them to be perfectly representative of the populations from which they had been drawn. And many of those samples that were not perfectly representative of their populations would produce mean-difference scores just a bit different from zero—some would be a bit higher, and some a bit lower. That is why we see some variability in the sampling distribution depicted in Figure 5.2.

**THE STANDARD ERROR.** There is good news, however. We can compute a statistic called the *standard error* which tells us how much variability appears in the sampling distribution.

A *standard error* (symbol: \(SE\)) is a special type of standard deviation: It is a standard deviation of a sampling distribution. The specific name given to a standard error is determined by the nature of the investigation. In the present case, Dr. O’Day’s sampling distribution consists of mean-difference scores, so she will call it the *standard error of the differences between means*. In a statistics textbook, this standard error is typically represented using the symbol \(s_{\bar{X}_E - \bar{X}_C}\), although this book will generally use \(SE_{\text{diff}}\) or just \(SE\).

**FORMULA.** The formula for computing a standard error is determined by the nature of the specific investigation. If one is performing an independent-samples \(t\) test (and if certain statistical assumptions are met), the standard error may be computed using the following formula (from Howell, 2002):

\[
SE_{\text{diff}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

With the above formula: \(SE_{\text{diff}} = \) the standard error of the differences between means; \(s_1^2 = \) the estimated population variance of scores on the criterion variable in the first treatment condition; \(s_2^2 = \) the estimated population variance of scores on the criterion variable in the second treatment condition; \(n_1 = \) the number of
participants in the first treatment condition; and \( n_2 \) = the number of participants in the second treatment condition.

Let’s assume that Dr. O’Day computes the standard error of the differences between means for the current investigation as \( SE_{\text{diff}} = 5.848 \). Now that we have this value, we can compute the obtained \( t \) statistic for her investigation.

### Computing the Obtained \( t \) Statistic

An **obtained statistic** is a value which indicates where the results from the current sample are located within the sampling distribution of the statistic. The way that researchers compute an obtained statistic depends on which statistic they are computing (e.g., a \( \chi^2 \) test, an \( F \) test, etc.). Dr. O’Day is performing an independent-samples \( t \) test, and the definitional formula for this \( t \) statistic (in its simplest form) is as follows:

\[
t_{\text{obt}} = \frac{\bar{X}_E - \bar{X}_C}{SE_{\text{diff}}}
\]

With the preceding formula: \( t_{\text{obt}} \) = the obtained \( t \) statistic; \( \bar{X}_E \) = the sample mean for the experimental condition; \( \bar{X}_C \) = the sample mean for the control condition; \( SE_{\text{diff}} \) = the standard error of the differences between means.

Given the nature of the formula above, it is clear that the \( t \) statistic from an independent-samples \( t \) test is a measure of **difference**: It represents the size of the difference between the sample mean for the experimental group (\( \bar{X}_E \)) versus the sample mean for the control group (\( \bar{X}_C \)) as measured in standard errors (\( SE_{\text{diff}} \)).

Table 5.1 (presented earlier) provided mean scores for the dependent variable, *self-rated irritability*. The table indicated that the mean for the high-caffeine condition was \( \bar{X}_E = 575.00 \) and the mean for the no-caffeine condition was \( \bar{X}_C = 544.73 \). The standard error of the difference (not shown in the table) was \( SE_{\text{diff}} = 5.848 \). Inserting these values into the formula produces the following obtained \( t \) statistic:

\[
t_{\text{obt}} = \frac{\bar{X}_E - \bar{X}_C}{SE_{\text{diff}}} = \frac{575.00 - 544.73}{5.848} = \frac{30.27}{5.848} = 5.176 = 5.18
\]

So the obtained \( t \) statistic for the self-reported irritability dependent variable is \( t_{\text{obt}} = 5.18 \). The size of the statistic tells us that the mean score for the high-caffeine condition is 5.18 standard errors higher than the mean score for the no-caffeine condition.

### Making a Decision Regarding the Null Hypothesis

Table 5.1 showed that the mean scores on self-rated irritability were in the direction that Dr. O’Day’s research hypothesis had predicted, with the high-caffeine condition scoring higher than the no-caffeine condition (mean self-rated irritability scores were 575.00 and 544.73, respectively). To determine whether the
difference between the means is large enough to be statistically significant, she must now review the results of the null-hypothesis significance test.

The preceding section showed that Dr. O’Day’s obtained statistic is $t_{\text{obt}} = 5.18$. An earlier section of this chapter had indicated that the critical value of the statistic for this analysis was $t_{\text{crit}} = \pm 1.96$. Since this obtained statistic of $t_{\text{obt}} = 5.18$ is larger than the critical value of $t_{\text{crit}} = \pm 1.96$, Dr. O’Day may:

- Reject the null hypothesis
- Tell the world that there was a statistically significant difference between the two sample means
- Tell the world that the results of her investigation were consistent with her research hypothesis

**What Statistical Significance Does and Does Not Mean**

People often misinterpret just what is meant by statistically significant findings. To sharpen our understanding, this section provides some correct and incorrect interpretations of Dr. O’Day’s current results.

**Correct:**

- We may reject the null hypothesis.
- If the null hypothesis were true, there is only a 5% probability that we would have obtained a $t$ statistic this large or larger (i.e., as large as or larger than $t_{\text{obt}} = 5.18$, in this case).

**Incorrect:**

- There is only a 5% chance that the mean score for the high-caffeine group is equal to the mean score for the no-caffeine group in the population (This is incorrect because the statistical procedures described here allow us only to make probability statements about samples, not about populations).
- There is only a 5% chance that the null hypothesis is true (Come on! This communicates exactly the same idea as the previous incorrect statement!).
- The amount of caffeine consumed had a strong effect on irritability scores (This is incorrect because significance tests do not reveal the strength of the effect that an independent variable has had on a dependent variable; to learn this, we must instead compute an index of effect size).

**Factors Affecting the Power of a Test**

The power of a significance test is the probability that you will reject the null hypothesis and conclude that you have significant results when there really is an “effect” in the population. With a test of group differences (such as the current $t$
test), the power of the test is the probability that you will conclude that you have significant results when there really is a difference between the means of the different conditions in the population.

Power is usually represented as a proportion or as a percentage, and an ideal level of power that many researchers strive for is .80. When power is equal to .80, the researcher has an 80% probability of obtaining significant results (if there really is an effect in the population, that is). The value of .80 is important because some funding agencies will award a research grant to investigators only if the investigators have demonstrated that the power of their significance tests will be .80 or higher.

The power of the test is influenced by a number of factors. Some of the more important factors are discussed here.

**Actual size of the effect in the population.** In empirical investigations, one definition for effect size is the strength of the relationship between a predictor variable and a criterion variable. Other things held constant, the stronger the “effect” in the population, the greater the power of the significance test performed on a sample of data taken from that population.

**Alpha level.** In this context, the alpha level (also called the significance level; symbol: $\alpha$) refers to the size of the region of rejection in a sampling distribution. You will recall that the region of rejection consists of the most extreme part of a sampling distribution, out in the tails. When the analysis is conducted correctly (i.e., when all assumptions are met), alpha is also the theoretical probability of making a Type I error (Type I errors will be discussed later).

The most popular alpha level is $\alpha = .05$ (meaning that the size of the region of rejection is the most extreme 5% of the distribution, out in the tails). Sometimes researchers set alpha at $\alpha = .01$ (meaning that the region of rejection is the most extreme 1% of the distribution), or $\alpha = .001$ (meaning that the region of rejection is the most extreme .1% of the distribution).

In general, researchers increase the power of the test (i.e., the likelihood that they will obtain significant results) if they set alpha at a large value, such as $\alpha = .05$. In general, they decrease the likelihood that they will obtain significant results if they set alpha at a small value, such as $\alpha = .001$.

**One-tailed test versus two-tailed test.** The significance test for the caffeine study discussed earlier was a two-tailed significance test (also called a nondirectional significance test). With a two-tailed test, the researcher makes a general prediction that one treatment condition will score higher than the other, but does not specifically predict which condition will score high and which will score low. With a nondirectional prediction, the region of rejection is divided into two tails. If the researcher has set alpha at $\alpha = .05$ and performs a two-tailed test, this means that the 5% that makes up the region of rejection is divided into two tails, with 2.5% in one tail, and 2.5% in the other.
The counterpart to a two-tailed test is a **one-tailed significance test** (also called a **directional significance test**). With a one-tailed test, the researcher makes a specific prediction as to which condition will score higher than the other. This means that the region of rejection is concentrated in just one tail. For example, if Dr. O’Day had specifically predicted that the high-caffeine condition would score higher than the no-caffeine condition, she could have performed a directional test by concentrating the entire region of rejection in the right-side tale of the sampling distribution in Figure 5.2.

In general, one-tailed tests are more powerful than two-tailed tests, as long as the researcher correctly predicts the direction of the results. One-tail tests tend to be more powerful because the region of rejection is concentrated in just one tail, and the critical value of the statistic is therefore somewhat lower (and is therefore is easier to reach).

**NUMBER OF SUBJECTS AND THE DEGREES OF FREEDOM.** Other things held constant, the greater the number of participants in the study, the greater the power of the test. This means that (assuming that there really is an effect in the population) Dr. O’Day would be more likely to obtain significant results if she conducted her study with a sample of 300 participants, rather than a sample of just 30 participants.

One of the reasons that more subjects results in more power is because more subjects result in a larger value of the **degrees of freedom** for the analysis. In this context, **degrees of freedom** (symbol: $df$) may be defined as the number of observations used in the computation of a specific statistic (such as the $t$ statistic), minus the number of restrictions that have been placed on the freedom of those observations to vary. For many statistical procedures, the degrees of freedom largely reflect the number of participants in the study. For example, the degrees of freedom for an independent-samples $t$ test are computed as:

$$df = N - 2$$

With the preceding formula, $N$ represents the total number of subjects in the study. In Dr. O’Day’s caffeine study, the total sample size was 150, so the $df$ were equal to 148.

**Two Approaches for Determining Statistical Significance**

Researchers typically determine whether their results are statistically significant by either (a) comparing their obtained statistic against a critical value of the statistic found in a table in the back of a statistics textbook, or (b) consulting the $p$ value provided by a statistical application such as SAS or SPSS. This book refers to these approaches as the **old-school approach** and the **new-school approach**, respectively.

**OLD SCHOOL: TABLES OF CRITICAL VALUES.** Before computer applications such as SAS or SPSS were widely available, researchers routinely computed obtained statistics (such as the $t$ statistic) by hand. To determine whether this obtained statistic was large enough to be in the region of rejection, they then looked up the critical value
of the statistic that was appropriate for their data set. To make this possible, statistics textbooks routinely include tables of critical values for the $t$ statistic, the $F$ statistic, the chi-square statistic, and other procedures.

The size of the critical value of a statistic is determined by a variety of factors. A previous section indicated that the size of the critical value for an independent-samples $t$ test is determined by (a) the degrees of freedom for the analysis, (b) the decision to perform either a one-tailed test or a two-tailed test, and (c) the alpha level selected by the researcher.

For example, assume that in Dr. O’Day’s study, the $df = 148$, and she chose to perform a two-tailed test with alpha set at $\alpha = .05$. A table of critical values in the back of a statistics textbook told her that, given these considerations, the critical value of the test ($t_{crit}$) was $\pm 1.96$.

One of the criterion variables in her investigation was self-reported outbursts (i.e., the number of temper tantrums that the subject displayed during the investigation, as reported by the subject). After analyzing the data, Dr. O’Day found that her obtained $t$ statistic for this criterion variable was $t(148) = 2.77$. Since this obtained statistic (2.77) was larger than the critical value ($\pm 1.96$), she rejected the null hypothesis and told the world that her results for this variable were statistically significant.

Another criterion variable in her investigation was other-reported irritability (i.e., how irritable the subject was, according to friends or relatives). The obtained $t$ statistic for this criterion variable computed as $t(148) = 1.73$. Since this obtained statistic (1.73) was smaller than the critical value ($\pm 1.96$), she failed to reject the null hypothesis. She told the world that her results for this variable were statistically nonsignificant.

**NEW SCHOOL: CONSULTING THE p VALUE FROM COMPUTER OUTPUT.** Today most researchers analyze their data using computer applications such as SAS or SPSS. Among the many advantages of this approach, is the fact that it is now seldom necessary to look up critical values of a statistic in the back of a statistics textbook. Instead, today’s computer applications routinely report (a) the obtained statistic (such as the $t_{obt}$ statistic), along with (b) the $p$ value for that statistic. The $p$ value tells the researcher whether the obtained statistic is significant.

A $p$ value (i.e., probability value) is the probability of obtaining a statistic the size of the current obtained statistic (or an even larger statistic) if the null hypothesis were true. In most studies, the null hypothesis is a hypothesis of no-differences between conditions, or no relationship between variables (in the population). This means that, in most cases, the $p$ value provides the probability that the researcher would have obtained the current sample results if the null hypothesis of “no effect in the population” were true. In most cases, statistical applications provide precise $p$ values, carried out to three or more decimal places, such as “$p = .492$,” “$p = .022$,” or “$p = .001$.” If the $p$ value is small enough, the researcher rejects the null hypothesis and tells the world that the results are statistically significant.
Just how small must the $p$ value be to reject the null hypothesis? It has to be smaller than the alpha level that the researcher has selected for the study. You will recall that the alpha level (symbol: $\alpha$) is the size of the region of rejection in the sampling distribution. Although alpha may be set at .10, .01, 001, or just about any other value, the most popular value is $\alpha = .05$. This means that, if an obtained $p$ value is less than .05, most researchers will conclude that the results are statistically significant. In fact, this book will just about always use $\alpha = .05$ as the criterion for making this decision.

![Diagram of normal distribution with rejection regions](image)

**Figure 5.3.** Location of an obtained $t$ statistic of $t_{obt} = 2.77$ within the sampling distribution (these are statistically significant results; the area shaded with stripes represents the area beyond the obtained $t$ statistic of $t_{obt} = 2.77$).

**NEW SCHOOL EXAMPLE 1: SIGNIFICANT RESULTS.** In addition to the definition provided above, here is an additional (and equivalent) way of defining a $p$ value:

The $p$ value for an obtained statistic indicates the proportion of area in the sampling distribution which lies beyond that obtained statistic, out in the tail or tails.

For example, Table 5.1 (presented much earlier in the chapter) indicated that, for the criterion variable *self-reported outbursts*, the obtained $t$ statistic was $t_{obt} = 2.77$ and the $p$ value for this statistic was $p = .006$. This $p$ value communicates the following message:

The obtained $t$ statistic ($t_{obt} = 2.77$) was so far away from the mean of the sampling distribution that only .006 of the sampling distribution was beyond this obtained statistic, out in the tails.

This is illustrated in Figure 5.3.
If the null hypothesis were true, we would expect Dr. O’Day’s obtained $t$ statistic to be $t_{\text{obs}} = 0$. This is represented by the solid vertical line in the center of the sampling distribution. But Dr. O’Day did not obtain a $t$ statistic equal to zero. Instead, her obtained statistic was $t_{\text{obs}} = 2.77$, as represented by the single-headed arrow on the right side of the figure. The shaded (cross-hatched) area represents the proportion of the sampling distribution which lies beyond this obtained statistic, out in the two tails (the same amount of area was shaded on both the left tail and right tail, because Dr. O’Day had performed a two-tailed test).

If the null hypothesis were true, the probability is just .006 that Dr. O’Day would have obtained a $t$ statistic as large (or larger) than 2.77. In the disciplines of research and statistics, this is considered a fairly low probability.

In summary, Dr. O’Day’s obtained $p$ value ($p = .006$) was smaller than the alpha level that she selected for the test ($\alpha = .05$). Therefore, she rejected the null hypothesis and told the world that there was a statistically significant difference between the two conditions with respect to their scores on self-reported outbursts.

![Figure 5.4. Location of an obtained $t$ statistic of $t_{\text{obs}} = 1.73$ within the sampling distribution (these are statistically nonsignificant results; the area shaded with stripes represents the area beyond the obtained $t$ statistic of $t_{\text{obs}} = 1.73$).](image)

**NEW SCHOOL EXAMPLE 2: NONSIGNIFICANT RESULTS.** In Dr. O’Day’s study, a different criterion variable was *other-rated irritability*. This variable indicated how irritable the subject was according to ratings made by the subject’s friends or relatives. Table 5.1 (presented earlier in this chapter) indicated that, for this criterion variable, the obtained $t$ statistic was $t_{\text{obs}} = 1.73$ and the corresponding $p$ value was $p = .085$. These results are graphically illustrated in Figure 5.4.
Once again, if the null hypothesis were true, we would expect Dr. O’Day’s obtained $t$ statistic to be $t_{obt} = 0$. The single-headed arrow shows that for other-rated irritability, the obtained statistic was $t_{obt} = 1.73$. As before, the shaded area represents the proportion of area in the sampling distribution that lies beyond this obtained statistic, out in the two tails. The amount of shaded area totals to .085 of the sampling distribution.

If the null hypothesis were true, the probability is .085 that Dr. O’Day would have obtained a statistic as large (or larger) than 1.73. In the disciplines of research and statistics, this is considered a fairly high probability.

For this criterion variable, Dr. O’Day’s obtained $p$ value ($p = .085$) was larger than the alpha level of $\alpha = .05$. This meant that she failed to reject the null hypothesis and was forced to tell the world that there was a statistically nonsignificant difference between the two conditions with respect to their mean scores on other-reported irritability.

**SUMMARY.** There is no difference between the old-school approach and the new-school approach with respect to the final decision on the null hypothesis—if one approach indicates that the results are statistically significant, the other approach will arrive at the same conclusion. The main difference involves precision: The new-school approach easily produces a precise $p$ value such as $p = .543$ or $p = .004$, whereas the old school approach does not. The Publication Manual of the American Psychological Association 6th ed. (2010) recommends reporting precise $p$ values out to two or three decimal places whenever possible, and this means that the new-school approach is generally the preferred paradigm for performing null-hypothesis significance tests.

**Type I Errors and Type II Errors**

When testing null hypotheses, researchers sometimes make mistakes called *Type I errors* or *Type II errors*. This section describes the differences between the two types of mistakes, and summarizes the steps that can be taken to avoid them.

**TYPE I ERRORS.** A *Type I error* can be defined in the following ways:

- A Type I error involves rejecting a true null hypothesis.

- A Type I error occurs when a researcher tells the world that the results are *statistically significant* when (unbeknownst to the researcher) this is the wrong conclusion.

To expand upon these definitions a bit, think about how Dr. O’Day might make a Type I error in her caffeine study. Imagine that, in the population, caffeine has no effect on irritability at all. In other words, in the population there is no difference between the mean self-reported irritability scores for the high-caffeine condition versus the no-caffeine condition.

Imagine that Dr. O’Day conducts her caffeine study with 150 subjects. Before analyzing the data, she sets alpha at $\alpha = .05$. 
Due to sampling error, the mean score for the high-caffeine condition happens to be much higher than the mean score for the no-caffeine condition—high enough so that the obtained $t$ statistic is way out in the region of rejection ($p = .002$). Because the obtained $p$ value (.002) was smaller than the selected level of alpha (.05), Dr. O'Day rejects the null hypothesis. She tells the world that there is a statistically significant difference between the two conditions.

Dr. O’Day rejected the null hypothesis when she should not have done so. She has committed a Type I error.

Making a Type I error is a bad thing, but researchers have a way of protecting themselves: They can select a conservative value for alpha ($\alpha$) prior to analyzing the data. You will recall that $\alpha$ is the size of the region of rejection, out in the extreme tails of the distribution.

In addition to being the size of the region of rejection, alpha also represents the probability of making a Type I error. When a researcher sets alpha at .05 (and conducts the statistical analysis correctly), the probability of making a Type I error is .05. By setting alpha at a relatively small, conservative value (such as $\alpha = .001$) rather than a large, liberal value (such as $\alpha = .05$), the researcher decreases the likelihood of making a Type I error. If the researcher sets alpha at $\alpha = .001$ and still obtains significant results, it is possible that she is making a Type I error, but the probability that she is making a Type I error is very small—it is less than .001.

**Comparison-wise versus Familywise Type I Error Rate.** The *comparison-wise Type I error rate* (also known as the *per-comparison error rate*; symbol: $\alpha_{PC}$) is the probability of making a Type I error when performing just one specific significance test (i.e., just one comparison). Imagine for a moment that Dr. O’Day has just one dependent variable in her study: scores on the self-reported irritability index. If she set alpha at $\alpha = .05$ for the $t$ test performed on this dependent variable, then the comparison-wise Type I error rate is equal to .05.

There is a problem, however. In the real world, researchers almost always have more than one dependent variable. Earlier in this chapter, you learned that Dr. O’Day actually measured four dependent variables in the caffeine study. She will perform a separate $t$ test for each of these dependent variables. Unfortunately, this means that the probability of making at least one Type I error out of all of these tests will probably be greater than .05.

The *familywise Type I error rate* (symbol: $\alpha_{FW}$; also known at the *experiment-wise Type I error rate*) is the probability of making at least one Type I error out of all of the comparison tests performed on the same data set (Hays, 1988). When the individual significance tests are independent of one another and the same level of alpha is used for each comparison, the familywise Type I error rate can be estimated as:

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^K$$
With the above formula, \( \alpha_{FW} = \) the familywise Type I error rate; \( \alpha_{PC} = \) the level of alpha selected for each comparison; and \( K = \) the total number of comparisons to be made.

For example, imagine that Dr. O’Day plans to perform four \( t \) tests (one for each dependent variable), and that she will set alpha at \( \alpha = .05 \) for each comparison. The familywise error rate may be estimated as:

\[
\alpha_{FW} = 1 - (1 - \alpha_{PC})^K = 1 - (1 - .05)^4 = 1 - (.95)^4 = 1 - (.815) = .185 = .19
\]

So the familywise Type I error rate is .19. This means that Dr. O’Day has a 19% probability of making at least one Type I error out of all of the individual \( t \) tests that she plans to perform. Ouch!

Fortunately, there are a number of ways to deal with a familywise Type I error rate that exceeds the conventional level of .05. One popular (and fairly easy) approach involves using the \textit{Bonferroni adjustment} (Howell, 2002; Warner, 2008). With the \textit{Bonferroni adjustment} (also known as the \textit{Bonferroni correction}), the researcher divides the desired familywise error rate (usually .05) by the number of comparisons to be made (four, in this case). The result is the \textit{adjusted alpha level} (\( \alpha_{ADJ} \)). In performing the individual comparisons, the researcher views a given comparison as being statistically significant only if the obtained \( p \) value is less than \( \alpha_{ADJ} \).

The general form for the Bonferroni adjustment is:

\[
\alpha_{ADJ} = \frac{\alpha_{FW}}{K}
\]

With the preceding formula, \( \alpha_{ADJ} = \) the adjusted value of alpha which will be used for each individual significance test; \( \alpha_{FW} = \) the desired value for the familywise Type I error rate; and \( K = \) the number of comparisons (i.e., the number of significance tests) to be performed.

Again, let’s assume that Dr. O’Day wants the familywise Type I error rate to be \( \alpha_{FW} = .05 \) and she must make four comparisons. Inserting these values into the preceding formula results in the following adjusted alpha level:

\[
\alpha_{ADJ} = \frac{\alpha_{FW}}{K} = \frac{.05}{4} = .0125
\]

This means that, when performing the individual \( t \) tests, Dr. O’Day would consider the results for a given dependent variable to be significant only if the \( p \) value for that test is less than .0125. By using this Bonferroni adjustment, the familywise Type I error rate will remain at \( \alpha_{FW} = .05 \).

**Type II Errors.** A \textit{Type II error} can be defined in the following ways:

- A Type II error involves failing to reject a false null hypothesis.
- A Type II error occurs when a researcher tells the world that the results are \textit{not} statistically significant when (unbeknownst to the researcher) this is the wrong conclusion.
To make this a bit more concrete, consider this twist of events in Dr. O’Day’s study: Imagine that, in the population, caffeine does have a significant and substantial effect on irritability. In other words, in the population there is a difference between the mean self-reported irritability scores for the high-caffeine condition versus the no-caffeine condition.

Imagine that Dr. O’Day conducts her caffeine study with 150 subjects and, due to sampling error, the mean score for the high-caffeine condition happens to be very similar to the mean score for the no-caffeine condition—so similar that the obtained $t$ statistic is close to zero, solidly in the region of nonsignificance. Let’s imagine that $t_{obt} = 0.20$, $p = 0.841$.

Because the obtained $p$ value (.841) is larger than the selected value of alpha (.05), Dr. O’Day fails to reject the null hypothesis. She tells the world that there is not a statistically significant difference between the two conditions.

Dr. O’Day failed to reject the null hypothesis when she should have rejected it. She has made a Type II error.

Making a Type II error is a bad thing. But researchers know that they can protect themselves against Type II errors by maximizing the power of their statistical procedures. In the previous section titled “Factors Affecting the Power of a Test” you learned about several strategies that researchers can use to increase the power of their tests. These strategies included using a larger sample, setting alpha at $\alpha = 0.05$ rather than $\alpha = 0.001$, and related tactics.

### How Significance is Reported in Articles

Researchers follow a fairly standard set of conventions when presenting the results of significance tests in published reports. This section introduces some of those conventions.

**USING PRECISE $p$ VALUES TO REPORT SIGNIFICANCE.** The *Publication Manual of the American Psychological Association* 6th ed. (2010) recommends that, when feasible, researchers should present the precise $p$ values from significance tests, carried to two or three decimal places. This approach had been used in Table 5.1, a revised version of which is reproduced here as Table 5.2.

The information relevant to the significance tests appears under the general heading, “Difference between means.” The column headed “$t(148)$” provides the obtained $t$ statistic for each dependent variable. The entry “(148)” in this heading indicates that the degrees of freedom for each $t$ test was $df = 148$. This column shows that the obtained $t$ statistic for self-rated irritability was $t_{obt} = 5.18$, with corresponding $p$ value of $p < .001$ (in the “$p$” column). The $p$ value is less than .05, so Dr. O’Day rejected the null hypothesis for self-rated irritability. The results for the remaining criterion variables can be interpreted in the same way.
Table 5.2

Scores on Four Measures of Irritability as a Function of Amount of Caffeine Consumed (High-Caffeine versus No-Caffeine): Results from Independent-samples t Tests with p Values

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>High caffeine $^a$</th>
<th>No caffeine $^b$</th>
<th>Difference between means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$     (SD$_1$)</td>
<td>$M_2$    (SD$_2$)</td>
<td>$M_1 - M_2$, $SE$, $t(148)$, $p$</td>
</tr>
<tr>
<td>Self-rated irritability</td>
<td>575.00 (35.49)</td>
<td>544.73 (36.13)</td>
<td>30.27, 5.85, 5.18, .001</td>
</tr>
<tr>
<td>Other-rated irritability</td>
<td>556.46 (35.86)</td>
<td>546.27 (36.10)</td>
<td>10.19, 5.88, 1.73, .085</td>
</tr>
<tr>
<td>Self-reported outbursts</td>
<td>6.24 (2.67)</td>
<td>5.03  (2.69)</td>
<td>1.21, 0.44, 2.77, .006</td>
</tr>
<tr>
<td>Other-reported outbursts</td>
<td>5.24 (2.67)</td>
<td>4.99  (2.65)</td>
<td>0.25, 0.43, 0.58, .561</td>
</tr>
</tbody>
</table>

Note. $N = 150$.

* $n = 75$. $^b n = 75$.

Within the text of the published article, authors are expected to highlight the investigation’s most important findings. Dr. O’Day does this in the following excerpt:

Because the study included four dependent variables, the Bonferroni adjustment was used to maintain a familywise Type I error rate of .05. The resulting Bonferroni-adjusted alpha level was $\alpha_{ADJ} = .0125$.

Using this adjusted level of alpha, an independent-samples $t$ test revealed a significant difference between the high-caffeine condition and the no-caffeine condition for self-rated irritability, $t(148) = 5.18, p < .001$, and self-reported outbursts, $t(148) = 2.77, p = .006$. However, the $t$ test revealed a nonsignificant difference for other-rated irritability, $t(148) = 1.73, p = .085$, and other-reported outbursts, $t(148) = 0.58, p = .561$.

Some notes regarding the preceding excerpt:

- Early in the Results section, investigators should indicate the level of alpha that was selected prior to performing the significance tests. Dr. O’Day does this in the first paragraph of the preceding excerpt.

- When reporting the obtained $p$ values, Dr. O’Day generally uses the “=” symbol and reports the $p$ value to three decimal places, as recommended by the *Publication Manual of the APA* (2010). For example, with self-reported outbursts, she indicates “$p = .006$.” The one exception is when she reports the $p$ value for self-rated irritability. The output from the statistical
application indicated that the actual $p$ value was a very small probability—smaller than .0001. The publication manual of the APA recommends reporting $p$ values to no more than three decimal places, so Dr. O’Day presented it as “$p < .001$.” Be warned that statistical applications sometimes report $p$ values as “$p = .0000$.” In these cases, the statistical application is attempting to tell you that the $p$ value is a very small number (such as .00000001). However, in these instances you should not report the $p$ value as “$p = .0000$” in your article. If you reported it in this way, you would be telling the reader “if the null hypothesis were true, the probability that I would have obtained the current sample results is zero.” Such a statement is not logical—in the field of statistics, certain events are unlikely, but no event is impossible. When statistical applications indicate that “$p = .0000$,” it is best to report it as “$p < .001$,” as Dr. O’Day did in her excerpt.

Table 5.3

Scores on Four Measures of Irritability as a Function of Amount of Caffeine Consumed (High-Caffeine versus No-Caffeine): Results from Independent-samples $t$ Tests with Asterisks

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>High caffeine $a$</th>
<th>No caffeine $b$</th>
<th>Difference between means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$ (SD$_1$)</td>
<td>$M_2$ (SD$_2$)</td>
<td>$M_1 − M_2$ SE $t(148)$</td>
</tr>
<tr>
<td>Self-rated irritability</td>
<td>575.00 (35.49)</td>
<td>544.73 (36.13)</td>
<td>30.27 5.85 5.18**</td>
</tr>
<tr>
<td>Other-rated irritability</td>
<td>556.46 (35.86)</td>
<td>546.27 (36.10)</td>
<td>10.19 5.88 1.73</td>
</tr>
<tr>
<td>Self-reported outbursts</td>
<td>6.24 (2.67)</td>
<td>5.03 (2.69)</td>
<td>1.21 0.44 2.77*</td>
</tr>
<tr>
<td>Other-reported outbursts</td>
<td>5.24 (2.67)</td>
<td>4.99 (2.65)</td>
<td>0.25 0.43 0.58</td>
</tr>
</tbody>
</table>

Note. $N = 150$.  
$a_n = 75$. $b_n = 75$.  
*$_p < .01$. **$_p < .001$.

**Using Asterisks to Report Significance.** Prior to the availability of computer programs that provided $p$ values, it was standard for researchers to use asterisks in tables to “flag” (i.e., to identify) statistics that were significant. Today, the use of asterisks is declining, in part because of the recommendation from the APA’s Publication Manual that researchers should report precise $p$ values whenever feasible. The Publication Manual does, however, allow researchers to use asterisks rather than $p$ values in certain cases (e.g., when there is not enough room in the table for...
precise $p$ values). Table 5.3 shows how Table 5.2 could be revised to make use of asterisks rather than $p$ values.

In Table 5.3, asterisks are placed next to obtained $t$ statistics that are significant. The note at the bottom of Table 5.3 shows that one asterisk (*) identifies statistics that are significant at the .01 level, whereas two asterisks (**) identify statistics that are significant at the .001 level. As the reader, you should always consult the notes at the bottom of the table to determine what level of significance is signified by the number of asterisks.

**USING THE $p <$ AND $p >$ CONVENTIONS TO REPORT SIGNIFICANCE.** Within the text of the published article, authors have traditionally used the “less than” symbol (“<”) and “greater than” symbol (“>”) to identify statistics that are significant or nonsignificant. The *Publication Manual of the American Psychological Association 6th ed.* (2010) discourages this practice, advising that researchers should instead report precise $p$ values using the “=” symbol whenever possible (e.g., “$p = .006$”). However, you should expect to encounter the use of the “$p <$” and “$p >$” conventions fairly often, especially when you read older articles. This section provides a short guide to their interpretation.

The “less than” symbol “<” is used to identify statistics that are significant at a given level of alpha. For example, if the researchers had set alpha at $\alpha = .05$ prior to analyzing the data and then found that the obtained statistic was significant, they might report it in this way:

The results of the independent-samples $t$ test were significant, $t(148) = 4.12, p < .05$

If the researchers had instead set alpha at $\alpha = .001$ prior to analyzing the data and then found that the obtained statistic was significant, they would report it in this way:

The results of the independent-samples $t$ test were significant, $t(148) = 4.12, p < .001$

In essence, the preceding sentence is communicating the following message:

My obtained statistic was $t_{obt} = 4.12$. If the null hypothesis were true, the probability that I would have obtained a sample statistic this large (or larger) is less than .001. Therefore, I rejected the null hypothesis.

In contrast, the “greater than” symbol (“>”) is used to identify statistics that are not significant at a given level of alpha. For example, if the researchers had set alpha at $\alpha = .05$ prior to analyzing the data and then found that the obtained statistic was not significant, they might report it this way:

The results of the independent-samples $t$ test were nonsignificant, $t(148) = 1.10, p > .05$

In essence, the preceding sentence communicates the following:

My obtained statistic was $t_{obt} = 1.10$. If the null hypothesis were true, the probability that I would have obtained a sample statistic this large (or larger) is greater than .05. Therefore, I failed to reject the null hypothesis.
I, for one, absolutely hate the use of “p <” and “p >” to identify significant versus nonsignificant findings. While reading an article, I am forced to stop in mid-sentence and try to recall whether the “>” symbol means significant or nonsignificant. In an act of great mercy, the Publication Manual of the APA also allows the use of the abbreviation ns to identify nonsignificant statistics. Authors using this abbreviation should be sure to indicate the level of alpha that had been chosen for the analysis. To use this ns abbreviation, a researcher might re-write one of the preceding sentences in this way:

With alpha set at $\alpha = .05$, the analysis revealed a nonsignificant difference between the means of the two conditions, $t(148) = 1.10$, ns.

The Controversy Regarding Significance Testing

Up until this point, most of this chapter has focused on just one type of statistical procedure: the null-hypothesis significance test (abbreviation: NHST). Most statistics textbooks spend a good deal of time on this procedure because it is so widely-reported in journal articles.

**Criticisms of Significance Tests.** In recent decades, however, many researchers have criticized the role that the NHST currently plays in the social and behavioral sciences. These critics have argued that significance testing is illogical, does not tell researchers what they really want to know, produces results that are often misinterpreted, and has therefore slowed the pace of scientific advancement (e.g., Cohen, 1994; Fidler & Cumming, 2008; Kirk, 1996; Kline, 2004). Some well-respected researchers have suggested that the practice of significance testing should be abolished entirely (e.g., Hunter, 1997).

**Example: Misinterpretation of Small p Values.** One of the most common criticisms of the NHST involves the fact that many people mistakenly believe that, if the obtained $p$ value for an analysis is very small (such as $p < .001$), it indicates that there is a strong relationship between the predictor variable and the criterion variable. In reality, a small $p$ value does not necessarily mean this at all.

For example, imagine that Dr. O’Day had performed her caffeine investigation using very large samples—say, 1,000 participants in each condition. Imagine that the mean number of self-reported outbursts per week was equal to 5.10 for the high-caffeine condition and 5.00 for the no-caffeine condition. The difference between the two means is computed as 5.10 $- 5.00 = 0.10$. This difference is fairly trivial in size—it shows that the difference between the two groups is equal to one-tenth of one temper tantrum per week. Big deal.

There is a problem, however. In an earlier section, you learned that researchers can increase the power of a significance test by using larger samples. By using 1,000 participants in each condition, Dr. O’Day might have made her test so powerful that even a trivial difference between the means could result in a large obtained $t$ statistic (say $t_{obt} = 4.00$) with a really small corresponding $p$ value (say, $p < .001$). Dr. O’Day might then summarize her findings in this way:
Analyses revealed a highly significant difference between the two means, \( t(1,998) = 4.00, \ p < .001 \).

Many readers would be misled by this finding. When they see the tiny \( p \) value of \( p < .001 \), many readers will think “Yikes! Caffeine must really have a big effect on self-reported outbursts.” And if they never bothered to check the means, they would never know any better.

Okay—no one would really say “Yikes!” But they would be misled.

By the way, you should just about never use expressions such as “highly significant” in an article. The decision regarding statistical significance is a dichotomous decision: The results are either significant, or they are not. Telling a reader that your results are “highly significant” is like telling someone that a mutual friend’s newborn baby not only is a boy, but “He is highly a boy!”

**Response to criticism from the APA.** In response to these and other criticisms, the American Psychological Association formed a task force of well-known researchers and statisticians to review the role of significance testing in psychological research and to make recommendations for the future. In its report (Wilkinson & the Task Force on Statistical Inference, 1999), the task force did not recommend that researchers abandon significance testing. Instead, it recommended that researchers should supplement significance tests with additional statistical procedures in order to provide a clearer, more comprehensive understanding of the results of investigations. These additional procedures include confidence intervals as well as indices of effect size. Much of the remainder of this chapter is devoted to these latter two procedures.

**An Alternative to the NHST: Replication Statistics and \( p_{rep} \)**

In part due to the controversy described above, many researchers have turned away from null-hypothesis significance testing, and have replaced these procedures with replication statistics: procedures which estimate the probability that another investigation using the same research method and population of participants would replicate the current investigation’s results (Killeen, 2005a, 2005b, 2008). This movement has been spurred, in part, by the fact that some prestigious journals (such as *Psychological Science*) encourage authors to report \( p_{rep} \) rather than the traditional \( p \) value.

The exact meaning of \( p_{rep} \) depends on how the researcher defines “replicate,” and this definition may vary from study to study. In general, the \( p_{rep} \) statistic that is computed for a specific, individual study estimates the long-run probability that an exact replication of that study will support a specific research claim—that it will produce an effect in a specified direction (Killeen, 2008). Values of \( p_{rep} \) may range from 0.00 (meaning there is no probability that another study would replicate these findings) to 1.00 (meaning that there is a 100% probability that another study would replicate these findings).

For example, imagine that Dr. O’Day predicts that consuming caffeine will cause an increase in self-reported irritability. She conducts the study reported earlier,
finds that the mean self-reported irritability score for the high-caffeine condition is in fact higher than the mean for the no-caffeine condition, and computes $p_{rep} = .93$. This indicates that, if another researcher conducted exactly the same study with a different sample from the same population, she estimates that there is a 93% probability that the second study would display the same trend in the results, with the high-caffeine condition displaying higher levels of irritability than the no-caffeine condition.

How large does $p_{rep}$ have to be for the researcher to conclude that the study’s results provide strong support for their research hypothesis? At present, there do not appear to be hard-and-fast criteria, although Killeen (2008) mentions $p_{rep} = .90$ as being in the “respectable range” (p. 114).

Many researchers are attracted to the $p_{rep}$ concept, as it is readily interpretable and addresses the issue of replication, which is always important in empirical research. As more graduate programs incorporate replication statistics as part of their curriculum (and as more computer applications make it easy to compute $p_{rep}$), you should see this alternative to the null-hypothesis significance test appear more and more frequently in research articles.

**Final Thoughts on Significance Testing**

Remember: when you think *statistical significance*, think *probability* or *likelihood*. When results are statistically significant, it does not necessarily tell us that there is a strong relationship between the predictor variable and the criterion variable, and it does not necessarily tell us that the statistics computed from this sample are precise estimates of the corresponding population parameters. It merely tells us that, if the null hypothesis were true, it is *unlikely* that we would have obtained sample results such as these.

**Big-Three Results, Part 2: Confidence Intervals**

When you think *confidence interval*, you should think *precision*. A confidence interval communicates whether your estimate of a population parameter (based on the sample data) is relatively precise, or is relatively imprecise.

**Confidence Intervals: Basic Concepts**

A *parameter* is some characteristic of a population (such as the population mean). In most cases, the actual value of a population parameter is unknown, and this forces us to rely on estimation. *Estimation* is the process through which we analyze sample data in order to arrive at a “best guess” as to what this population parameter is likely to be.

When estimating population parameters, researchers may report point estimates, interval estimates, or both. A *point estimate* is a single number that is used as the best estimate of the corresponding population parameter (Hays, 1988). In contrast, a *confidence interval* (symbol: CI) is a range of plausible values for the
The sample mean was \( M = 110 \), indicating that the best estimate for the population mean was also 110. Given the 95% confidence interval, it was plausible that the sample was drawn from a population whose mean was somewhere between 105 and 115.

Be warned that the interpretation of confidence intervals is a tricky business. A later section will discuss correct (as well as incorrect) interpretations of these interval estimates.

**WHY CONFIDENCE INTERVALS ARE USEFUL.** Interval estimates (i.e. confidence intervals) tend to be more useful than point estimates because only interval estimates communicate precision. A confidence interval communicates whether the parameter estimates you have computed from your sample data are relatively precise—or relatively imprecise—estimates.

The problem with point estimates (when presented alone, without confidence intervals) is that they provide no information about precision, and can therefore be misleading. For example, imagine that we rewrote the preceding excerpt so that it presented only the point estimate, and left out the confidence interval:

The sample mean was \( M = 110 \), indicating that the best estimate for the population mean was also 110.
Many readers would interpret the preceding as saying that mean in the population is exactly equal to 110, period. They would not realize that some error was almost certainly involved in this estimate.

In contrast, confidence intervals are very up-front in acknowledging the amount of error that is involved when we estimate population parameters from sample data. For example, consider the IQ study described above: If your sample mean was 110 and the 95% CI extended from 108 to 112, this would be seen as a fairly narrow range of scores. You could, therefore, assume that your estimate of the mean in the population (110) was a relatively precise estimate.

For purposes of contrast, consider a different outcome in which the mean was the same number (110) but the 95% CI was much wider—extending from 90 to 130. Because your confidence interval covered such a wide range of values, you would be forced to assume that there was a good deal of error associated with your point estimate. In other words, your point estimate of 110 could no longer be viewed as a terribly precise estimate of the population mean.

**Correct (and incorrect) interpretations of a confidence interval.** The concept of confidence interval can be defined in a number of different ways. This section provides some correct definitions and/or interpretations for confidence interval from well-regarded sources. These definitions refer to situations in which the researcher is using sample data to estimate \( \mu \), a population’s mean.

Cumming and Finch (2005) provide the following: “Our CI is a range of plausible values for \( \mu \). Values outside the CI are relatively implausible” (p.174). Aron et al. (2006) offer this: “Roughly speaking, [the CI is] the region of scores...that is likely to include the true population mean; more precisely, the region of possible population means for which it is not highly unlikely that one could have obtained one’s sample” (p. 711). Warner (2008) defines a confidence interval as “…a range of values above and below a sample statistic that is used as an interval estimate of a corresponding population parameter” (p.1001).

Based on these definitions, a correct interpretation of the 95% CI for the IQ study described above might go something like this:

**Correct:**

The sample mean was \( M = 110 \), 95% CI [105, 115]. Given the confidence interval, it was plausible that the sample was drawn from a population whose mean was somewhere between 105 and 115.

Notice that the preceding included the word “plausible;” it did not include the word “probability.” This is because it is generally a bad idea to use the word “probability” any time that you interpret a confidence interval. For example, below is an incorrect interpretation of the CI for the IQ study:
Incorrect:

The sample mean was $M = 110$, 95% CI [105, 115]. Given the confidence interval, there was a 95% probability that the true mean in the population ($\mu$) was somewhere between 105 and 115.

What was wrong with the second interpretation? The problem is that the “95%” in “95% confidence interval” does not tell you that there is a 95% probability that the true population parameter $\mu$ is included within the interval. A confidence interval does not make a probability statement about a population—it makes a probability statement about your sample. In this case, the confidence interval tells you that there is a 95% probability that this confidence interval (based on the current sample) is one of the confidence intervals that actually captured the true population parameter, out of the infinite number of confidence intervals that theoretically could be computed if the study were conducted an infinite number of times (Field, 2009a; Warner, 2008). Understanding this concept requires a little imagination:

- Imagine that, in the population of people taking this IQ tonic, the population mean really is 110 (that is, $\mu = 110$).

- Assume that you draw a sample of 20 people from this population and compute their mean IQ score. Assume that the mean for this sample is $\bar{X} = 108$, and the 95% confidence interval extends from 103 to 113. In Figure 5.5, the mean and confidence interval for this mini-study are represented by the small box-and-whisker plot that appears to the right of the label “Study 1.” The small black box represents the mean of 108, and the whiskers extending out from the box represents the confidence interval.

- Assume that you draw a new sample of 20 different people from this same population and compute their mean IQ score. Assume that the mean for this sample is $\bar{X} = 106$, and the 95% confidence interval extends from 101 to 111. This mean and confidence interval appears in Figure 5.5, to the right of the label “Study 2.”

- Finally, imagine that you repeat this process a total of 20 times, each time computing the sample mean and the 95% confidence interval. These means their confidence intervals are represented by the remaining box-and-whisker plots in Figure 5.5.

Remember that (unbeknownst to you), the actual mean in the population was $\mu = 110$. If you closely inspect the 20 confidence intervals in Figure 5.5, you will see that:

- 19 of the 20 confidence intervals (that is, 95% of the intervals) actually “capture” the true population mean of $\mu = 110$, and
1 of the 20 confidence intervals (that is, 5% of the intervals) does not “capture” the true population mean of \( \mu = 110 \). The study that did not capture the true population mean was Study 10—notice that the whiskers for this boxplot do not overlap the vertical line that represents a score of 110.

![Figure 5.5. Mean IQ scores (represented by boxes) and their 95% confidence intervals (represented by whiskers) for 20 studies; \( n = 20 \) for each study.](image)

When we use a 95% confidence interval and conduct the study an infinite number of times, we expect that 95% of the confidence intervals will capture the true population mean. That is why 95% (i.e., 19 out of 20) of the confidence intervals in Figure 5.5 capture the true population mean of 110.

Remember that a confidence interval allows you to make a probability statement about your sample confidence interval. It does not allow you to make a probability statement about the population parameter (such as \( \mu \)) that you are trying to estimate. This is why this text recommends that you use words such as “plausible” or “confidence” rather than the word “probability” when discussing a CI.

**Computing the 95% CI for a Sample Mean.** The specific nature of the formula that will be used to compute a confidence interval is determined by a variety of factors (e.g., the desired level of precision, the nature of the investigation, as well as other considerations). Below is the formula for computing the 95% CI for a sample mean:

\[
95\% \text{ CI} = \bar{X} \pm (SE_M) (t_{crit})
\]
With the preceding formula, $\bar{X}$ = the observed sample mean; $SE_m$ = the standard error of the mean (i.e., the standard deviation of the sampling distribution of means); and $t_{crit}$ = the critical value of the $t$ statistic, found in a table in the back of a statistics textbook.

**Using Confidence Intervals to Test the Null Hypothesis**

Seasoned consumers of research can review a confidence interval and immediately know whether the results are statistically significant. This section shows how it is done.

**The Rule.** Using a confidence interval to determine significance is fairly simple, as long as you know the null hypothesis being tested. Here is the rule:

If the population parameter described in the null hypothesis is outside of the obtained confidence interval, the results are statistically significant.

In other words, outside = significant. This following section explains why this is so.

**An Example with Significant Results.** For example, consider the single-sample study in which you developed a tonic that you believed would increase IQ scores. In your investigation, you measured intelligence with a test that produces a mean IQ score of 100 in the population of “normal” adults (that is, adults who have not taken your tonic). This means that, in the population of normal adults, the mean IQ score is $\mu = 100$.

You are now ready to conduct your study in which you will administer the tonic to a random sample of 20 adults, wait two weeks (for the tonic to have its effect), and then measure the IQ of each subject in your sample. You hope that your tonic will cause your sample to display a mean IQ score that is higher than the mean of the normal population. Therefore, when you analyze your data, you will test the following statistical null hypothesis:

$H_0: \mu = 100$. In the population, the mean IQ score is equal to 100.

Remember that the null hypothesis is the hypothesis that you (the researcher) hope to reject once you have analyzed your data. In this case, you hope that the mean IQ for the 20 subjects who took the tonic will be substantially higher than the mean IQ of 100 that is typically seen in the no-tonic population.

At the end of the investigation, you assess the IQ of each participant in the sample who took the tonic. To your delight, you find that the sample mean IQ for these 20 individuals is $\bar{X} = 120$, which is very high. You compute the 95% confidence interval around this sample mean, and find that the 95% CI extends from 110 to 130.

Knowing only these facts, most researchers and statisticians would understand that your results are statistically significant ($\alpha = .05$). They would understand that the sample of 20 individuals who took the tonic displayed a mean IQ score that was significantly higher than the population of regular people who have not taken the tonic.
How would they know this? Because the confidence interval did not contain the population parameter that had been described in the null hypothesis. This state of affairs is illustrated graphically in Figure 5.6.

The horizontal line in Figure 5.6 represents scores on a measure of IQ. In the population of “normal” adults (adults who have not taken your tonic), the mean IQ score is $\mu = 100$. In the figure, a dashed arrow points to this population mean of 100. You began with the null hypothesis that your 20 subjects came from a population in which the mean IQ score is $\mu = 100$. By the end of your investigation, you hope to reject this null hypothesis. By the end of your study, you hope to have evidence that the mean IQ score for the 20 people who took your tonic is significantly different from a mere 100 points.

For the 20 participants who took the IQ tonic, the sample mean was a fairly high value: $\bar{X} = 120$. This sample mean is represented by the solid arrow. The 95% confidence interval around the sample mean is represented by the horizontal bracket centered above $\bar{X}$. Notice that this bracket extends from the lower limit of the confidence interval ($LL = 110$) to the upper limit ($UL = 130$). Notice also that the bracket representing the confidence interval does not contain the null-hypothesis population mean of 100. This tells you that it is unlikely that your sample of 20 tonic-drinkers were drawn from a population that displayed a mean IQ score of just 100. In other words, it tells you that there is a statistically significant difference between the sample mean of $\bar{X} = 120$ versus the null-hypothesis population mean of $\mu = 100$. Because of this, you may reject the null hypothesis and tell the world that you have statistically significant results.
AN EXAMPLE WITH NONSIGNIFICANT RESULTS. For the sake of contrast, imagine that you conducted the same study but instead obtained very different results. Imagine that the sample of 20 participants who took the IQ tonic now display a lower mean score—a mean score of just $\bar{X} = 105$. The 95% confidence interval for this sample mean extends from 95 to 115. In this second case, the confidence interval now contains the population mean described under the null hypothesis ($\mu = 100$).

This state of affairs is illustrated in Figure 5.7. In the figure, the obtained sample-mean for the 20 tonic-drinkers ($\bar{X} = 105$) is now pretty close to the null-hypothesis population mean ($\mu = 100$). The two means are so close to one another that the null-hypothesis population mean ($\mu = 100$) is now contained within the 95% confidence interval for the sample mean. This tells you that there is not a statistically significant difference between the sample mean of $\bar{X} = 105$ versus the null-hypothesis population mean of $\mu = 100$. Because of all this, you fail to reject the null hypothesis—you tell the world that you do not have statistically significant results.

Figure 5.7. Statistically nonsignificant results: Mean IQ score for the null-hypothesis population (represented by the dashed arrow), mean IQ score for the sample of 20 participants who took the tonic (represented by the solid arrow), and 95% CI for this sample mean (represented by the bracket).

CONFIDENCE INTERVALS AND CORRESPONDING ALPHA LEVELS. The first of the two preceding examples illustrated statistically significant results. In a research article, you might have summarized these results in this way:
For the 20 participants who drank the IQ tonic, the sample mean relatively high, $M = 120$, 95% CI [110, 130]. The 95% CI did not capture the population mean described by the null hypothesis ($\mu = 100$), indicating that the results were statistically significant ($\alpha = .05$).

This sentence ends with the symbol, “$\alpha = .05$,” indicating that the results were statistically significant with alpha set at .05. You will recall that the alpha level (or significance level) refers to the size of the region of rejection in a sampling distribution. When you use confidence intervals to test a null hypothesis, the size of the CI determines the alpha level for the significance test. In general, it works like this:

- A 90% confidence interval corresponds to a significance test with $\alpha = .10$
- A 95% confidence interval corresponds to a significance test with $\alpha = .05$
- A 99% confidence interval corresponds to a significance test with $\alpha = .01$

**Confidence Interval for the Difference Between Two Means**

Preceding sections of this chapter have discussed confidence intervals that have been computed around sample means. Researchers also sometimes compute confidence intervals around differences between sample means. This is typically done when subjects have been assigned to two or more treatment conditions, and the researcher has computed the mean score for each condition on some criterion variable. The researcher then computes a mean-difference score: a single number that represents the size of the difference between the two treatment means. Finally, the researcher computes the confidence interval around this mean-difference score. This section discusses how these confidence intervals for differences may be interpreted, and shows how you may use them to determine whether the observed difference between the two means is statistically significant.

**Point estimate for the difference score.** To illustrate these concepts, this section will refer to the caffeine study that was described earlier in this chapter. You may recall that one of the criterion variables in the caffeine study was self-reported outbursts: the number of angry outbursts (i.e., temper tantrums) per week that were reported by each participant. In this fictitious study, participants in the high-caffeine condition reported 6.24 outbursts per week, whereas subjects in the no-caffeine condition reported 5.03 outbursts. The observed mean-difference score is computed as:

$$\bar{X}_{\text{diff}} = \bar{X}_E - \bar{X}_C$$

$$\bar{X}_{\text{diff}} = 6.24 - 5.03$$

$$\bar{X}_{\text{diff}} = 1.21$$
So the observed difference between sample means was equal to 1.21. It is useful to think of this “1.21” as a point estimate. Given this finding in the sample, you estimate that the mean-difference score is also equal to 1.21 in the population.

**CONFIDENCE INTERVAL FOR THE DIFFERENCE SCORE.** A confidence interval can be constructed around a mean-difference score, just as one can be constructed around a mean. Assume that Dr. O’Day does this and reports it in the following excerpt:

> On the criterion variable self-reported outbursts, the mean score for the high caffeine condition was 1.21 units higher than the mean for the no-caffeine condition, 95% CI [0.36, 2.08].

The above indicates that the observed difference between the means of the two samples was 1.21, and the 95% confidence interval for this difference extends from 0.36 to 2.08. We can now use this confidence interval to determine whether there is a significant difference between the two sample means.

**USING THE CI FOR THE MEAN DIFFERENCE TO TEST THE NULL HYPOTHESIS.** The same rule that had been used to perform the significance test for a single-sample mean is also used to test the significance of the difference between two means. The rule is once again reproduced below:

> If the population parameter described in the null hypothesis is outside of the obtained confidence interval, the results are statistically significant.

The null hypothesis for the current caffeine study could be presented in this way:

\[
H_0 : \mu_E - \mu_C = 0. \text{ In the population, the difference between the mean outburst score for the experimental condition versus the mean outburst score for the control condition is equal to zero.}
\]

This null hypothesis says that the actual difference between the two means is equal to zero. If we compute a confidence interval for the differences between means in the sample and this CI does not contain the value of zero, then we may reject this null hypothesis.

**AN EXAMPLE WITH SIGNIFICANT RESULTS.** Table 5.4 presents the some of the same results that had appeared in other tables earlier in this chapter. We will use the results in this table to perform some significance tests.

The third dependent variable in Table 5.4 is self-reported outbursts. Table 5.4 shows that subtracting the mean outburst score for the no-caffeine condition from the mean for the high-caffeine condition produces a difference score of 1.21, and the 95% confidence interval for this difference extends from 0.35 to 2.08. This confidence interval does not contain the value of zero. Using the rule provided above, this tells us that there was a statistically significant difference between the two conditions.
Table 5.4

Scores on Four Measures of Irritability as a Function of Amount of Caffeine Consumed (High-Caffeine versus No-Caffeine): Observed Differences Between Means with 95% Confidence Intervals

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>High-caffeine</th>
<th>No-caffeine</th>
<th>Differencea</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported irritability</td>
<td>575.00</td>
<td>544.73</td>
<td>30.27**</td>
<td>18.71 - 41.82</td>
</tr>
<tr>
<td>Other-reported irritability</td>
<td>556.46</td>
<td>546.27</td>
<td>10.19</td>
<td>-1.42 - 21.80</td>
</tr>
<tr>
<td>Self-reported outbursts</td>
<td>6.24</td>
<td>5.03</td>
<td>1.21*</td>
<td>0.35 - 2.08</td>
</tr>
<tr>
<td>Other-reported outbursts</td>
<td>5.24</td>
<td>4.99</td>
<td>0.25</td>
<td>-0.61 - 1.11</td>
</tr>
</tbody>
</table>

Note. *N* = 150. CI = confidence interval for difference between means; LL = lower confidence limit; UL = upper confidence limit.

*Differences were computed by subtracting mean for the no-caffeine sample from the mean for the high-caffeine sample. Differences flagged with an asterisk are statistically significant based on a *t* test for independent samples (*df* = 148).

*p < .01; **p < .001.

These concepts are easiest to understand if we can see them illustrated graphically. Figure 5.8 displays the mean-difference score and the confidence interval just described.

The horizontal line in Figure 5.8 represents *difference scores*—scores obtained when the mean for the control condition is subtracted from the mean for the experimental condition ($\bar{X}_E - \bar{X}_C$). In the figure, the dashed line points to the difference score that we would expect to see if there was really no difference between the two means in the population: 0.00. Think about why this makes sense: If the high-caffeine groups did not display any more outbursts than the no-caffeine group, then the difference between the means for these two groups should be equal to 0.00.

However, that is not what Dr. O’Day found. In her study, the difference between the two sample means was actually 1.21. This observed difference is represented by the solid line in the figure. A mean difference of 1.21 is pretty far away from the value of 0.00 which had been predicted by the null hypothesis. But is it far enough
away for Dr. O’Day to conclude that her results are statistically significant? To find out, she reviewed the confidence interval for the difference.

![Figure 5.8](image)

**Figure 5.8.** Statistically significant results from Dr. O’Day’s analysis of self-reported outbursts: observed difference between means, 95% confidence interval for the observed difference, and difference that would be expected if the null hypothesis were true.

The confidence interval for the difference is represented by the bracket that appears above the horizontal line. You can see that the bracket extends from a low difference score of 0.35 (the lower limit for the CI) to a high difference score of 2.08 (the upper limit). This confidence interval tells us that it is plausible that the actual difference in the population might be as low as 0.35, or as high as 2.08. Notice that the bracket does not contain a difference score of 0.00. This tells us that Dr. O’Day’s difference of 1.21 is significantly different from 0.00. In other words, there was a statistically significant difference between the two sample means.

**AN EXAMPLE WITH NONSIGNIFICANT RESULTS.** The fourth dependent variable listed in Table 5.4 is *other-reported outbursts*: the number of temper tantrums displayed by the participant, according to people who know the participant. The table shows that subtracting the mean outburst score for the no-caffeine condition from the mean for the high-caffeine condition produced a difference score of 0.25, and that the 95% confidence interval for this difference extended from –0.61 to 1.11. This confidence interval does contain the value of 0.00—notice that it extends from a negative number (–0.61) to a positive number (+1.11). Using the general rule provided above, the fact that this confidence interval contains the value of zero tells us that there was not a statistically significant difference between the two conditions. These results are illustrated graphically in Figure 5.9.
The solid line shows that, for this new criterion variable (other-reported outbursts), the difference between the mean for the two groups was only 0.25. This is not terribly far away from the difference of zero that had been predicted by the null hypothesis.

To determine whether it represents a statistically significant difference, Dr. O’Day reviewed the confidence interval represented by the bracket above the horizontal line. This bracket extends from –0.61 to 1.11. The bracket very clearly “captures” the null-hypothesis value of 0.00, and this tells Dr. O’Day that the difference between means is not statistically significant.

In summary: Whenever you are reviewing a confidence interval for the difference between two means, you can use the following simplified rule for determining whether the observed difference is statistically significant:

If the confidence interval contains zero, it is plausible that the actual difference is zero.

**How Confidence Intervals are Presented in Articles**

feasible. It indicates that researchers should always report the confidence level for
the interval (i.e., researchers should indicate whether it is a 90%, a 95%, or a 99%
confidence interval). When feasible, researchers should use the same confidence
level for each statistic in the Results section (i.e., if you report a 95% CI for one
statistic, in most cases you should report a 95% CI for all statistics in that article).

**Presenting confidence intervals in tables.** Many research articles present
confidence intervals in tables under the heading “95% CI,” as was done in Table
5.4, above. This table also illustrates the standard convention of using the
headings “LL” and “UL” to identify lower limits and upper limits, respectively.

An alternative approach for presenting confidence intervals in a table is used with
Table 5.1, which appeared near the beginning of this chapter. For convenience, it
is reproduced below.

Table 5.5

Scores on Four Measures of Irritability as a Function of Amount of Caffeine Consumed (High-Caffeine
versus No-Caffeine): Significance Tests, Confidence Intervals, and Point-Biserial Correlations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>High caffeine$^a$</th>
<th>No caffeine$^b$</th>
<th>Difference between means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_1$ (SD$_1$)</td>
<td>$M_2$ (SD$_2$)</td>
<td>$M_1-M_2$ [95% CI]</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Self-rated irritability</td>
<td>575.00 (35.49)</td>
<td>544.73 (36.13)</td>
<td>30.27 [18.71, 41.82]</td>
</tr>
<tr>
<td>Other-rated irritability</td>
<td>556.46 (35.86)</td>
<td>546.27 (36.10)</td>
<td>10.19 [-1.42, 21.80]</td>
</tr>
<tr>
<td>Self-reported outbursts</td>
<td>6.24 (2.67)</td>
<td>5.03 (2.69)</td>
<td>1.21 [0.35, 2.08]</td>
</tr>
<tr>
<td>Other-reported outbursts</td>
<td>5.24 (2.67)</td>
<td>4.99 (2.65)</td>
<td>0.25 [-0.61, 1.11]</td>
</tr>
</tbody>
</table>

Note. $N=150$. CI = Confidence interval; $r_{pb}$ = Point-biserial correlation between the independent variable
and dependent variable.

$^a n = 75$. $^b n = 75$.

In Table 5.5, the observed difference between means for each criterion variable
appears below the heading “$M_1-M_2$.” The confidence intervals for these
differences appear below the heading “[95% CI].” Notice that the confidence
intervals themselves are presented within brackets so that they are easily
distinguished from other results in the table.

**Presenting confidence intervals in text.** The Publication Manual of the American
Psychological Association 6th ed. (2010) recommends that brackets be used to
enclose confidence intervals when they are presented within the text of an article.
Below is an example of how Dr. O’Day might summarize one of the statistically
significant results from her caffeine study (notice that the lower limit of the confidence interval is separated from the upper limit by a comma):

One of the study’s dependent variables was self-reported irritability. For this variable, the mean score displayed in the high-caffeine condition ($M = 575.00, SD = 35.49$) was higher than the mean displayed in the no-caffeine condition ($M = 544.73, SD = 36.13$). The observed difference between means was $30.27$, 95% CI $[18.71, 41.82]$, and this difference was statistically significant, $t(148) = 5.18$, $p < .001$.

In the same study, there was a nonsignificant difference between conditions on the criterion variable labeled other-reported irritability. Below is an example of how Dr. O’Day might have described this in the text of an article:

One variable in the investigation was other-reported irritability. For this variable, the mean score displayed in the high-caffeine condition ($M = 556.46, SD = 35.86$) was only slightly higher than the mean displayed in the no-caffeine condition ($M = 546.27, SD = 36.10$). The observed difference between means was $10.19$, 95% CI $[-1.42, 21.80]$, and this difference was not statistically significant, $t(148) = 1.73$, $p = .085$.

**Final Thoughts on Confidence Intervals**

Remember: when you think confidence interval, think precision. A narrow confidence interval tells you that the estimate is fairly precise, while a wide confidence interval tells you that it is relatively imprecise.

**Big-Three Results, Part 3: Effect Size**

When you think effect size, you should think strength. An index of effect size is a number that represents the strength of the relationship between predictor variables and criterion variables. With most indices, higher values (in absolute terms) indicate a stronger association between the predictor variables and the criterion variables.

Statisticians and researchers have developed many different indices of effect size. Examples include Cohen’s $d$ statistic, eta squared ($\eta^2$), partial eta squared ($\eta_p^2$), omega squared ($\omega^2$), and the $r^2$ and $R^2$ indices from regression. These indices differ with respect to (a) the types of data sets for which they are appropriate, and (b) the information that they provide.

**Effect Size: Basic Concepts**

Indices of effect size are important because they answer some of the questions that are of greatest interest to researchers:

- In this experiment, did the independent variable have a powerful effect on the dependent variable?
• In this correlational study, how strong was the relationship between these variables?

• If I use scores on this X variable to predict scores on this Y variable, to what extent will it decrease my errors of prediction?

In addition, indices of effect size may be less likely to be misinterpreted, compared to the null-hypothesis significance tests that have traditionally been emphasized in research. For example, an earlier section of this chapter pointed out that, when a study includes a large number of participants, the analysis might indicate that there is a statistically significant difference between two groups even when the actual size of the difference is quite small, and that this “statistically significant” outcome can mislead readers into believing that the difference is substantial when it is actually quite trivial. Because indices of effect size are not influenced by sample size in the way that significance tests are, many researchers feel that they are less likely to be misinterpreted. Because of this and other advantages, the Publication Manual of the American Psychological Association 6th ed. (2010) now strongly recommends that researchers should report effect size as part of their results.

A CAVEAT REGARDING THE WORD “EFFECT.” Do not be misled by the word “effect,” as it is used in this book. If a researcher has conducted a true experiment with strong internal validity, then an index of effect size may indeed be interpreted as the size of the causal effect that the independent variable has had on the dependent variable—the larger the value, the stronger the causal effect. However, if the study was a nonexperimental study, then the index of effect size should merely be interpreted as indicating the strength of the association (i.e., the correlation) between the predictor variable and the criterion variable. Although this book may use the term “effect size” with respect to correlational studies as well as experimental studies, remember that this term merely refers to the strength of the association between variables, not necessarily the size of a causal effect.

TWO APPROACHES TO COMPUTING EFFECT SIZE. An earlier section has already indicated that there are many different indices of effect size that are regularly reported in research articles: $d$, $\eta^2$, $\omega^2$, $r^2$, $R^2$, and others. Most of these indices may be classified as falling into one of two categories. The first category includes the standardized difference indices, for which Cohen’s $d$ statistic is probably the best-known example. Rosenthal (1994) refers to this category as the $d$ family of indices. The second category includes the variance accounted for indices, such as $r^2$, $R^2$, and eta-squared ($\eta^2$). Rosenthal (1994) refers to this as the $r$ family. The next two sections discuss these two categories in more detail.

Standardized-Difference Indices of Effect Size

A standardized difference index of effect size indicates the size of the difference between two means, as measured in standard deviations. Examples include Cohen’s $d$ statistic, Hedges’ $g$ statistic, and Glass’s $\Delta$ statistic (Rosenthal, 1991; 1994; 1995).
**Example: Cohen’s d statistic.** The precise way that a standardized difference index is computed depends on the nature and purpose of the investigation. However, most standardized difference indices are computed by (a) subtracting the mean obtained in one treatment condition from the mean obtained in a different treatment condition, and (b) dividing this difference by an estimate of the population standard deviation.

A good example of a standardized-difference index of effect size is Cohen’s d statistic (Cohen, 1988). For a study with two independent samples (such as the caffeine study, described above), the d statistic is computed as:

\[
d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}
\]

With this formula, \(d = \) Cohen’s d statistic; \(\bar{X}_1\) = the sample mean for one group; \(\bar{X}_2\) = the sample mean for the second group; and \(s_p\) = pooled estimate of the population standard deviation (this is also called the within-group standard deviation, symbol: \(s_{\text{within}}\)).

**Labeling an effect as small, medium, or large.** The late Jacob Cohen did much to educate researchers on the importance of reporting effect size in research articles. He provided the following advice to researchers interested in interpreting the size of the effect obtained in a given study (e.g., Cohen, 1988, p. 12, 25):

1) In general, researchers should review theory and previous research findings that are relevant to the current study, and interpret the index of effect size obtained in the current study by comparing it to the indices observed in these previous investigations.

2) When previous research does not provide sufficient guidance along these lines, researchers can compare the size of the obtained effect against Cohen’s own operational definitions for small, medium, and large effects.

As an illustration, Cohen’s criteria for evaluating his d statistic are presented in Table 5.6.

To help researchers put these criteria in perspective, Cohen provided several real-world examples of differences that would be classified as being small, medium, or large according to this system. One example involves measurements of the height of girls, where the population standard deviation is about \(\sigma = 2.1\) inches. Using this data set, Cohen (1988) offered the following examples:

- An example of a small effect would be the difference between the height of the average 16-year-old girl versus the average 15-year-old girl (\(d = 0.24\)). Cohen indicated that a small effect should be a difference that is “difficult to detect” (Cohen, 1988, p. 25).

- An example of a medium effect would be the difference between the height of the average 18-year-old girl versus the average 14-year-old girl (\(d = 0.48\)). He indicated that a medium effect should be “one large enough to be visible to the naked eye” (Cohen, 1988, p. 25).
• An example of a *large effect* would be the difference between the height of the average 18-year-old girl versus the average 13-year-old girl \((d = 0.81)\). He felt that a large effect should be a difference that is “grossly perceptible” (Cohen, 1988, p. 27).

Table 5.6

*Cohen’s (1988) Criteria for Interpreting the Size of the d Statistic*

<table>
<thead>
<tr>
<th>Size of the effect</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>±.20</td>
</tr>
<tr>
<td>Medium</td>
<td>±.50</td>
</tr>
<tr>
<td>Large</td>
<td>±.80</td>
</tr>
</tbody>
</table>

**HEDGE’S \(g\) STATISTIC.** In small samples, Cohen’s \(d\) statistic tends to overestimate the actual effect size in the population. Multiplying Cohen’s \(d\) statistic by a correction factor produces Hedge’s \(g\) statistic: an *unbiased* estimate of the standardized difference in the population.

The correction causes Hedges’ \(g\) to be slightly smaller than Cohen’s \(d\) in absolute value. However, unless the sample is very small, the correction is slight and results in very little change in the size of Cohen’s \(d\) (Borenstein, Hedges, Higgins, & Rothstein, 2009).

**“Variance Accounted for” Indices of Effect Size**

A *variance accounted for* index of effect size is a statistic that communicates the percent of variance in the criterion variable(s) that is accounted for by the predictor variable(s). Examples include the squared Pearson correlation coefficient \((r^2)\), the squared point-biserial correlation coefficient \((r_{pb}^2)\), the squared multiple correlation coefficient \((R^2)\), eta-squared \((\eta^2)\), and omega-squared \((\omega^2)\), among others. The following sections use the squared point-biserial correlation coefficient as an example of a variance-accounted-for index.

**Example: The Point-Biserial Correlation.** The point-biserial correlation coefficient (symbol: \(r_{pb}\)) is used to investigate the relationship between a dichotomous variable and a continuous quantitative variable. Under ideal conditions, values of \(r_{pb}\) may range from –1.00 through 0.00 through +1.00, with values closer to zero indicating weaker relationships.
This is illustrated in Table 5.7, which presents results from Dr. O’Day’s caffeine study. The column headed $r_{pb}$ presents the point-biserial correlations between the dichotomous independent variable (“high-caffeine” group versus “no-caffeine” group) and each of the four quantitative dependent variables (self-reported irritability, other-reported irritability, etc.). A separate correlation coefficient is reported for each variable.

Table 5.7

Scores on Four Measures of Irritability as a Function of Amount of Caffeine Consumed (High-Caffeine versus No-Caffeine): Means, t Tests, and Point-Biserial Correlations

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Means</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High-</td>
<td>No-</td>
<td>$t(148)$</td>
<td>$\rho$</td>
<td>$r_{pb}$</td>
<td>$r_{pb}^2$</td>
</tr>
<tr>
<td></td>
<td>caffeine</td>
<td>caffeine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-reported irritability</td>
<td>575.00</td>
<td>544.73</td>
<td>5.18</td>
<td>&lt;.001</td>
<td>.39</td>
<td>.15</td>
</tr>
<tr>
<td>Other-reported irritability</td>
<td>556.46</td>
<td>546.27</td>
<td>1.73</td>
<td>.085</td>
<td>.14</td>
<td>.02</td>
</tr>
<tr>
<td>Self-reported outbursts</td>
<td>6.24</td>
<td>5.03</td>
<td>2.77</td>
<td>.006</td>
<td>.22</td>
<td>.05</td>
</tr>
<tr>
<td>Other-reported outbursts</td>
<td>5.24</td>
<td>4.99</td>
<td>0.58</td>
<td>.561</td>
<td>.05</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note. $N = 150$; $r_{pb}$ = point-biserial correlation between the dichotomous independent variable (amount of caffeine consumed) and quantitative dependent variable.

When the point-biserial correlation coefficient is squared, it indicates the percent of variance in the continuous criterion variable that is accounted for by the dichotomous predictor variable. When the squared point-biserial correlation is equal to zero, it means that there is no relationship between the two variables; larger absolute values indicate stronger relationships.

In Table 5.7, the column headed $r_{pb}^2$ contains the squared point-biserial correlation for each of the four criterion variables. The values in this column were computed by squaring the correlation coefficients in the $r_{pb}$ column. Given the relative size of the $r_{pb}^2$ values, it appears that the amount of caffeine consumed had the strongest effect on self-reported irritability ($r_{pb}^2 = .15$), and the weakest effect on other-reported outbursts ($r_{pb}^2 = .00$).
PERCENT IMPROVEMENT IN PREDICTIVE ACCURACY. The numerical value of a variance accounted for index communicates the proportional improvement in accuracy that is attained by capitalizing on the observed relationship between the predictor variable and the criterion variable (Heiman, 2006). This value will range from 0% (when the predictor does a poor job of predicting scores on the criterion variable) to 100% (when the predictor does a perfect job of predicting scores on the criterion).

For example, consider the current study in which one criterion variable was self-reported irritability, and the predictor variable was the amount of caffeine consumed (high-caffeine versus no-caffeine). At the end of the investigation, the combined mean on self-reported irritability was $M_{combined} = 559.87$ (here, combined mean indicates the mean based on all 150 subjects, making no distinction between the high-caffeine sample versus the no-caffeine sample).

Imagine that you now wish to predict the score that was displayed by each of the 150 participants in the study. That is, for Participant 1, you must predict where he or she scored on self-rated irritability; for Participant 2, you must predict where he or she scored, and so forth. Imagine further that you must make these predictions under two different circumstances: Situation 1 and Situation 2. These circumstances are described below.

**SITUATION 1.** In Situation 1, you have no idea whether a given subject was in the high-caffeine condition or the no-caffeine condition. Given this limitation, if you wish to minimize the average amount of error that is associated with the predictions that you are making, your “best guess” for each of 150 subjects would be the combined mean: The overall mean based on all 150 subjects. An earlier paragraph indicated that this combined mean was $M_{combined} = 559.87$. This indicates that, for Participant 1, you will guess that this subject’s score on the criterion variable was 559.87. For Participant 2, you will guess that this subject’s score was also 559.87. In this way, you will guess each subject’s score on the criterion variable, and your guess will be exactly the same value (559.87) for each person.

If you use the combined mean of 559.87 as your guess for each participant’s score, your predictions will display a good deal of error. For many of your subjects, your guess of 559.87 will be too low and with other subjects, your guess of 559.87 will be too high.

**SITUATION 2.** Imagine that, in Situation 2, your circumstances have now changed: Imagine that—for each subject—you have now been told whether that participant was in the high-caffeine condition or the no-caffeine condition. Further, imagine that you now know that the mean irritability score for the high-caffeine (experimental) condition was $M_e = 575.00$, and the mean irritability score for the no-caffeine (control) condition was $M_c = 544.73$. Once again, for each participant you are required to guess that participant’s score on the criterion variable (self-reported irritability). However, now that you know which condition each participant was in, it is no longer wise for you to use the combined mean score ($M_{combined} = 559.87$) as you are making your guesses. Instead, for each subject, you will now use the mean score of the treatment condition that the subject was in: That is, for each of the 75 subjects who were in the high-caffeine condition, your
guess will be 575.00 (because the mean score for the high-caffeine condition was $M_h = 575.00$). Similarly, for each of the 75 subjects who were in the no-caffeine condition, your guess will be 544.73 (because the mean score for the no-caffeine condition was $M_c = 544.73$).

It makes sense that the accuracy of your predictions should be better in Situation 2 than they had been in Situation 1. This is because (a) in Situation 2, you are using the mean score of the subject’s treatment condition as your best guess for that subject, and (b) we know that there was a significant difference between the two treatment conditions with respect to their scores on self-reported irritability. That is, we have already established that the high-caffeine condition scored significantly higher on self-reported irritability, compared to the no-caffeine condition.

So your predictions in Situation 2 will be more accurate than your predictions from Situation 1, but exactly how much more accurate will they be? In this case, we know that you have increased the accuracy of your predictions by 15%. Why 15%? Because the squared point-biserial correlation between this study’s predictor variable (amount of caffeine consumed) and the criterion variable (self-reported irritability) was $r_{pb}^2 = .15$. This was shown in Table 5.7. The squared point-biserial correlation tells us the percent improvement in predictive accuracy that is achieved by capitalizing on the observed relationship between the predictor variable and the criterion variable. In this case, you “capitalized on the relationship” by using the mean of the condition that a subject was in as your best guess for that subject (rather than the combined mean based on all 150 participants).

If there had been a larger difference between the mean scores displayed by the two treatment conditions, the squared point-biserial correlation would have been a larger value, perhaps $r_{pb}^2 = .25$. If this had been the case, you would have increased the accuracy of your predictions by 25% by capitalizing on the relationship between the predictor and the criterion.

By the same reasoning, if there had been no difference between the mean scores for the two conditions, the squared point-biserial correlation would have been $r_{pb}^2 = .00$. In this case, capitalizing on the relationship between the two variables would not have increased the accuracy of your prediction at all.

**OTHER VARIANCE-ACCOUNTED-FOR INDICES.** This section has focused on just one “variance-accounted-for” index of effect size: the squared point-biserial correlation coefficient. However, it is important to remember that the basic concepts discussed here apply with the other variance-accounted-for indices listed at the beginning of this section: the squared Pearson correlation coefficient ($r^2$), the squared multiple correlation coefficient ($R^2$), eta-squared ($\eta^2$), and omega-squared ($\omega^2$).

**LABELING AN EFFECT AS SMALL, MEDIUM, OR LARGE.** An earlier section reviewed Cohen’s criteria for evaluating standardized-difference indices of effect size (such as the $d$ statistic) to determine whether they should be labeled as representing a small,
medium, or large effect. In the same book, Cohen (1988) also provides a different set of criteria for evaluating various types of correlation coefficients. For example, Table 5.8 presents criteria for evaluating the Pearson correlation coefficient and the point-biserial correlation coefficient (Cohen, 1988; Kline, 2004). Notice that these criteria apply to correlation coefficients that have not been squared.

Table 5.8

<table>
<thead>
<tr>
<th>Size of r (the Pearson Correlation Coefficient) and rpb (the Point-Biserial Correlation Coefficient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
</tr>
<tr>
<td>±.10</td>
</tr>
<tr>
<td>±.30</td>
</tr>
<tr>
<td>±.50</td>
</tr>
</tbody>
</table>

Interpreting the Statistics Other than d or r

This section has provided guidelines for interpreting the relative size of the d statistic, the r statistic, and the rpb statistic. But what if you are reading about a study that did not produce a simple statistic such as d or r? What if the study reports an F statistic (as is obtained with analysis of variance) or some other statistic? Is it still possible to interpret the size of the effect?

The answer is yes. First, Cohen’s (1988) oft-cited book on power analysis is the most obvious reference for interpreting effect size based on a wide variety of statistics. In addition, some statistics textbooks provide guidelines (often adapted from Cohen, 1988) that can be used to interpret statistics from more advanced research designs. For example, Newton and Rudestam (1999) provide guidelines for interpreting the size of $R^2$ (from multiple regression) and eta-squared (from ANOVA), along with other statistics (see, for example, the table on page 76 of their book).

In addition, a number of publications provide formulas that can be used to convert certain statistics (such as F) into the r statistic, the d statistic, or some other common metric. See, for example, Hedges and Becker (1986) and Rosnow and Rosenthal (1996). Books on meta-analysis are particularly rich sources for these conversion formulas (e.g., Borenstein et al., 2009; Hedges & Olkin, 1985; Hunter & Schmidt, 2004).
But it may not be necessary to turn to other references at all. Most of the chapters in this book (the one that you are now reading) deal with specific statistical procedures, and many of these chapters discuss the indices of effect size that are typically used for that procedure. Wherever possible, these chapters also provide guidelines for labeling a given effect as small, medium, or large.

**Final Thoughts on Effect Size**

Remember: When you think *effect size*, think *strength*. An index of effect size communicates the strength of the relationship between predictor variables and criterion variables. When research articles report indices of effect size, it decreases the likelihood that readers will be misled by the results: Although a researcher might describe a result from a large-sample study as being “highly-significant” according to a significance test, you will be able to use the index of effect size to evaluate the actual strength of the relationship between the variables in question.

**Chapter Conclusion**

The big-three results covered in this chapter are important, both individually and collectively. Individually, they each provide a different type of information regarding the investigation’s findings: the significance test provides information about probability, the confidence intervals provide information about the precision of the estimates, and the indices of effect size provide information about the strength of the relationships. Collectively, they allow us to see the big picture regarding the statistical and practical significance of the results. When conducting your own research, you should report all three whenever feasible.

When reading about the empirical research conducted by others, however, don’t be shocked to see that they have omitted one or more of the big-three. In some cases, this is appropriate, as it is not possible to compute all three types of results with some statistical procedures. In other cases, however, it is likely that the researchers just didn’t know how to compute all three results. This is particularly likely with older articles published prior to the development of statistical applications such as SPSS or SAS.

The vast majority of empirical research articles will present null-hypothesis significance tests. Despite the current controversy, most graduate programs continue to teach significance tests, and most journals continue to require them.

Depending on the type of journal you are reading, you may have only a 50-50 chance of encountering confidence intervals. Confidence intervals provide a lot of information: They not only tell you whether the results are statistically significant, but also convey the precision of the sample-based estimates. Because of this, the *Publication Manual of the American Psychological Association 6th ed.* (2010) strongly endorses their use. However, graduate programs have not emphasized confidence intervals to the extent that they have emphasized significance testing, and many data-analysis applications do not compute confidence intervals for many statistics.
Indices of effect size occupy a position somewhere between significance tests and confidence intervals. Popular statistical applications compute indices of effect size for many statistics, and some journals have required them for decades. As a result, you are fairly likely to encounter indices such as $R^2$ and $\eta_p^2$ in your reading, and the popularity of these indices is only likely to increase.

All of this is in transition. With the passage of time, graduate programs may increasingly emphasize confidence intervals and effect size, and statistical applications may make it easier to compute these indices for a wide variety of statistical procedures. If this occurs, tomorrow’s researchers are likely to be as comfortable with the big-three results as today’s researchers are with simple significance tests.