8 Independent-Samples *t* Test

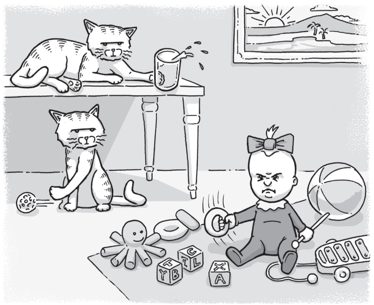
LEARNING OBJECTIVES

- Differentiate between independent samples and paired samples.
- Conduct the steps for an independent-samples *t* test.
- Interpret an independent-samples *t* test.

CHAPTER OVERVIEW

Chapters 6 and 7 covered the single-sample *z* test and the single-sample *t* test. Both are difference tests, used to compare the mean of a sample to the mean of a population. They were the first tests covered because they are good for introducing the logic of hypothesis testing and the six-step procedure for completing a hypothesis test. But researchers rarely ask whether this sample mean differs from that population mean.

This chapter introduces a test, the independent-samples *t* test, that researchers do use regularly. The independent-samples *t* test is a two-sample difference test. This is just what it sounds like—a test that is used to see if the average score in one population is better or worse than the average score in a second population. Why is it commonly used? Because classic experiments involve two groups—an experimental group and a control group. Nothing is done to the control group, something is done to the experimental group, and then the outcomes of the two groups are compared to see if the independent variable (the "something") had an effect on the dependent variable.



Tom and Harry despise crabby infants

8.1 Types of Two-Sample *t* Tests

Two-sample *t* tests are used to compare the mean of one population to the mean of another population. To conduct a two-sample *t* test, a researcher needs two samples of cases and an interval- or ratio-level dependent variable. Many experiments in psychology follow this format. For example, a group of researchers was interested in seeing how stereotypes affect behavior (Bargh, Chen, & Burrows, 1996). They divided undergraduate students into two groups, a control group and an experimental group. Both groups were given words that they had to put into sentences. The control group received neutral words and the experimental group was given words like *Florida, wrinkle,* and *forgetful*—words that were meant to prime a stereotype of elderly people. When the participants thought the experiment was over, they exited through a hallway and the experimenters secretly timed how long each person took to walk down the hall. The experimenters believed that being primed with words that are associated with the elderly would lead participants to act like elderly people and walk more slowly.

The results showed that the control group walked down a 30-foot hall, on average, 1 second faster than the experimental group. One second does not sound like much of a difference, but when the experimenters analyzed the data using a two-sample *t* test, they found that the difference was a statistically significant one.

The difference meant that by the time an average person in the control group reached the end of the hall, an average person in the experimental group wasn't yet 90% of the way down the hall (Figure 8.1). The experimental group averaged 2.6 mph compared to 2.9 mph in the control group. Being primed with an elderly stereotype did lead participants to act more like elderly people and a two-sample *t* test helped the researchers reach this conclusion.

There are two different types of two-sample *t* tests, the independent-samples *t* test (the focus of this chapter) and the paired-samples *t* test (see Chapter 9). Obviously, they differ in terms of the samples being analyzed, whether the samples are independent or paired. What's the difference between *independent samples* and *paired samples*?

With **independent samples,** how cases are selected for one sample has no influence on (*is independent of*) the case selection for the other sample. Each sample could be a random sample from its population. With **paired samples,** often called dependent samples, the cases selected for one sample are connected to (*depend on*) the cases in the other sample. The cases in the samples are pairs of cases, yoked together in some way.

To clarify the difference between independent samples and paired samples, here is an example of two different ways to compare the intelligence of men and women to see which sex is smarter:

1. Dr. Smith obtains a random sample of men from the population of men in the world and a random sample of women from the population of women. Each person in the samples takes an IQ test, and Dr. Smith compares the mean of the men to the mean of the women.

2. Dr. Jones gets a random sample of married heterosexual couples. Each man and each woman takes an IQ test, and Dr. Jones compares the mean of the men to the mean of the women.

In the first example, the cases selected for one group have no influence on the cases selected for the other group, so that is an example of independent samples.

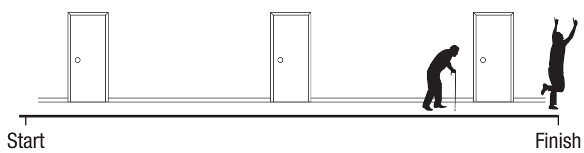


Figure 8.1 Results of a Study with a Control Group and an Experimental Group
By the time the average person in the neutral condition had walked the length of the hallway, the average person in the experimental condition had walked less than 90% of the same distance. Being primed with words that activate a stereotype of elderly people led participants to walk more slowly (Bargh, Chen, & Burrows, 1996).

In the second scenario, who the women are is dependent on the men, so that is an example of paired samples.

It is important to know if the two samples being analyzed are independent or paired so that the researcher can choose the right two-sample test. The example above is one way that subjects can be paired. Table 8.1 contains some guidelines for determining if subjects are paired.

TABLE 8.1 How to Choose: Guidelines for Determining if Samples Are Independent or Paired

Independent Samples	Paired Samples
If each sample is a random sample from its respective population *or* If $n_1 \neq n_2$, that is, each sample has a different number of cases.*	If samples consist of the same cases measured at more than one point in time or in more than one condition *or* If the selection of cases for one sample determines the selection of cases for another sample *or* If the cases in the samples are matched, yoked, or paired together in some way.

Note: These guidelines help decide if samples are independent or paired.
*For paired-samples the two sample sizes must be equal. As they can be equal for independent samples, having equal sample sizes doesn't provide any information about whether the samples are paired or independent.

Practice Problems 8.1

Apply Your Knowledge

8.01 A gym owner is offering a strength training course for women. Eight women have signed up. The gym owner measures how many pounds they can bench press at the first class. He plans to measure this again, 12 weeks later, at the last class, in order to see if their strength has changed. Should he use an independent-samples *t* test or a paired-samples *t* test?

8.02 A public health researcher wants to compare the rate of cigarette smoking for a sample of eastern states vs. a sample of western states. Should she use an independent-samples *t* test or a paired-samples *t* test?

8.03 A nutritionist wants to compare calories for meals at restaurants with tablecloths to restaurants without tablecloths. He gets a sample of each type of restaurant and finds out the calorie count for the most popular meal at each restaurant. Should he use an independent-samples *t* test or a paired-samples *t* test?

8.2 Calculating the Independent-Samples *t* Test

The two-sample *t* test covered in this chapter is the independent-samples *t* test. The **independent-samples *t* test** is used when seeing if there is a difference between the means of two independent populations. For our example, let's imagine following Dr. Villanova as he replicates part of a classic experiment about factors that influence how well one remembers information (Craik & Tulving, 1975).

Dr. Villanova's participants were 38 introductory psychology students, randomly assigned to two groups. He ended up with 18 in the control group (n_1) and 20 in the experimental group (n_2). The lowercase n is a new abbreviation that will be used to indicate the sample size for a specific group. Subscripts are used to distinguish the two groups. Here, one might say $n_1 = 18$ and $n_2 = 20$ or $n_{\text{Control}} = 18$ and $n_{\text{Experimental}} = 20$. An uppercase N will still be used to indicate the total sample size. For this experiment, $N = n_1 + n_2 = 18 + 20 = 38$.

Each participant was tested individually. All participants were shown 20 words (e.g., *giraffe, DOG, mirror*), one at a time, and asked a question about each word. The control group was asked whether the word appeared in capital letters or not. The experimental group was asked whether the word would make sense in this sentence, "The passenger carried a _______ onto the bus." The first question doesn't require much thought, so the control group was called the "shallow" processing group. Answering the second question required more mental effort, so the experimental group was called the "deep" processing group. After all 20 words had been presented, the participants were asked to write down as many words as they could remember. This recall task was unexpected and the number of words recalled was the dependent variable. The shallow processing group recalled a mean of 3.50 words ($s = 1.54$ words), and the deep processing group a mean of 8.30 words ($s = 2.74$ words). Here is Dr. Villanova's research question: Does deep processing lead to better recall than shallow processing?

Box-and-whisker plots are an excellent way to provide a visual comparison of two or more groups. Box-and-whisker plots are data-rich, providing information about central tendency and variability. In **Figure 8.2**, the number of words recalled is the dependent variable on the *Y*-axis and the two groups, shallow and deep, are on the *X*-axis. Both groups are in the same graph, making it easy to compare them on three things, the median, the interquartile range, and the range:

- The line in the middle of the box is the median, a measure of central tendency. The median number of words recalled is around 8 for the deep processing group and around 4 for the shallow processing group.
- The box that surrounds the line for the median represents the interquartile range, the range within which the middle 50% of scores fall. These middle 50% of scores are often referred to as the average scores for a group, so the interquartile range is a measure of central tendency. Note that in this study, no overlap occurs between these average scores for the two groups.
- The interquartile range is also a measure of variability, measured by the distance from the bottom to the top of the box. Taller boxes mean more variability. The box for the deep processing group looks a little taller than the box for the shallow processing group, indicating more variability for the experimental group than for the control group.
- The whiskers that extend from the box represent the range of scores, another measure of variability. One can see that participants in the shallow processing group recalled from 1 to 6 words and those in the deep processing group recalled from 4 to 12 words. The whiskers show that there is some overlap in the number of words recalled between the two groups and that there is more variability in the deep processing group than in the shallow processing group.

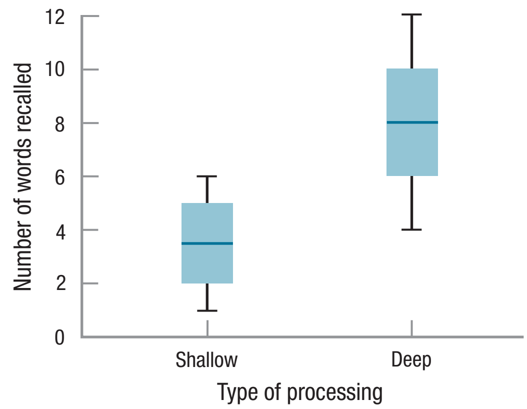


Figure 8.2 Box-and-Whisker Plots Showing the Results of the Depth of Processing Study Note that the middle 50% of cases, represented by the box for the deep processors, recalled more words than the average cases for the shallow processors. This graph appears to show that deep processing leads to better recall than shallow processing. To make sure that the result is a statistically significant one, a researcher would need to complete the appropriate hypothesis test, an independent-samples *t* test.

It certainly appears as if deeper processing leads to better memory, and this was the result that Craik and Tulving found back in 1975. But, in order to see if he replicated their results, Dr. Villanova will need to complete a hypothesis test to see if the difference is a statistically significant one.

Step 1 Pick a Test. The first step in completing a hypothesis test is choosing the correct test. Here, Dr. Villanova is comparing the mean of one population to the mean of another population, so he will use a two-sample *t* test. The two sample sizes are different ($n_{\text{Shallow}} = 18$ and $n_{\text{Deep}} = 20$), which means they are independent samples (see Table 8.1). In addition, the cases in one sample aren't paired with the cases in the other sample, another indicator of independent samples. As the samples are independent, his planned test is the independent-samples *t* test.

Step 2 Check the Assumptions. Several of the assumptions for the independent-samples *t* test, listed in Table 8.2, are familiar from the single-sample *t* test, but one is new.

TABLE 8.2 Assumptions for the Independent-Samples *t* Test

	Assumption	Robustness
1. Random samples	Each sample is a random sample from its population.	Robust
2. Independence of observations	Cases within a group are not influenced by other cases within the group.	Not robust
3. Normality	The dependent variable is normally distributed in each population.	Robust, especially if *N* is large
4. Homogeneity of variance	Degree of variability in the two populations is equivalent.	Robust, especially if *N* is large

Note: N is considered large if it is 50 or greater.

The first assumption, random samples, is one seen before. For the depth of processing study, the participants are intro psych students, specifically ones who volunteered to participate in this study. Dr. Villanova may have used random *assignment* to place the participants into a control group and an experimental group, but he doesn't have a random *sample* of participants. Thus, this assumption was violated and this will have an impact on interpreting the results. However, such an assumption is robust, so he can proceed with the *t* test.

With participants being randomly assigned to groups, each case participating by him- or herself, and no one participating twice, the second assumption, that each participant's responding was not influenced by any other participant, was not violated. Dr. Villanova was willing to assume that the third assumption, normality, is not violated. He's willing to make this assumption because memory is a cognitive ability and most psychological variables like this are considered to be normally distributed. In addition, this assumption is robust as long as $N > 30$.[1]

The final assumption is new, homogeneity of variance. This fourth assumption says that the amounts of variability in the two populations should be about equal. Dr. Villanova will assess this by comparing the two sample standard deviations. If the larger standard deviation is not more than twice the smaller standard deviation, then the amounts of variability are considered about equal (Bartz, 1999). For the depth of processing study, the larger standard deviation is 2.74, and the smaller standard deviation is 1.54. The larger standard deviation is not more than twice the smaller standard deviation, so the fourth assumption has not been violated. Dr. Villanova can proceed with the planned independent-samples *t* test.

A Common Question

Q What's the point of the homogeneity of variance assumption? What can I do if it is violated?

A As a step in calculating the independent-samples *t* test, the two sample variances are averaged together. This only gives a meaningful result if the two variances are about the same, homogeneous, to begin with. If the only two siblings in a family are a 192-pound 15-year-old and an 8-pound newborn, it doesn't mean much to report that the average weight of kids in that family is 100 pounds.

If this assumption is violated, there is a way to reduce the degrees of freedom. This correction causes the critical value of *t* to fall further out in the rare zone, making it harder to reject the null hypothesis.

Step 3 List the Hypotheses. Nondirectional, or two-tailed, hypotheses are more common than directional, one-tailed hypotheses, so let's mention those first. Nondirectional hypotheses are used when there is no prediction about which group will have a higher score than the other. In this situation, the null hypothesis will be a negative statement about the two *populations* represented by the two samples. It will say that no difference exists between the mean of one population, μ_1, and the mean of the other population, μ_2.

[1] There are objective ways to assess the normality assumption, but they are beyond the scope of an introductory statistics book like this one. Those who go on to take more classes in statistics will learn these techniques and will be freed from having to assume that variables are normally distributed.

This doesn't mean that the two sample means, M_1 and M_2, will be exactly the same. But, the difference between the sample means should be small enough that it can be explained by sampling error if the two samples were drawn from the same population.

The alternative hypothesis will state that the two *population* means are different. When using a two-tailed test, the alternative hypothesis will just say that the two population means are different from each other; it won't state the direction of the difference. The implication of the alternative hypothesis is that the two *sample* means will be different enough that sampling error is not a likely explanation for the difference.

Written mathematically and using the names of the conditions, the hypotheses are

$$H_0: \mu_{\text{Shallow}} = \mu_{\text{Deep}}$$
$$H_1: \mu_{\text{Shallow}} \neq \mu_{\text{Deep}}$$

Hypotheses like these will always be the null and alternative hypotheses used for a two-tailed independent-samples *t* test.

Dr. Villanova, however, has directional hypotheses and is doing a one-tailed test. Why is he doing a one-tailed test? Because the original study by Craik and Tulving in 1975 found that deep processing worked better than shallow processing, and Dr. Villanova expects that to be the case in his replication. He has made a prediction about the direction of the results in advance of collecting data, so he can do a one-tailed test.

One-tailed tests require a little more thought on the experimenter's part in formulating the hypotheses. With one-tailed tests, it is easier to state the alternative hypothesis first. Dr. Villanova's theory is that deep processing leads to better recall and the alternative hypothesis should reflect this: $H_1: \mu_{\text{Deep}} > \mu_{\text{Shallow}}$.

Once the alternative hypothesis is stated, the null hypothesis is formed by making sure that the two hypotheses are all-inclusive and mutually exclusive. If the alternative hypothesis says that deep processing is better than shallow, then the null hypothesis has to say that shallow processing is as good as, or better than, deep processing. The null hypothesis for the one-tailed independent-samples *t* test would be $H_0: \mu_{\text{Deep}} \leq \mu_{\text{Shallow}}$.

Dr. Villanova, then, would state his hypotheses as

$$H_0: \mu_{\text{Deep}} \leq \mu_{\text{Shallow}}$$
$$H_1: \mu_{\text{Deep}} > \mu_{\text{Shallow}}$$

Step 4 **Set the Decision Rule.** The critical value of *t*, t_{cv}, is the border between the rare and common zone for the sampling distribution of *t*. t_{cv} is used to set the decision rule and decide whether to reject the null hypothesis for an independent-samples *t* test.

To find the critical value of *t*, Dr. Villanova will use Appendix Table 3. To use this table of critical values of *t*, he needs to know three things: (1) whether he is doing a one-tailed test or a two-tailed test, (2) what alpha level he wants to use, and (3) how many degrees of freedom the test has.

Dr. Villanova predicted in advance of collecting the data that the deep processing participants would do better than the shallow processing participants. And, because he had directional hypotheses, he has a one-tailed test. So, he will be looking at one-tailed values of t_{cv}.

By convention, most researchers are willing to have a 5% chance of making a Type I error and set alpha at .05. (Type I error occurs when one erroneously rejects the null hypothesis.) As Dr. Villanova is replicating previous work, he wants to make sure that if he rejects the null hypothesis, it should have been rejected. He wants to have only a 1% chance of Type I error, so he's setting $\alpha = .01$.

Finally, the degrees of freedom need to be determined. The formula for calculating degrees of freedom (df) for an independent-samples t test is given in Equation 8.1.

Equation 8.1 Formula for Calculating Degrees of Freedom (*df*) for an Independent-Samples *t* Test

$$df = N - 2$$

where df = degrees of freedom
N = total number of cases in the two groups

For the depth of processing study, Dr. Villanova calculates $df = 38 - 2 = 36$. Looking in Appendix Table 3 at the row where $df = 36$ and under the column where $\alpha = .01$, one-tailed, we find $t_{cv} = 2.434$.

The t distribution is symmetric, so this value could be either -2.434 or $+2.434$. As Dr. Villanova is doing a one-tailed test, he needs to decide whether his critical value of t will be a positive or a negative number. To do so requires knowing: (1) which mean will be subtracted from which in the numerator of Equation 8.1 when it is used to find the t value; and (2) what the sign of the difference *should* be. The expected sign of the difference is the sign to be associated with the critical value of t for a one-tailed test.

Dr. Villanova has decided that he'll subtract the shallow processing group's mean from the deep processing group's mean. That is, it will be $M_{Deep} - M_{Shallow}$. If his theory is correct, the deep processing group will remember more words than the shallow group, and the difference will have a positive sign. This means that the critical value of t will be positive: 2.434. This value separates the rare zone of the sampling distribution of t from the common zone. The decision rule is shown in **Figure 8.3,** with the

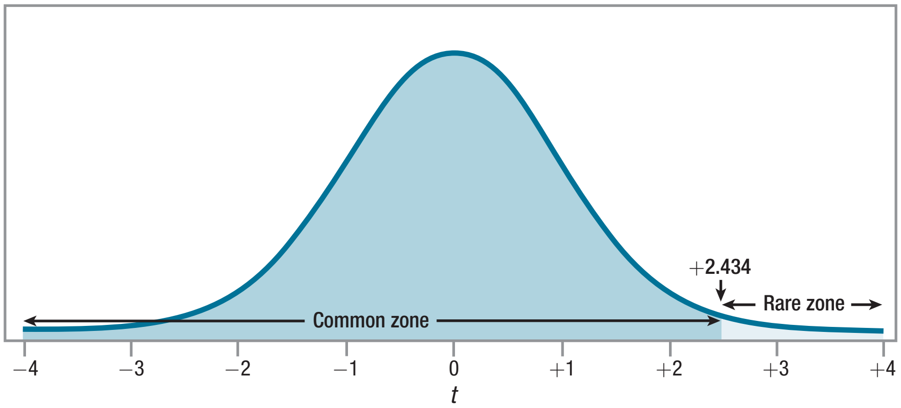


Figure 8.3 The Decision Rule for an Independent-Samples *t* Test, *df* = 36 For a one-tailed t test with the alpha set at .01 and 36 degrees of freedom, the critical value of t is 2.434.

rare and common zones labeled. Here is how Dr. Villanova states the decision rule for his one-tailed test:

- If $t \geq 2.434$, reject the null hypothesis.
- If $t < 2.434$, fail to reject the null hypothesis.

How would things be different if Dr. Villanova were doing a two-tailed test and had set alpha at .05? The degrees of freedom would still be 36, but t_{cv} would be 2.028 and the decision rule would be:

- If $t \leq -2.028$ *or* if $t \geq 2.028$, reject the null hypothesis.
- If $-2.028 < t < 2.028$, fail to reject the null hypothesis.

This critical value of *t*, 2.028, is closer to zero, the midpoint of the *t* distribution, than is the critical value of *t* that Dr. Villanova is actually using, 2.434. This means it would be easier to reject the null hypothesis if Dr. Villanova were using the two-tailed test with $\alpha = .05$ than with the one-tailed test set at .01 that he is actually using because the rare zone would be larger. He wanted to reduce the likelihood of Type I error, so he has achieved his goal. Table 8.3 summarizes the decision rules for one-tailed and two-tailed two-sample *t* tests.

Step 5 Calculate the Test Statistic. It is now time for Dr. Villanova to calculate the test statistic, *t*. In the last chapter, it was pointed out that the formula for a single-sample *t* test, $t = \frac{M - \mu}{S_M}$, was similar to the *z* score formula, $z = \frac{X - \mu}{\sigma}$. Though it won't look like it by the time we see it, the independent-samples *t* test formula is also like a *z* score formula. The numerator is a deviation score, the deviation of the difference between the two sample means from the difference between the two population means: $(M_1 - M_2) - (\mu_1 - \mu_2)$. Because hypothesis testing proceeds under the assumption that the null hypothesis is true, if we assume $\mu_1 = \mu_2$, we can simplify the numerator to $M_1 - M_2$.

The denominator is a standard deviation, the standard deviation of the sampling distribution of the difference scores. Standard deviations of sampling distributions are called standard errors and this one, called the standard error of the difference, is abbreviated $s_{M_1 - M_2}$. Calculating the standard error of the difference proceeds in two steps. First we need to calculate variance for the two samples combined, or pooled, and then we need to adjust this **pooled variance** to turn it into a standard error. Equation 8.2 gives the formula for calculating the pooled variance.

TABLE 8.3 Decision Rules for Independent-Samples *t* Tests

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

Note: *t* is the value of the test statistic, which is calculated in Step 5. t_{cv} is the critical value, which is found in Appendix Table 3. For a one-tailed test, the researcher needs to decide in advance whether the *t* value should be negative or positive in order to reject the null hypothesis.

Equation 8.2 Formula for Calculating the Pooled Variance for an Independent-Samples *t* Test

$$s^2_{\text{Pooled}} = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{df}$$

where s^2_{Pooled} = the pooled variance
n_1 = the sample size for Group (sample) 1
s_1^2 = the variance for Group 1
n_2 = the sample size for Group (sample) 2
s_2^2 = the variance for Group 2
df = the degrees of freedom ($N - 2$)

This equation says the pooled variance is calculated by multiplying each sample variance by 1 less than the number of cases that are in its sample and adding together these products. That sum is then divided by 2 less than the total number of cases, which is the same as the degrees of freedom.

Let's follow Dr. Villanova as he plugs in the values from his depth of processing study into Equation 8.2. Remember, the shallow processing group had 18 cases, with a standard deviation of 1.54; the sample size and standard deviation for the deep processing group were 20 and 2.74, respectively. The equation calls for variances, not standard deviations, but a variance is simply a squared standard deviation, so we are OK. And the equation calls for the degrees of freedom, which we calculated in the previous step as $N - 2 = 38 - 2 = 36$:

$$\begin{aligned}
s^2_{\text{Pooled}} &= \frac{s_1{}^2(n_1 - 1) + s_2{}^2(n_2 - 1)}{df} \\
&= \frac{1.54^2(18 - 1) + 2.74^2(20 - 1)}{36} \\
&= 2.3716(17) + 7.5076(19) \\
&= \frac{40.3172 + 142.6444}{36} \\
&= \frac{182.9616}{36} \\
&= 5.0823 \\
&= 5.08
\end{aligned}$$

The pooled variance is 5.08. We can now use the pooled variance to find the standard error of the mean by using Equation 8.3.

Equation 8.3 Formula for the Standard Error of the Mean, $s_{M_1 - M_2}$, for an Independent-Samples *t* test

$$s_{M_1 - M_2} = \sqrt{s^2_{\text{Pooled}}\left(\frac{N}{n_1 \times n_2}\right)}$$

where $s_{M_1 - M_2}$ = the standard error of the difference
s^2_{Pooled} = the pooled variance (from Equation 8.2)
N = the total number of cases
n_1 = the number of cases in Group 1
n_2 = the number of cases in Group 2

Corty_3e_Ch08_257-292_Final HR.indd 266 21/01/16 8:52 pm

This formula says that to find the standard error of the difference, one finds the quotient of the total sample size divided by the two individual sample sizes multiplied together. This quotient is multiplied by the pooled variance. Finally, in a step that is often overlooked, the square root of the product is found. In essence, this is similar to what was done in calculating the standard error of the mean, where the standard deviation was divided by the square root of *N*.

Let's do the math for the shallow vs. deep processing study:

$$\begin{aligned} s_{M_1-M_2} &= \sqrt{s^2_{\text{Pooled}}\left(\frac{N}{n_1 \times n_2}\right)} \\ &= \sqrt{5.08\left(\frac{38}{18 \times 20}\right)} \\ &= \sqrt{5.08\left(\frac{38}{360}\right)} \\ &= \sqrt{5.08 \times 0.1056} \\ &= \sqrt{0.5364} \\ &= .7324 \\ &= .73 \end{aligned}$$

At this point, the hardest part of calculating an independent-samples *t* test is over and Dr. Villanova knows that the standard error of the mean is 0.73. All that is left to do is use the formula in Equation 8.4 to find *t*.

Equation 8.4 Formula for an Independent-Samples *t* Test

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

where t = the independent-samples *t* test value

M_1 = the mean of Group (sample) 1

M_2 = the mean of Group (sample) 2

$s_{M_1-M_2}$ = the standard error of the difference (Equation 8.3)

Equation 8.4 says that an independent-samples *t* value is calculated by dividing the numerator, the difference between the two sample means, by the denominator, the standard error of the difference. Dr. Villanova, because he is doing a one-tailed test, has already decided that he will be subtracting the shallow processing group's mean, 3.50, from the deep processing group's mean, 8.30:

$$\begin{aligned} t &= \frac{M_1 - M_2}{s_{M_1-M_2}} \\ &= \frac{8.30 - 3.50}{0.73} \\ &= \frac{4.8000}{0.73} \\ &= 6.5753 \\ &= 6.58 \end{aligned}$$

The test statistic *t* that Dr. Villanova calculated for his independent-samples *t* test is 6.58 and Step 5 of the hypothesis test is over. We'll follow Dr. Villanova as he covers Step 6, interpretation, after working through the first five steps with another example.

Worked Example 8.1

For practice with an independent-samples *t* test, let's use an urban example. Dr. Risen, an environmental psychologist, wondered if temperature affected the pace of life. She went to Fifth Avenue in New York City, randomly selected pedestrians who were walking alone, timed how long it took them to walk a block, and converted this into miles per hour (mph).

She did this on two days, one a 20°F day in January and the other a 72°F day in June. Each time, she used the same day of the week and the same hour of the day. She also made sure that on both days there were blue skies and no obstructions, like snow or trash, on the sidewalk. On the cold day she timed 33 people, and on the warm day she timed 28. Her total sample size, *N*, was 61.

The results, displayed in Figure 8.4, show that people walked faster on the cold day ($M = 3.05$ mph, $s = 0.40$ mph) than on the warm day ($M = 2.90$ mph, $s = 0.39$ mph). The results suggest that people pick up their pace when it is cold outside. Let's follow Dr. Risen as she determines if the effect is a statistically significant one. To do so, she'll use the six steps of hypothesis testing. We'll follow her through the first five steps in this section and then tag along for the sixth step, interpretation, in the next part of this chapter.

Step 1 Pick a Test. Two groups of people, those walking on a cold day vs. those walking on a warm day, are being compared in terms of mean walking speed. This calls for a two-sample *t* test. Using Table 8.2, Dr. Risen concludes that the samples are independent—each sample is a random sample from its respective population, the cases in the two samples aren't paired, and the two sample sizes are different. Thus, the appropriate test is the independent-samples *t* test.

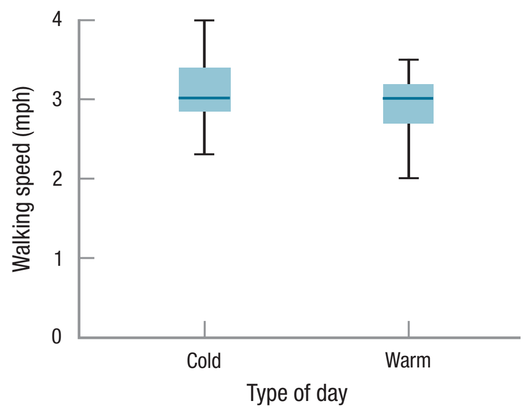


Figure 8.4 Box-and-Whisker Plots Showing the Effect of Temperature on Walking Time The middle 50% of cases (those in the box) look like they are walking a little faster on cold days than on warm days. In order to determine if the effect is a statistically significant one, however, a hypothesis test will need to be completed.

Step 2 Check the Assumptions.

- The random samples assumption is not violated as pedestrians were randomly selected.
- The independence of observations assumption is not violated. Only one case was observed at a time, each person was walking alone, and no person was timed twice. So, no other cases influenced a case.
- The normality assumption is not violated. Dr. Risen is willing to assume that a physical trait, like walking speed, is normally distributed. Plus, her sample size is large, greater than 50, and this assumption is robust if the sample size is large.
- The homogeneity of variance assumption is not violated. The two standard deviations are almost exactly the same ($s_{\text{Cold}} = 0.40$ mph and $s_{\text{Warm}} = 0.39$ mph).

Step 3 List the Hypotheses. Dr. Risen is doing an exploratory study. She's investigating whether temperature affects the pace of life. Hence, her test is two-tailed, and her hypotheses are nondirectional:

$$H_0: \mu_{\text{Cold}} = \mu_{\text{Warm}}$$

$$H_1: \mu_{\text{Cold}} \neq \mu_{\text{Warm}}$$

The null hypothesis says that there is no difference in walking speed in cold weather vs. walking speed in warm weather for the two populations. The alternative hypothesis says that the two population means—walking speed in cold weather vs. walking speed in warm weather—are different.

Step 4 Set the Decision Rule. To set the decision rule, Dr. Risen must find a critical value of t in Appendix Table 3. To do so, she needs three pieces of information: (1) whether the test is one-tailed or two-tailed, (2) what alpha level is selected, and (3) how many degrees of freedom there are.

1. The hypotheses were nondirectional, so she's doing a two-tailed test.
2. She's willing to run the standard, 5%, risk of making a Type I error, so $\alpha = .05$.
3. Applying Equation 8.1 to her total sample size, $N = 61$, she calculates degrees of freedom as $df = 61 - 2 = 59$.

Turning to the table of critical values of t, she looks for the intersection of the column for a two-tailed test with $\alpha = .05$, the bolded column, and the row for $df = 59$. However, there is no row for $df = 59$. What should she do? She'll follow *The Price Is Right* rule (see Chapter 7) and use the critical value found in the row for the degrees of freedom that are closest to her actual degrees of freedom without going over it. In this instance, that means the row with $df = 55$. The critical value of t is ± 2.004. The sampling distribution of t with this critical value of t is shown in Figure 8.5.

Here's her decision rule:

- If $t \leq -2.004$ *or* if $t \geq 2.004$, reject the null hypothesis.
- If $-2.004 < t < 2.004$, fail to reject the null hypothesis.

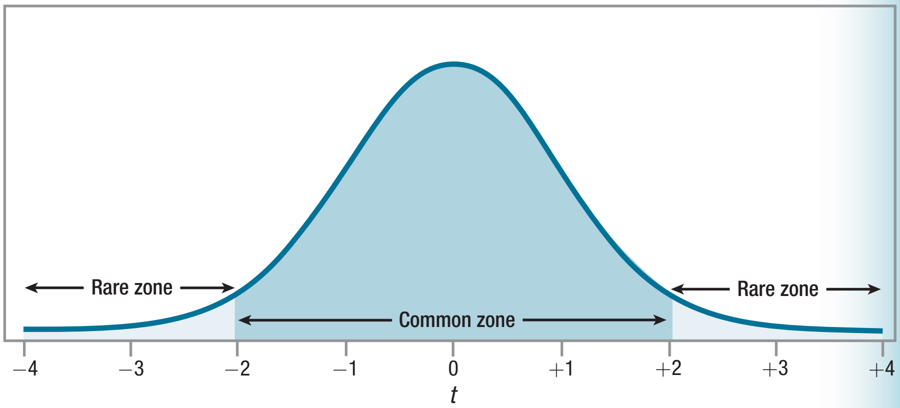


Figure 8.5 The Critical Value of *t* for the Walking Speed Study For a two-tailed *t* test with the alpha set at .05 and 55 degrees of freedom, the critical value of *t* is ±2.004. Note that the critical value associated with $df = 55$ is being used because this is the value closest to, without going over, the actual degrees of freedom of 59.

A Common Question

Q Why do statisticians use *The Price Is Right* rule and, if the actual degrees of freedom aren't in the table of critical values, use the value that is closest without going over the actual value?

A Statisticians like to make it difficult to reject the null hypothesis. Using a critical value associated with a smaller number of degrees of freedom means that the rare zone is smaller and so it is harder to reject the null hypothesis.

Step 5 Calculate the Test Statistic. The first step in finding the *t* value is to use Equation 8.2 to calculate the pooled variance. Once that is done, Dr. Risen can use Equation 8.3 to find the standard error of the difference and Equation 8.4 to calculate the value of the test statistic.

As a reminder, here is the information Dr. Risen has to work with:

- Cold day: $n = 33$, $M = 3.05$ (mph), $s = 0.40$ (mph)
- Warm day: $n = 28$, $M = 2.90$ (mph), $s = 0.39$ (mph)

Let's follow as Dr. Risen plugs the values into Equation 8.2 to find s^2_{Pooled}:

$$
\begin{aligned}
s^2_{\text{Pooled}} &= \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{df} \\
&= \frac{0.40^2(33 - 1) + 0.39^2(28 - 1)}{59} \\
&= \frac{0.1600(32) + 0.1521(27)}{59} \\
&= \frac{5.1200 + 4.1067}{59} \\
&= \frac{9.2267}{59} \\
&= 0.1564 \\
&= 0.16
\end{aligned}
$$

The next step is to use Equation 8.3 to find the standard error of the difference:

$$
\begin{aligned}
s_{M_1-M_2} &= \sqrt{s^2_{\text{Pooled}}\left(\frac{N}{n_1 \times n_2}\right)} \\
&= \sqrt{0.16\left(\frac{61}{33 \times 28}\right)} \\
&= \sqrt{0.16\left(\frac{61}{924.0000}\right)} \\
&= \sqrt{0.16 \times .0660} \\
&= \sqrt{0.0106} \\
&= 0.1030 \\
&= 0.10
\end{aligned}
$$

Now that she knows $s_{M_1-M_2} = 0.10$, Dr. Risen can go on to complete Equation 8.4 and find the *t* value:

$$
\begin{aligned}
t &= \frac{M_1 - M_2}{s_{M_1-M_2}} \\
&= \frac{3.05 - 2.90}{0.10} \\
&= \frac{0.1500}{0.10} \\
&= 1.5000 \\
&= 1.50
\end{aligned}
$$

The value of the test statistic, *t*, is 1.50. This completes Step 5.

A Common Question

Q Is it possible to calculate the standard error of the estimate without computing the pooled variance first?

A With a little algebraic rearranging, almost anything is possible. Here is the combination of Equations 8.2 and 8.3:

$$
s_{M_1-M_2} = \sqrt{\left[\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{df}\right]\left[\frac{N}{n_1 \times n_2}\right]}
$$

Practice Problems 8.2

Apply Your Knowledge

8.04 Previous research has shown that people who have served in the U.S. Armed Forces feel more patriotic about America. A researcher obtains a sample of veterans and a sample of nonveterans, and administers the interval level Sense of Patriotism Scale (SPS). Higher scores on the SPS indicate greater patriotism.

The researcher expects to replicate previous research. Write the researcher's null and alternative hypotheses.

8.05 If $n_1 = 12$, $s_1 = 4$, $n_2 = 16$, and $s_2 = 3$, calculate s^2_{Pooled} and $s_{M_1-M_2}$.

8.06 If $M_1 = 99$, $M_2 = 86$, and $s_{M_1-M_2} = 8.64$, calculate t.

8.3 Interpreting the Independent-Samples *t* Test

The final step in hypothesis testing, Step 6, is interpretation. This will follow the same format for the independent-samples *t* test as for previous hypothesis tests, addressing a series of questions:

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

The questions should be answered in order and each one adds additional information. After answering the first question, a researcher will have enough information for a basic interpretation. Answering all three questions, however, gives a deeper understanding of what the results mean and allows a researcher to write a more nuanced interpretation.

Was the Null Hypothesis Rejected?

The null and alternative hypotheses are set up to be all-inclusive and mutually exclusive. If a researcher can reject one hypothesis, he or she will have to accept the other hypothesis. If Dr. Villanova can reject the null hypothesis (that shallow processing is better than or equal to deep processing), he will be forced to accept the alternative hypothesis (that deep processing is better than shallow processing). With a one-tailed test, as this one is, if the null hypothesis is rejected, then the direction of the difference between the populations is known. With a two-tailed test, the researcher would need to examine the sample means in order to tell the direction of the probable population difference.

Here is what we know so far about Dr. Villanova's study on depth of processing:

- Shallow processing: $M = 3.50$, $s = 1.54$, $n = 18$
- Deep processing: $M = 8.30$, $s = 2.74$, $n = 20$
- $df = 36$
- $t_{cv} = 2.434$
- $s^2_{\text{Pooled}} = 5.08$
- $s_{M_1-M_2} = 0.73$
- $t = 6.58$

Dr. Villanova's first move is to plug the observed value of the test statistic, 6.58, into the decision rule generated in Step 4 and decide which statement is true:

- Is $6.58 \geq 2.434$?
- Is $6.58 < 2.434$?

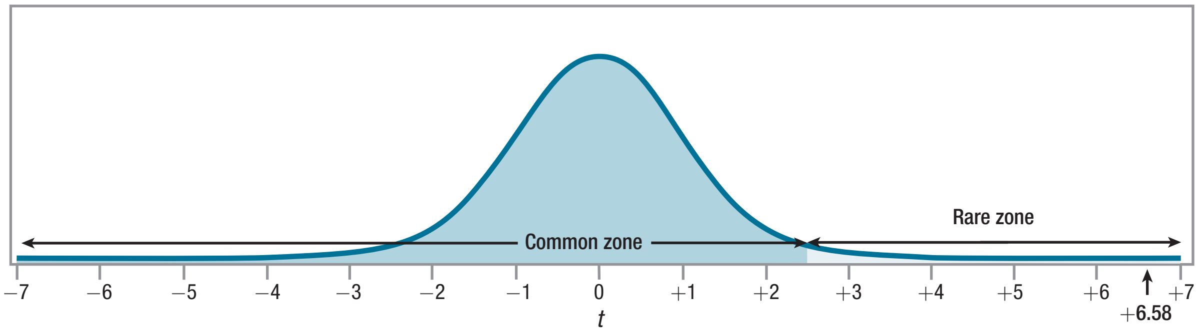


Figure 8.6 The Test Statistic for an Independent-Samples *t* Test for the Depth of Processing Data The test statistic *t*, 6.58, falls in the rare zone, so the null hypothesis is rejected. The alternative hypothesis is accepted. Because this is a one-tailed test, the alternative hypothesis states the direction of the difference.

6.58 is greater than or equal to 2.434, so the first statement is true and Dr. Villanova will reject the null hypothesis (see Figure 8.6) and call the results statistically significant. This means he accepts the alternative hypothesis that the mean number of words recalled by the population of people who use deep processing is greater than the mean number of words recalled by people who use shallow processing.

In APA format, Dr. Villanova would write the results as

$$t(36) = 6.58, p < .01 \text{ (one-tailed)}$$

- The *t* says that he is reporting the results of a *t* test.
- The 36 in parentheses, the degrees of freedom, gives information about how many cases were in the study. Because the degrees of freedom for an independent-samples *t* test is the total sample size minus 2, that means there were a total of 36 + 2, or 38, participants in the study.
- 6.58 is the value of the test statistic that was calculated.
- $p < .01$ indicates two things. The .01 tells that alpha was set at .01 because Dr. Villanova was willing to make a Type I error 1% of the time. The $p < .01$ reveals that the null hypothesis was rejected because the test statistic of 6.58 is a rare occurrence (it happens less than 1% of the time) when the null hypothesis is true.
- Finally, the phrase in parentheses at the end, one-tailed, tells the reader that Dr. Villanova conducted a one-tailed test. As most hypothesis tests are two-tailed, it is only when the test is not two-tailed that this fact is noted.

If Dr. Villanova chose to stop the interpretation at this point, he would have enough information to make a meaningful statement about the results. Here's what he could say:

> In a study comparing deep processing to shallow processing, a statistically significant effect was found [$t(36) = 6.58$, $p < .01$ (one-tailed)]. People who were randomly assigned to use deep processing recalled more words ($M = 8.30$) than did people who used shallow processing ($M = 3.50$).

How Big Is the Effect?

Cohen's *d*

Cohen's *d* and r^2, the same effect sizes used for the single-sample *t* test in Chapter 7, will be used for the independent-samples *t* test to tell how much impact the explanatory variable has on the outcome variable. The same standards can be used to judge the size of the effect for the independent-samples *t* test as were used for the single-sample *t* test:

- 0.00 means there is absolutely no effect.
- $d \approx 0.20$ or $r^2 = 1\%$ is a small effect.
- $d \approx 0.50$ or $r^2 = 9\%$ is a medium effect.
- $d \geq 0.80$ or $r^2 \geq 25\%$ is a large effect.

Equation 8.5 shows how to calculate Cohen's *d* for the independent-samples *t* test. Note that it makes use of the pooled variance, s^2_{Pooled}, which was 5.08.

Equation 8.5 Formula for Calculating Cohen's *d* for an Independent-Samples *t* Test

$$d = \frac{M_1 - M_2}{\sqrt{s^2_{\text{Pooled}}}}$$

where d = Cohen's *d* value
M_1 = the mean for Group (sample) 1
M_2 = the mean for Group (sample) 2
s^2_{Pooled} = the pooled variance (from Equation 8.2)

Here are Dr. Villanova's calculations for the effect size for the depth of processing study. He substitutes in 8.30 as the mean of the deep processing group, 3.50 as the mean of the shallow processing group, and 5.08 as the pooled variance:

$$\begin{aligned} d &= \frac{M_1 - M_2}{\sqrt{s^2_{\text{Pooled}}}} \\ &= \frac{8.30 - 3.50}{\sqrt{5.08}} \\ &= \frac{4.8000}{\sqrt{5.08}} \\ &= \frac{4.8000}{2.2539} \\ &= 2.1296 \\ &= 2.13 \end{aligned}$$

Cohen's *d* value, 2.13, is greater than 0.80, so Dr. Villanova can consider that the effect of the independent variable (type of processing) on the dependent variable (number of words recalled) is large. Now, in his interpretation, he can note more than the fact that deep processing leads to significantly better recall than shallow processing. He can say that the effect size is large, that how people process information *does* matter in how well they recall information.

r^2

The same formula, Equation 7.4, is used to calculate r^2 for the independent-samples *t* test as was used for the single-sample *t* test. It makes use of two values, *t* and *df*:

$$\begin{aligned} r^2 &= \frac{t^2}{t^2 + df} \times 100 \\ &= \frac{6.58^2}{6.58^2 + 36} \times 100 \\ &= \frac{43.2964}{43.2964 + 36} \times 100 \\ &= \frac{43.2964}{79.2964} \times 100 \\ &= 0.5460 \times 100 \\ &= 54.60\ \% \end{aligned}$$

r^2, remember, calculates the percentage of variability in the outcome variable that is accounted for by the explanatory variable. Here, r^2 tells how much of the variability in the number of words recalled is accounted for by the group, shallow vs. deep processing, subjects were assigned to. r^2 varies from 0% to 100%; the higher the percentage, the stronger the effect. Here, the effect is quite strong, with over 50% of the variability explained by group status.

How Wide Is the Confidence Interval?

To determine the impact of the independent variable on the dependent variable in the population, a confidence interval is used. For an independent-samples *t* test, a researcher calculates a confidence interval for the difference between population means, the same type of confidence interval calculated for the single-sample *t* test. This confidence interval estimates how close together (or how far apart) the two population means may be. This tells how much of an effect may, or may not, exist in the population.

Though any level of confidence, from greater than 0% to less than 100%, can be used for a confidence interval, the most commonly calculated is a 95% confidence interval. The formula for that is found in Equation 8.6. Two other common confidence intervals are 90% and 99%.

Equation 8.6 Formula for Calculating the 95% Confidence Interval for the Difference Between Population Means

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

where $95\%\text{CI}\mu_{\text{Diff}}$ = the 95% confidence interval for the difference between population means

M_1 = the mean of Group (sample) 1

M_2 = the mean of Group (sample) 2

t_{cv} = the critical value of *t*, two-tailed, $\alpha = .05$, $df = N - 2$ (Appendix Table 3)

$s_{M_1 - M_2}$ = the standard error of the difference (Equation 8.3)

For the depth of processing study, Dr. Villanova is going to calculate the 95% confidence interval. The two sample means are 8.30 and 3.50; the critical value of *t*, two-tailed, with $\alpha = .05$, and 36 degrees of freedom is 2.028; and the standard error of the difference is 0.73:

$$
\begin{aligned}
95\%\text{CI}\mu_{\text{Diff}} &= (M_1 - M_2) \pm (t_{cv} \times s_{M_1 - M_2}) \\
&= (8.30 - 3.50) \pm (2.028 \times 0.73) \\
&= 4.8000 \pm 1.4804 \\
&= \text{from } 3.3196 \text{ to } 6.2804 \\
&= [3.32, 6.28]
\end{aligned}
$$

The 95% confidence interval for the difference between population means ranges from 3.32 to 6.28. In APA format, it would be written as 95% CI [3.32, 6.28]. This confidence interval tells what the effect of the type of processing is on recall in the larger population. It says that the effect probably falls somewhere in the range from deep processing, leading to an average of anywhere from 3.32 to 6.28 more words being recalled over shallow processing.

Just as with the one-sample *t* test, there are three aspects of the confidence interval to pay attention to: (1) whether it captures zero; (2) how close it is to zero; and (3) how wide it is:

1. If the confidence interval captures zero, then it is plausible that no difference exists between the two population means. Thus, a confidence interval that captures zero occurs when the researcher has failed to reject the null hypothesis, as long as he or she is using a two-tailed test with $\alpha = .05$ *and* as long as he or she is calculating a 95% confidence interval.
2. When the confidence interval comes close to zero, then it is possible that there is little difference between the two population means. When it ranges farther away from zero, then it is possible that the difference between the two populations is more meaningful. In this way, a confidence interval is helpful in thinking about the size of the effect.
3. The width of the confidence interval tells how precisely a researcher can specify the effect in the population. A narrower confidence interval means the researcher can be reasonably certain of the size and meaningfulness of the difference. A wider confidence interval leaves the researcher uncertain of the size and meaningfulness of the difference. In such a situation, it is often reasonable to recommend replicating the study with a larger sample size in order to obtain more precision.

Dr. Villanova's confidence interval ranges from 3.32 to 6.28 for the depth of processing study. With regard to the three points above: (1) The confidence interval doesn't capture zero. It is unlikely that there is no difference between the two population means. Dr. Villanova expected this result as the null hypothesis had been rejected. (2) The low end of the confidence interval, the end that is closer to zero, is 3.32. In Dr. Villanova's opinion, a difference of 3.32 words is still a meaningful difference. Dr. Villanova, who planned and conducted this study, has expertise in this area and with this dependent variable. As a result, his opinion carries some weight. (3) The width of a confidence interval can be calculated by subtracting one side from the other:

$$6.28 - 3.32 = 2.96$$

The confidence interval is almost three words wide. In Dr. Villanova's opinion, based on his expertise, this is a reasonably narrow confidence interval. Thus, it provides a precise-enough estimate of what the population difference is. He feels little need to replicate with a larger sample size to obtain a narrower confidence interval.

A Common Question

Q How are d, r^2, and a confidence interval alike? How do they differ?

A d and r^2 are officially called effect sizes, but a confidence interval also gives information about how strong the effect of the explanatory variable is. d and r^2 reflect the size of the effect as observed in the actual sample; a confidence interval extrapolates the effect to the population. Cohen's d is not affected by sample size. If the group means and standard deviations stay the same but the sample size increases, d will be unchanged, but the confidence interval will narrow and offer a more precise estimate of the population value. r^2 is inversely affected by sample size—as N increases, r^2 decreases.

Putting It All Together

Dr. Villanova has addressed all three of the interpretation questions and is ready to use the information to write an interpretation that explains the results. In the interpretation, he addresses four points:

1. He starts with a brief explanation of the study.
2. He states the main results.
3. He explains what the results mean.
4. He makes suggestions for future research.

There's one more very important thing Dr. Villanova does. He did a lot of calculations—t, d, r^2, and a confidence interval—in order to understand the results. But, he doesn't feel obligated to report them all just because he calculated them. He limits what he reports in order to give a clear and concise report:

> This study compared how deep processing of words vs. shallow processing of words affected recall on an unexpected memory test. The deep processing group ($M = 8.30$ words, $s = 2.74$ words) recalled more words than the shallow processing group ($M = 3.50$ words, $s = 1.54$ words). This effect was statistically significant [$t(36) = 6.58$, $p < .01$ (one-tailed)] and it is a large effect. Using deep processing leads to markedly better recall when a person is not trying to memorize the words. Whether the effect exists when a person is purposefully trying to learn a list of words should be examined in a subsequent study.

Worked Example 8.2 For practice interpreting results for an independent-samples t test, let's return to Dr. Risen's study about how temperature affects the pace of city life. She measured the walking speed for 33 people walking alone on a cold day (20°F) and for 28 people on a warm day (72°F). Here is what is already known:

- For the cold day: $M = 3.05$ mph, $s = 0.40$
- For the warm day: $M = 2.90$ mph, $s = 0.39$
- $df = 59$
- t_{cv}, two-tailed, $\alpha = .05$ is ± 2.004
- $s^2_{\text{Pooled}} = 0.16$
- $s_{M_1 - M_2} = 0.10$
- $t = 1.50$

Was the null hypothesis rejected? The first step is applying the decision rule. Which is true?

- Is $1.50 \leq -2.004$ *or* is $1.50 \geq 2.004$?
- Is $-2.004 < 1.50 < 2.004$?

The second statement is true and the value of the test statistic falls in the common zone, as shown in Figure 8.7. Insufficient evidence exists to reject the null hypothesis, so there is no reason to conclude that temperature affects walking speed. The results are called "not statistically significant." In APA format, the results would be written like this:

$$t(59) = 1.50, p > .05$$

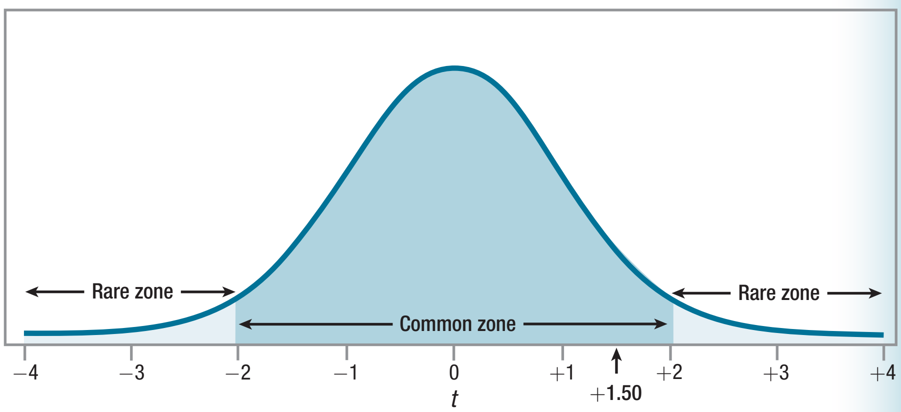


Figure 8.7 The Test Statistic *t* for the Walking Speed Study The test statistic *t*, 1.50, falls in the common zone, so the null hypothesis is not rejected. There is not enough evidence to conclude that the two population means differ.

A Common Question

Q When she looked up the critical value of *t* in Appendix Table 3, Dr. Risen had to use the row with 55 degrees of freedom because there was no row for $df = 59$. But, in reporting the results in APA format, she used $df = 59$, not $df = 55$. Why?

A The degrees of freedom within the parentheses in APA format provide information about how many cases there are, so they should reflect that.

How big is the effect? Using Equation 8.5, Dr. Risen calculated Cohen's *d:*

$$
\begin{aligned}
d &= \frac{M_1 - M_2}{\sqrt{s^2_{\text{Pooled}}}} \\
&= \frac{3.05 - 2.90}{\sqrt{0.16}} \\
&= \frac{0.1500}{\sqrt{0.16}} \\
&= \frac{0.1500}{0.4000} \\
&= 0.3750 \\
&= 0.38
\end{aligned}
$$

Equation 8.6 is used to calculate r^2:

$$
\begin{aligned}
r^2 &= \frac{t^2}{t^2 + df} \times 100 \\
&= \frac{1.50^2}{1.50^2 + 59} \times 100 \\
&= \frac{2.2500}{2.2500 + 36} \times 100 \\
&= \frac{2.2500}{38.2500} \times 100 \\
&= 0.0588 \times 100 \\
&= 5.88\,\%
\end{aligned}
$$

Hypothesis testing says there is not enough evidence to conclude an effect has occurred, but the two effect sizes suggest that a small to moderate effect may be present. Dr. Risen might want to replicate the study with a larger sample size in order to have a better chance of rejecting the null hypothesis and seeing whether temperature does affect the pace of urban life.

How wide is the confidence interval? Applying Equation 8.6, Dr. Risen calculated the 95% confidence interval:

$$
\begin{aligned}
95\%\text{CI}\mu_{\text{Diff}} &= (M_1 - M_2) \pm (t_{cv} \times s_{M_1 - M_2}) \\
&= (3.05 - 2.90) \pm (2.004 \times 0.10) \\
&= 0.1500 \pm 0.2004 \\
&= \text{from } -0.0504 \text{ to } 0.3504 \\
&= [-0.05, 0.35]
\end{aligned}
$$

As expected, the confidence interval, -0.05 to 0.35, captures zero. So, the possibility that there is zero difference between the two population means—walking speed on cold days vs. walking speed on warm days—is plausible. What other information is contained in the confidence interval? It tells Dr. Risen that comparing the walking speed of the population of warm-day pedestrians to the population of cold-day pedestrians, the difference may be from 0.05 mph *slower* on cold days to 0.35 mph *faster* on cold days. In other words, from this study she can't tell if, how much, or in what direction the two populations differ.

But, the confidence interval raises the possibility that the null hypothesis is false and pedestrians walk about a third of a mile per hour faster on cold days. In the language of hypothesis testing, she is worried about Type II error, that there may be an effect of temperature on walking speed that she failed to find. To ease her concern, she's going to recommend replicating the study with a larger sample size. This would increase power, making it easier to reject the null hypothesis if it should be rejected, and it would give a narrower confidence interval.

Putting it all together. Here's Dr. Risen's four-point interpretation. She (1) tells what the study was about, (2) gives the main results, (3) explains what the results mean, and (4) makes suggestions for future research:

> This study was conducted to see if temperature affected the pace of life in a city. The walking speed of pedestrians was measured on a cold day (20°F) and on a warm day (72°F). The mean speed on the cold day was 3.05 mph; on the warm day it was 2.90 mph. The difference in speed between the two days was not statistically significant [$t(59) = 1.50$, $p > .05$]. These results do not provide sufficient evidence to conclude that temperature affects the pace of urban life. However, the confidence interval for the difference between population means raised the possibility that there might be a small effect of temperature on walking speed. Therefore, replication of this study with a larger sample size is recommended.

Application Demonstration

One of the most famous studies in psychology was conducted by Festinger and Carlsmith in 1959. In that study, they explored cognitive dissonance, a phenomenon whereby people change their attitudes to make attitudes consistent with behavior. Their study used an independent-samples *t* test to analyze the results. Let's see what they did and then apply this book's interpretative techniques to update Festinger's and Carlsmith's findings.

Festinger and Carlsmith brought male college students into a laboratory and asked them to perform very boring tasks. For example, one task involved a board with 48 pegs. The participant was asked to give each peg a one-quarter clockwise turn, and to keep doing this, peg after peg. After an hour of such boring activity, each participant was asked to help out by telling the next participant that the experiment was enjoyable and a lot of fun. In other words, the participants were asked to lie.

This is where the experimental manipulation took place: 20 participants were paid $1 for lying and 20 were paid $20 for lying. After helping out, each participant was then asked to rate how interesting he had actually found the experiment. The scale used ranged from −5 (extremely boring) to +5 (extremely interesting). Obviously, the participants should rate the experiment in the negative zone because it was quite boring.

Did how much participants get paid influence their rating? It turned out that it did. The participants who were paid \$1 gave a mean rating in the positive zone (+1.35), and the participants who were paid \$20 gave a mean rating in the negative zone (−0.05). This difference was statistically significant [$t(38) = 2.22$, $p < .05$]. The participants paid less to lie gave statistically significantly higher ratings about their level of interest in the study.

The researchers used cognitive dissonance to explain the results. The participants who had been paid \$20 (\$160 in 2015 dollars) had a ready explanation for why they had lied to the next participant—they had been well paid to do so. The participants who had been paid \$1 had no such easy explanation. The \$1 participants, in the researchers' view, had thoughts like this in their minds: "The task was very boring, but I just told someone it was fun. I wouldn't lie for a dollar, so the task must have been more fun than I thought it was." They reduced the dissonance between their attitude and behavior by changing their attitude and rating the experiment more positively. Festinger's and Carlsmith's final sentence in the study was, "The results strongly corroborate the theory that was tested."

Statistical standards were different 60 years ago. Festinger and Carlsmith reported the results of the *t* test, but they didn't report an effect size or a confidence interval. Fortunately, it is possible to take some of their results and work backward to find that $s^2_{\text{Pooled}} = 3.98$ and $s_{M_1-M_2} = 0.63$. Armed with these values, an effect size and a confidence interval can be calculated.

First, here are the calculations for both effect sizes:

$$
\begin{aligned}
d &= \frac{M_1 - M_2}{\sqrt{s^2_{\text{Pooled}}}} \\
&= \frac{1.35 - (-0.05)}{\sqrt{3.98}} \\
&= \frac{1.4000}{1.9950} \\
&= 0.7018 \\
&= 0.70
\end{aligned}
$$

$$
\begin{aligned}
r^2 &= \frac{t^2}{t^2 + df} \times 100 \\
&= \frac{2.22^2}{2.22^2 + 38} \times 100 \\
&= \frac{4.9284}{4.9284 + 38} \times 100 \\
&= \frac{4.9284}{42.9284} \times 100 \\
&= 0.1148 \times 100 \\
&= 11.48\,\%
\end{aligned}
$$

Now, here are the calculations for the confidence interval:

$$\begin{aligned} 95\%\text{CI}\mu_{\text{Diff}} &= (M_1 - M_2) \pm (t_{cv} \times s_{M_1-M_2}) \\ &= (1.35 - (-0.05)) \pm (2.024 \times 0.63) \\ &= 1.4000 \pm 1.2751 \\ &= \text{from } 0.1249 \text{ to } 2.6751 \\ &= [0.12, 2.68] \end{aligned}$$

From d and r^2, it is apparent that the effect size falls in the medium range. From the confidence interval, one could conclude that though an effect exists, it could be a small one. Based on this study, it would be reasonable to conclude that cognitive dissonance has an effect. But, how much of an effect isn't clear. Cognitive dissonance has gone on to be a well-accepted phenomenon in psychology. However, if this early cognitive dissonance study were being reported by a twenty-first-century researcher, he or she shouldn't say that the results "strongly corroborated" the theory.

Practice Problems 8.3

Apply Your Knowledge

8.07 A dermatologist obtained a random sample of people who don't use sunscreen and a random sample of people who use sunscreen of SPF 15 or higher. She examined each person's skin using the Skin Cancer Risk Index, on which higher scores indicate a greater risk of developing skin cancer. For the 31 people in the no sunscreen (control) condition, the mean was 17.00, with a standard deviation of 2.50. For the 36 people in the sunscreen (experimental) condition, the mean was 10.00, with a standard deviation of 2.80. With $\alpha = .05$, two-tailed, and $df = 65$, $t_{cv} = 1.997$. Calculations showed that $s^2_{\text{Pooled}} = 7.11$, $s_{M_1-M_2} = 0.65$, and $t = 10.77$. Use the results to (a) decide whether to reject the null hypothesis, (b) determine if the difference between sample means is statistically significant, (c) decide which population mean is larger than the other, (d) report the results in APA format, (e) calculate Cohen's d and r^2, and (f) report the 95% confidence interval for the difference between population means.

8.08 An exercise physiologist wondered whether doing one's own chores (yard work, cleaning the house, etc.) had any impact on the resting heart rate. (A lower resting heart rate indicates better physical shape.) He wasn't sure, for example, whether chores would function as exercise (which would lower the resting heart rate) or keep people from having time for more strenuous exercise, increasing the heart rate. The 18 people in the control group (doing their own chores) had a mean of 72.00, with a standard deviation of 14.00. The 17 in the experimental group (who paid others to do their chores) had a mean of 76.00, with a standard deviation of 16.00. Here are the results: $\alpha = .05$, two-tailed, $df = 33$, $t_{cv} = 2.035$, $s^2_{\text{Pooled}} = 225.09$, $s_{M_1-M_2} = 5.07$, and $t = 0.79$. Further, $d = 0.27$, $r^2 = 1.86\%$, and the 95% $\text{CI}\mu_{\text{Diff}}$ ranges from -6.32 to 14.32. Use these results to write a four-point interpretation.

SUMMARY

Differentiate between independent samples and paired samples.

- Two-sample *t* tests compare the mean of one sample (e.g., an experimental group) to the mean of another sample (e.g., a control group) in order to conclude whether the population means differ. There are two types of two-sample *t* tests: independent-samples *t* tests and paired-samples *t* tests. With an independent-samples *t* test, how the cases are selected for one sample has no impact on case selection for the other sample. With paired samples, the selection of cases for one sample influences or determines case selection for the other sample.

Conduct the steps for an independent-samples *t* test.

- To conduct an independent-samples *t* test, the assumptions must be checked, the hypotheses generated, and the decision rule formulated. To find *t*, the difference between the sample means is divided by the standard error of the difference.

Interpret an independent-samples *t* test.

- The decision rule is applied to decide if the null hypothesis is rejected. If it is rejected, the researcher concludes that the difference between *sample* means probably represents a difference between *population* means. Next, Cohen's *d* and r^2 are used to calculate effect size and categorize it as small, medium, or large. If the effect size is meaningful and the null hypothesis was not rejected, the researcher should consider the possibility of Type II error. Finally, the researcher should calculate the confidence interval for the difference and interpret it based on whether (1) it captures zero, (2) how close to zero it comes, and (3) how wide it is.

DIY

Put states into two categories on some basis. You could categorize them into whatever groups you wish—for example, states with nice climates vs. those without, or southern states vs. northern states, or states you would like to live in vs. those you would like to avoid. Include about 5–8 states in each group. Pick some outcome variable on which you want to compare the two groups. For example, is the murder rate different in southern states vs. northern states? Do an online search to find the necessary data. Then use an independent-samples *t* test to analyze the data. Report the results in APA format. Don't forget to calculate an effect size.

KEY TERMS

independent samples – the selection of cases for one sample has no impact on the selection of cases for another sample.

independent-samples *t* test – an inferential statistical test used to compare two independent samples on an interval- or ratio-level dependent variable.

paired samples – case selection for one sample is influenced by, depends on, the cases selected for another sample.

pooled variance – the average variance for two samples.

two-samples *t* test – an inferential statistical test used to compare the mean of one sample to the mean of another sample.

CHAPTER EXERCISES

Review Your Knowledge

8.01 To compute either a single-sample *z* test or a single-sample *t* test, one must know the population ____.

8.02 Two sample *t* tests compare the ____ of one sample to the ____ of another sample.

8.03 Two sample *t* tests use ____ means to draw a conclusion about ____ means.

8.04 A classic experiment might use a two-sample *t* test to compare a ____ group to an ____ group.

8.05 Two different types of two-sample *t* tests are the ____ -samples *t* test and the ____ -samples *t* test.

8.06 If each sample in a two-sample *t* test is a random sample from its population, then the test is an ____-samples *t* test.

8.07 If the selection of cases for one sample determines the cases selected for the other sample, then the samples are ____ samples.

8.08 ____ is the abbreviation for the total sample size in an independent-samples *t* test; ____ and ____ are the abbreviations for the sizes of the samples in the two groups.

8.09 In order to use an independent-samples *t* test to analyze data from two samples, the samples have to be ____ and one needs to know the ____ for each sample.

8.10 The nonrobust assumption for an independent-samples *t* test is ____.

8.11 The ____ assumption for the independent-samples *t* test is the one that allows a researcher to generalize the results back to the larger population.

8.12 The ____ assumption for the independent-samples *t* test says that the amount of variability in the two populations is about equal.

8.13 Researchers are often willing, for an independent-samples *t* test, to assume that the dependent variable is ____.

8.14 One tests the ____ assumption for the independent-samples *t* test by comparing the ____ of the two samples.

8.15 The hypotheses for an independent-samples *t* test are either directional or ____ directional.

8.16 The null hypothesis for a two-tailed independent-samples *t* test, expressed mathematically, is ____.

8.17 The alternative hypothesis for a two-tailed independent-samples *t* test, expressed mathematically, is ____.

8.18 If the null hypothesis for an independent-samples *t* test is true, then the observed difference between the sample means is due to ____.

8.19 The critical value of *t* is the border between the ____ and the ____ zones of the sampling distribution of ____.

8.20 *t* tests commonly are ____ tailed and have ____ set at .05.

8.21 To calculate the degrees of freedom for an independent-samples *t* test, subtract ____ from *N*.

8.22 For a two-tailed test, if t ____ $-t_{cv}$, reject H_0.

8.23 For a two-tailed test, if t ____ t_{cv}, reject H_0.

8.24 When writing the hypotheses for a one-way test, it is easier if one formulates the ____ hypothesis first.

8.25 ____ is the variance of the two samples combined.

8.26 To calculate $s_{M_1-M_2}$, one needs to know the ____ variance, the total sample ____, and the sample ____ of each group individually.

8.27 The numerator in the *t* equation is the difference between the sample ____.

8.28 If a researcher rejects the ____, the researcher is forced to accept the ____.

8.29 If a researcher reports the results of an independent-samples *t* test as showing a statistically significant difference, the researcher has ____ the null hypothesis.

8.30 If one rejects the null hypothesis for an independent-samples *t* test, then look at the sample ____ in order to comment on the ____ of the difference.

8.31 The .05 in APA format indicates that there is a ____% chance of a Type I error.

8.32 If the result of an independent-samples *t* test is written as $t(23) = 5.98$, $p < .05$, then N was ____.

8.33 In APA format, ____ means one rejected a null hypothesis with alpha set at .05 and ____ means one failed to reject it.

8.34 For an independent-samples *t* test, calculate ____ or _____ to quantify the size of the effect.

8.35 If there is absolutely no effect of the independent variable on the dependent variable, then d equals ____.

8.36 A d of $\approx$ ____ is considered a medium effect.

8.37 r^2 calculates the percentage of variability in the ____ that is accounted for by the ____.

8.38 The 95% confidence interval for the difference between population means estimates how ____ or how ____ the difference between the population means might be.

8.39 The 95% confidence interval for the difference between population means *probably* captures the real difference between the ____.

8.40 If the 95% confidence interval for the difference between population means fails to capture zero for a two-tailed test with $\alpha = .05$, then one has ____ the null hypothesis.

8.41 If the 95% confidence interval for the difference between population means comes close to zero, the size of the effect in the population may be ____.

8.42 If the 95% confidence interval for the difference between population means is wide, a reasonable suggestion is to ____ the study with a larger ____.

8.43 If sample size increases but the sample means and standard deviations don't change, then of the three values calculated for the interpretation of the independent-samples *t* test, the one that will not change is ______.

8.44 _____ calculate the size of the effect in the sample; a confidence interval calculates it for the ____.

Apply Your Knowledge

Selecting a test

8.45 A theology professor was curious whether children were as religious as their parents. He obtained a random sample of students at his school and administered an interval-level religiosity scale to them. Using the same scale, he collected information from the same-sex parent for each student. What statistical test should he use to see if there is a difference between a parent's and a child's level of religiosity?

8.46 A demographer working for the U.S. Census Bureau wants to compare salaries for urban vs. rural areas. She gets a sample of psychologists who live in urban areas and a sample of psychologists who live in rural areas. From each, she finds out his or her annual income. What statistical test should she use to see if a difference exists in a psychologist's income as a function of residential status?

8.47 An exercise physiologist classifies people—on the basis of their body mass index, heart rate, and lung capacity—as (a) above average in terms of fitness or (b) below average in terms of fitness. He then directs the same people to walk on a treadmill, individually, at an increasing speed until they can no longer walk. The speed of the treadmill when a person maxes out on walking is the dependent variable. What statistical test should the physiologist use to see if there is a difference in maximum walking speed based on fitness level?

8.48 Some people have white coat hypertension. That is, they grow anxious when a person with a white coat and a stethoscope walks into the examining room to take their blood pressure. As a result, their blood pressure increases.

A family practitioner believes this is quite common. To test her theory, she puts together a random sample of 50 patients and takes two blood pressure measurements, one when each patient first walks into the room and a second, unexpected one after about 15 minutes. What statistical test should she use to see if the two blood pressures differ?

Checking the assumptions

8.49 A developmental psychologist has randomly assigned men to two different groups. After sitting alone at a computer monitor to read a series of stories, each man is asked to rate his level of acceptance of gender typing. (Gender-typed people believe that there are certain roles men should fulfill and certain roles women should fulfill.) Higher scores on the scale indicate higher levels of gender typing; scores on the scale are normally distributed. One group of men, the control group, read gender-typed stories and the other group, the experimental group, read non-gender-typed stories. The researcher found $M_C = 75$, $s_C = 35$, $M_E = 55$, $s_E = 8$. The psychologist is planning to use an independent-samples t test to see if the experimental manipulation has had an impact. Check the assumptions and decide if it is OK to proceed with the planned test.

8.50 A clinical psychologist is studying the effects of an experimental medication on depression. He randomly assigns the next 50 patients at his clinic to receive either (a) Prozac or (b) a placebo. Each patient is treated individually. After eight weeks of treatment, each patient completes an interval-level depression scale. The standard deviations for the two groups are similar and the psychologist believes depression level is normally distributed. The psychologist is planning to use an independent-samples t test to see if there are differences between the two groups at the eight-week mark. Check the assumptions and decide if it is OK to proceed with the planned test.

Writing hypotheses

8.51 An infectious disease specialist is using an independent-samples t test to compare the effectiveness of two treatments for the common cold. (a) Write out H_0 and H_1 and (b) explain what they mean.

8.52 A medical educator is using an independent-samples t test to compare the age of physicians who complete the minimum number of continuing education hours per year vs. those who complete extra hours of continuing education. (a) Write out H_0 and H_1 and (b) explain what they mean.

8.53 There is a lot of evidence that fluoride reduces cavities but not all communities add it to their drinking water. A dentist, who expects to replicate this earlier work, classifies randomly selected communities in his state as (1) adding fluoride to their drinking water or (2) not adding fluoride to their drinking water. He then goes to high schools in these communities, inspects the mouths of all high school seniors, and calculates, for each community, the percentage of these students with cavities. He will compare the means for these values between the two types of communities. (a) Write out H_0 and H_1 and (b) explain what they mean.

8.54 The health department physician in a suburban community warned cat owners about a risk of infection. Cats leave their litter boxes and then jump on counters, trailing bacteria behind them. To highlight the greater risk to the health of cat owners than dog owners, the physician went to the homes of cat owners and dog owners, swabbed kitchen counters, cultured the swabs, and counted the number of bacteria that grew. She planned to use a t test to compare the mean number of bacteria in cat-owning households vs. dog-owning houses. (a) Write out H_0 and H_1 and (b) explain what they mean.

Finding t_{cv}

8.55 If $n_1 = 223$ and $n_2 = 252$, determine the critical value of t for an independent-samples t test, two-tailed, $\alpha = .05$.

8.56 If $n_1 = 17$ and $n_2 = 18$, determine the critical value of t for an independent-samples t test, two-tailed, $\alpha = .05$.

8.57 If $n_1 = 46$ and $n_2 = 46$, determine the critical value of t for an independent-samples t test, two-tailed, $\alpha = .01$.

8.58 If $n_1 = 13$ and $n_2 = 15$, determine the critical value of t for an independent-samples t test, one-tailed, $\alpha = .05$, where the numerator of the t equation is expected to be negative.

Writing the decision rule

8.59 If $t_{cv} = 2.086$, write the decision rule for a two-tailed test for (a) when to reject the null hypothesis and (b) when to fail to reject the null hypothesis.

8.60 If $t_{cv} = 2.396$, write the decision rule for a one-tailed test for (a) when to reject the null hypothesis and (b) when to fail to reject the null hypothesis. (*Hint:* Contemplate the sign of t_{cv} and what that means about what the researcher believes.)

Calculating the pooled variance

8.61 Given $n_1 = 12$, $s_1 = 7.4$, $n_2 = 13$, and $s_2 = 8.2$, calculate s^2_{Pooled}.

8.62 Given $n_1 = 15$, $s_1 = 3.6$, $n_2 = 16$, and $s_2 = 4.3$, calculate s^2_{Pooled}.

Calculating the standard error of the difference

8.63 Given $n_1 = 45$, $n_2 = 58$, and $s^2_{\text{Pooled}} = 5.63$, calculate $s_{M_1-M_2}$.

8.64 Given $n_1 = 23$, $n_2 = 19$, and $s^2_{\text{Pooled}} = 12.88$, calculate $s_{M_1-M_2}$.

8.65 Given $n_1 = 45$, $n_2 = 58$, $s_1 = 5.98$, and $s_2 = 7.83$, calculate $s_{M_1-M_2}$.

8.66 Given $n_1 = 22$, $n_2 = 28$, $s_1 = 9.58$, and $s_2 = 11.13$, calculate $s_{M_1-M_2}$.

Calculating t

8.67 Given $M_1 = 57$, $M_2 = 68$, and $s_{M_1-M_2} = 2.34$, calculate t.

8.68 Given $M_1 = 5.5$, $M_2 = 4.5$, and $s_{M_1-M_2} = 1.23$, calculate t.

8.69 Given $M_1 = -5$, $s_1 = 4.6$, $n_1 = 72$, $M_2 = -1$, $s_2 = 3.3$, and $n_2 = 60$, calculate t.

8.70 Given $M_1 = 48$, $s_1 = 15.0$, $n_1 = 8$, $M_2 = 52$, $s_2 = 14.2$, and $n_2 = 11$, calculate t.

Deciding whether the null hypothesis was rejected

8.71 Given $M_1 = 98$, $M_2 = 103$, $t_{cv} = 2.060$, $t = 2.060$, and a two-tailed test with $\alpha = .05$, (a) decide whether the null hypothesis was rejected or not, (b) tell whether the difference between sample means is a statistically significant one or not, and (c) make a statement about the direction of the difference between the sample means.

8.72 Given $M_1 = 88$, $M_2 = 83$, $t_{cv} = 2.042$, $t = 2.040$, and a two-tailed test with $\alpha = .05$, (a) decide whether the null hypothesis was rejected or not, (b) tell whether the difference between sample means is a statistically significant one or not, and (c) make a statement about the direction of the difference between the population means.

Using APA format

8.73 Given $N = 23$ and $t = 2.0723$, report the results in APA format for a two-tailed test, $\alpha = .05$.

8.74 Given $N = 35$ and $t = 2.0321$, report the results in APA format for a two-tailed test, $\alpha = .05$.

8.75 Given $N = 10$ and $t = 2.3147$, report the results in APA format for a two-tailed test, $\alpha = .05$.

8.76 Given $N = 73$, $\alpha = .05$, one-tailed, t expected to be negative, and $t = -1.65$, report the results in APA format.

Calculating effect sizes

8.77 Given $M_1 = 12$, $M_2 = 17$, and $s^2_{\text{Pooled}} = 4.00$, (a) calculate d and (b) classify the size of the effect.

8.78 Given $M_1 = 88$, $M_2 = 85$, and $s^2_{\text{Pooled}} = 81.00$, (a) calculate d and (b) classify the size of the effect.

8.79 Given $t = 9.87$ and $N = 73$, calculate r^2.

8.80 Given $t = 1.34$ and $N = 49$, calculate r^2.

Calculating confidence intervals

8.81 Given $M_1 = 31$, $M_2 = 24$, $s_{M_1-M_2} = 2.88$, and $t_{cv} = 2.045$, (a) calculate the 95% confidence interval for the difference between population means, and (b) based on the confidence interval, decide if the null hypothesis should be rejected for a nondirectional test with $\alpha = .05$.

8.82 Given $M_1 = -13$, $M_2 = -18$, $s_{M_1-M_2} = 1.48$, and $t_{cv} = 2.015$, (a) calculate the 95% confidence interval for the difference between population means, and (b) based on the confidence interval, decide if the null hypothesis should be rejected for a nondirectional test with $\alpha = .05$.

Writing a four-point interpretation

8.83 An elementary education researcher was interested in seeing how the color used to make corrections on students' papers affected their self-esteem. He assembled first graders and asked them to take a third-grade math test. He told the first graders that the test would be very difficult for them and they might not get very many answers right, but he needed their help. After the students were each called into a room to take the test alone, he pretended to grade it. Everyone had 25% of their answers marked wrong. For half the kids, these answers were marked with red ink, and for the other half, the "incorrect" answers were marked with pencil. Each child then took a self-esteem inventory on which higher scores indicate more self-esteem. The 17 red ink (control group) kids had a mean of 23.00 ($s = 5.00$); the 10 pencil (experimental group) kids had a mean score of 29.00 ($s = 5.00$). Given that information and the rest of the results (below), write a paragraph interpreting the results:

- $\alpha = .05$, two-tailed
- $t_{cv} = 2.060$
- $s^2_{\text{Pooled}} = 25.00$
- $s_{M_1-M_2} = 1.99$
- $t = 3.02$
- $d = 1.20$
- $r^2 = 26.61\%$
- 95% $CI\mu_{\text{Diff}}$ [1.90, 10.10]

8.84 A nutritionist compared the effectiveness of an online diet program to that of an in-person diet program. After three months, she compared the number of pounds of weight lost. The control group (in-person) lost a mean of 18.00 pounds ($s = 14.50$, $n = 16$) and the experimental group (online) lost 16.00 pounds ($s = 13.30$, $n = 21$). Using that and the information below, write a paragraph interpreting the results:

- $\alpha = .05$, two-tailed
- $t_{cv} = 2.030$
- $s^2_{\text{Pooled}} = 191.19$
- $s_{M_1-M_2} = 4.59$
- $t = 0.44$
- $d = -0.14$
- $r = 0.54\%$
- 95% $CI\mu_{\text{Diff}}$ [−11.32, 7.32]

Completing all six steps of hypothesis testing

8.85 A physician compared the cholesterol levels of a representative sample of Americans who ate an American diet vs. a representative sample of those who followed a Mediterranean diet. Below are the means, standard deviations, and sample sizes for both samples. Though in some studies a Mediterranean diet has been shown to be beneficial, this was one of the first studies on an American population and the physician had made no advance predictions about the outcome.

- American diet (control): $M = 230$, $s = 24$, $n = 36$
- Mediterranean (experimental): $M = 190$, $s = 26$, $n = 36$

8.86 An addictions researcher measured tolerance to alcohol in first-year and fourth-year college students. She gave participants a standard dose of alcohol and then had them walk along a narrow line painted on the floor. The higher the percentage of the distance that they were on the line, the greater their tolerance to alcohol. The researcher expected that the older students would show more tolerance to alcohol. Here is the relevant information:

- 1st year (control): $M = 30$, $s = 12.5$, $n = 20$
- 4th year (experimental): $M = 48$, $s = 14.6$, $n = 16$

Expand Your Knowledge

8.87 A researcher completes an independent-samples t test and finds that the probability of two sample means being this far apart, if the null hypothesis is true, is less than .05. Which of the following is true?

a. $\mu_1 = \mu_2$
b. $M_1 \neq \mu_1$
c. There probably is no difference between the two population means.
d. There probably is a difference between the two population means.
e. Sufficient evidence does not exist to draw any conclusion about the population means.

8.88 Which result cannot be true for an independent-samples *t* test?

a. A researcher has rejected the null hypothesis and found $d = 1.50$.
b. A researcher has failed to reject the null hypothesis and found $d = 1.50$.
c. A researcher has rejected the null hypothesis and found $d = 0.10$.
d. A researcher has failed to reject the null hypothesis and found $d = 0.10$.
e. Any of these results can be true.
f. None of these results can be true.

8.89 A consumer group is planning to do 2 two-sample *t* tests. In Test 1, they are going to put together a random sample of items at a jewelry store and compare the prices to a random sample of items at a bookstore, *in order to see which store is more expensive.* In Test 2, they are planning to compare a random sample of textbooks purchased at a campus bookstore to the same books purchased through an online bookseller, *in order to see which store is more expensive.* (a) Determine which test is an independent-samples *t* test and which a paired-samples *t* test. (b) It sounds like each test is answering the same question, "Which store is more expensive?" Rewrite the questions so that they more accurately pose the question that the test answers.

8.90 Explain why the critical value of *t*, one-tailed, $\alpha = .05$ is the same as the critical value of *t*, two-tailed, $\alpha = .10$.

8.91 Margery collected some data from two independent groups and analyzed them with an independent-samples *t* test. No assumptions were violated and she rejected the null hypothesis. Yet, when she calculated a confidence interval for the difference between population means, the confidence interval captured zero. Explain how this is possible.

8.92 Dr. Goddard developed a technique that he thought would increase IQ in adults. He obtained a random sample of 52 adult Americans and randomly assigned half of them to a control group and half to the experimental group. He did nothing to the control group, but he administered his IQ-increasing treatment to the 26 in the experimental group. Afterward, he measured IQs and found that the mean for the control group was 100, while the experimental group had a mean IQ of 102. The standard deviation in both groups was 15. Dr. Goddard found $s_{M_1-M_2} = 4.16$, $t = 0.48$, $r^2 = 0.46\%$, and that the 95% for the difference between population means ranged from –6.36 to 10.36. (a) Did Dr. Goddard reject the null hypothesis? (b) What conclusion should he reach about whether his treatment works to increase IQ? (c) How big is the size of the effect as determined by r^2? (d) What information does the confidence interval give on how sure we are about the impact of the IQ-increasing technique? (e) How worried are you that Dr. Goddard made a Type II error? (f) Do you recommend replicating with a larger sample size?

8.93 Dr. Brigham decided to replicate Dr. Goddard's study (see Exercise 8.92) with a larger sample. She did exactly what Dr. Goddard did, but had 1,002 subjects (501 in the control group and 501 in the experimental group). The means for the two groups were exactly the same, 100 and 102, and both groups again had standard deviations of 15. Dr. Brigham found $s_{M_1-M_2} = 0.95$, $t = 2.11$, $r^2 = 0.44\%$, and the 95% confidence interval ranged from .14 to 3.86. (a) Did Dr. Brigham reject the null hypothesis? (b) What conclusion should she reach about whether Dr. Goddard's treatment increases IQ? (c) How big is the size of the effect as determined by r^2? (d) What information does the confidence interval give on how sure we are about the impact of the IQ-increasing technique? (e) How worried are you that Dr. Brigham made a Type I error?

8.94 Compare your answers for Exercises 8.92 and 8.93. (a) What is the impact of sample size on rejecting the null hypothesis? (b) On conclusions about the effectiveness of treatment? (c) On the size of the effect as determined by r^2? (d) On the confidence interval?

SPSS

Data entry for an independent-samples *t* test in SPSS takes two columns. An example using data for the depth of processing study can be seen in **Figure 8.8**.

	Depth	Num_recall	va
1	1.00	1.00	
2	1.00	3.00	
3	2.00	11.00	
4	1.00	3.00	
5	1.00	5.00	
6	2.00	6.00	
7	2.00	4.00	
8	2.00	13.00	
9			
10			

Figure 8.8 Data Entry in SPSS for an Independent-Samples *t* Test "Depth" is the independent variable used to classify cases into groups, 1 for shallow processing and 2 for deep processing. The dependent variable is "Num_recall." (Source: SPSS)

The first column, with the variable named "Depth," contains information about each case's status on the independent variable. SPSS calls this the "Grouping Variable" because it is used to assign cases into groups, either into the shallow processing group or the deep processing group. SPSS uses numbers, not words, to classify cases. Here, "1" means the case belongs to the shallow processing group and "2" the deep processing group. Note that all the shallow processing cases don't have to be next to each other. As long as a case has the right group number associated with it, SPSS will correctly classify it.

The second column, with the variable named "Num_recall," contains the dependent variable, how many words were recalled. The first case recalled 1 word, the second 3 words, the third 11 words, and so on.

Figure 8.9 shows where the independent-samples *t* test is located in SPSS, under "Analyze," then "Compare Means," and finally, "Independent-Samples T Test. . . ." When one clicks on Independent-Samples T-Test, the box shown in **Figure 8.10** opens up.

In Figure 8.10, the arrow button has already been used to move the dependent variable, Num_recall, into the box for the "Test Variable(s)." The independent variable, Depth, has been moved into the "Grouping Variable" box. Note that the grouping variable now appears as "Depth(? ?)," to indicate that one needs to define the groups.

Figure 8.11 shows the box that opens up when one clicks on the "Define Groups" button in Figure 8.10. A value of 1 for "Group 1" and a value of 2 for "Group 2" were entered. Then click on the "Continue" button, which brings up the box seen in **Figure 8.12**.

Figure 8.12 is like Figure 8.10, but the grouping variable has been defined. We can tell SPSS to complete the *t* test by clicking the "OK" button.

Figure 8.13 shows all the output that SPSS provides. The first box gives descriptive statistics for the two groups. The SPSS results match Dr. Villanova's closely, though SPSS reports the standard deviations to three decimal places.

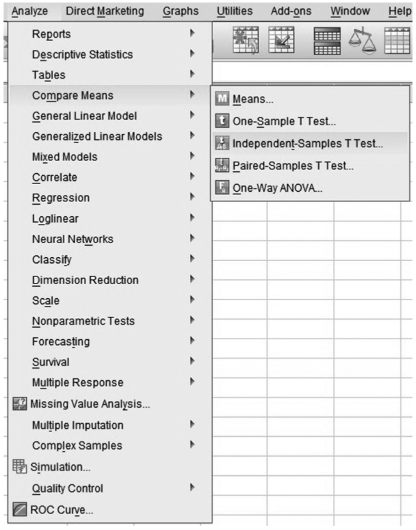


Figure 8.9 Starting an Independent-Samples *t* Test in SPSS The commands for an independent-samples *t* test in SPSS can be found under "Analyze," then "Compare Means." (Source: SPSS)

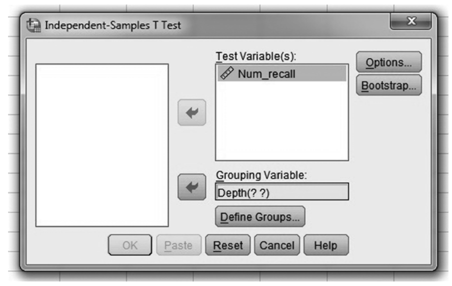


Figure 8.10 Defining Variables as Dependent and Independent in SPSS The dependent variable is called a "Test Variable" in SPSS and the explanatory variable a "Grouping Variable." (Source: SPSS)

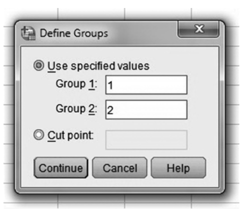


Figure 8.11 Defining the Grouping Variable for an Independent-Samples *t* Test in SPSS Once the grouping variable has been defined, SPSS has to be informed which value is for which group. Group 1 represented by the value "1" is the shallow processing group; Group 2 with a value of "2" is the deep processing group. (Source: SPSS)

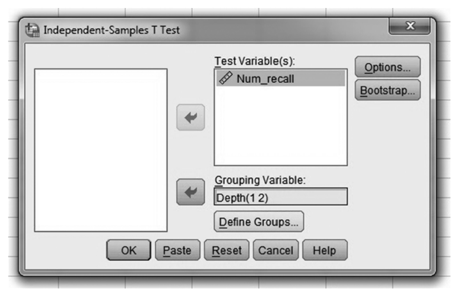


Figure 8.12 Running an Independent-Samples *t* Test in SPSS Once the dependent variable has been set as a test variable and the explanatory variable has been defined as the grouping variable, all that remains is to click on the "OK" button. (Source: SPSS)

Group Statistics

	Depth	N	Mean	Std. Deviation	Std. Error Mean
Num_recall	Shallow	18	3.50	1.543	.364
	Deep	20	8.30	2.736	.612

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Num_recall	Equal variances assumed	6.432	.016	-6.558	36	.000	-4.800	.732	-6.284	-3.316
	Equal variances not assumed			-6.744	30.546	.000	-4.800	.712	-6.252	-3.348

Figure 8.13 Printout for an Independent-Samples *t* Test in SPSS SPSS provides a lot of printout. The first table gives descriptive statistics for the two groups and the second the results of the *t* test. Look in the row labeled "Equal variances assumed" to find the results. (Source: SPSS)

The second box of output presents the results of the *t* test. SPSS gives more information than we need. Just pay attention to the first row, the one that says "Equal variances assumed." SPSS reports a negative value for the test statistic, while Dr. Villanova found a positive value. The different sign for the *t* values doesn't matter because it is merely a result of which mean is subtracted from the other. Because SPSS carries more decimal places than Dr. Villanova did, its *t* value (6.558) is more accurate than his (6.58). Similarly, the SPSS value for the standard error of the difference (.732) is more accurate than Dr. Villanova's (.73).

SPSS also reports degrees of freedom (36) and the exact, two-tailed significance level. If this value (here, .000) is less than or equal to .05, then reject the null hypothesis. If the value is greater than .05, then fail to reject the null hypothesis.

SPSS reports the 95% confidence interval for the difference between population means. Because SPSS subtracted the means in a different order than Dr. Villanova did, it reports the confidence interval as negative numbers, from −6.284 to −3.316. (The SPSS confidence interval is also reported with more decimal places.) Don't let the sign become a concern—by referring back to the population means, one can figure out the direction of the confidence interval.

Finally, SPSS does not report Cohen's *d*. And, unfortunately, it does not report the pooled variance so that Cohen's *d* can be calculated by hand. To calculate *d*, go through the first five steps of Equation 8.3 to calculate s^2_{Pooled}.