CHAPTER 9

Hypothesis Tests Involving Two Population Means or Proportions

> *Basic research is what I am doing when I don't know what I'm doing.*
>
> —Wernher von Braun

LEARNING OBJECTIVES

1. Identify the appropriate sampling distribution to use for a hypothesis test for large samples when you have a two-category independent variable and an interval- or ratio-level dependent variable.
2. Explain why you must use a slightly different hypothesis test when you cannot assume the variances for two groups are equal.
3. Describe the difference between independent groups and matched groups.
4. Conduct a hypothesis for the difference between means using the three variations of the *t* test and interpret the results.
5. Conduct a hypothesis for the difference between proportions and interpret the results.

2 Introduction

Imagine you are working for a state prison system and are in charge of testing a new batterer intervention program (BIP) designed to reduce recidivism for those convicted of intimate partner assaults. You have several individuals in the prison who were convicted of assaulting their intimate partners whom you randomly assign to either get the new BIP (called an experimental group) or be in a control group, whose group members do not receive the program. To measure individuals' approval of using violence to settle conflicts, you give all participants a survey both before the treatment (called a pretest) and after the treatment (called a posttest). To determine whether the BIP actually decreased the participants' approval of violence, you would have to conduct a hypothesis test between mean approval ratings for the pretest and posttest. In this case, the dependent variable would be the approval of violence scores and the independent variable would be a two-category variable indicating whether individuals were in the BIP or the control group. Luckily for you, you will be an expert on these types of tests after reading this chapter!

In this chapter, we will examine the statistical procedures that enable us to test hypotheses about the differences between two population means and two population proportions. We examine two different types of mean difference tests: one for independent samples and one for dependent or matched samples. The independent-samples test is designed to measure mean differences between *two* samples that are independent of each other or *two* subsets within the same sample (i.e., males and females). The key here is that we have two samples that are assumed to be independent. There are two different statistical tests for the difference between two independent samples; the appropriate one to use depends on what assumption we can make about the population variances. In contrast, the matched-groups or dependent-samples test is designed to measure the difference between means obtained for the same sample at two points in time or between two samples that are matched on certain characteristics so as to be as much alike as possible. This is the scenario we presented at the beginning of this section regarding the batterer intervention program. We also examine a test for the difference between two proportions in this chapter, which is a special case of a test for mean differences. In this chapter, we may use the terms "sample" and "group" interchangeably. Let's get started.

2 Explaining the Difference Between Two Sample Means

The hypothesis tests in this chapter are appropriate for the following variables: The independent variable is a two-level or binary categorical variable, and the dependent variable is continuous (interval/ratio). For example, one of the most persistent findings in criminology is the relationship between gender and the number of delinquent offenses committed. Consistently, males report having committed more delinquent acts than females. In a random sample of young males and females, then, the mean number of delinquent acts committed by the males is expected to be greater than the mean for females. In the language of causal analysis, gender is the independent variable that is predicted to "cause" high levels of delinquency, the dependent variable. In this example, gender is the dichotomous independent variable (male/female) and the number of committed delinquent acts is the dependent variable. Let's follow this example through to illustrate the kinds of problems we will encounter in this chapter.

If we were to take a random sample of 70 young males from some population, independently select an equal number of young females, and then ask each youth to self-report the number of times during the past year that he or she committed each of four delinquent offenses (theft, vandalism, fighting, and use of drugs), we would have two means: a mean for the sample of young men ($\overline{X}_{\text{men}}$) and a mean for the sample of young women ($\overline{X}_{\text{women}}$). We also would have two population means—one from the population of men (μ_{men}) and one from the population of women (μ_{women})—that we have not directly measured. Both the sample and the population also have standard deviations (remember, the sample standard deviation is *s* and the population standard deviation is σ). To keep these different components of samples and populations straight, we show them and their respective notations in Table 9.1.

Table 9.1 Characteristics and Notations for Two-Sample Problems

	Population 1	*Population 2*
Population mean	m_1	m_2
Population standard deviation	s_1	s_2
Sample mean	$\overline{X}_1$	$\overline{X}_2$
Sample standard deviation	s_1	s_2
Sample size	n_1	n_2

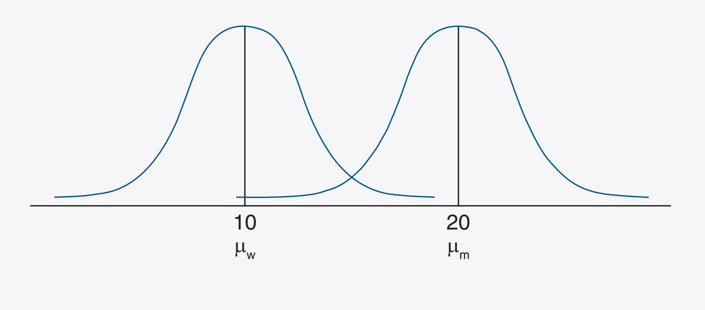
Let's say that, consistent with previous research, the mean number of delinquent offenses committed for the sample of young males is greater than the mean for the sample of young females $(\overline{X}_m > \overline{X}_w)$. As we learned in Chapter 9, we can account for the difference between these two sample means in two very different ways.

One possible explanation for the difference between the male and female sample means is that there really is a difference between the rate at which young men and young women offend. What this explanation implies is that males and females come from different offending populations with different population means (Figure 9.1). This means that there are two distributions of the rate of delinquent offending: one for females (on the left) and one for males (on the right). The population mean for the number of delinquent acts committed is greater for men ($m_m = 20$) than it is for women ($m_w = 10$). Note that if this is true, then when we randomly select a sample of men and record their mean, and randomly select a sample of women and record their mean, more frequently than not the sample mean for men will be greater than the sample mean for women.

A second explanation for the observed difference in sample means between young men and young women is that the two population means are equal ($\mu = 17$), and it was just by chance that we happened to select a sample of males with a higher mean of delinquent offending than our female mean. This is illustrated in Figure 9.2, which shows two distributions of offending: one for the population of males and one for the population of females. Although they may differ in some respects (e.g., their respective standard deviations may be different), the two population means are the same, implying that the mean level of offending is the same for both genders. If this explanation is true, when we draw random samples from both populations, the two sample means will sometimes differ by chance alone, with the male mean sometimes being the larger of the two and the female mean other times being the larger of the two, and sometimes the two means will be equal. Over a large number of mean differences, the mean or average of those differences will be zero. The important point is that if the two population means are equal, the two sample means we obtained in this sample are different because of random sample variation and chance alone.

These two explanations have very different substantive implications. If the first explanation is true, then we will conclude that the mean number of delinquent offenses committed by males is significantly different from the mean number of offenses committed by females. Because the frequency of committing delinquent acts is significantly different between males and females, we will say that there is a "statistically significant" relationship between the independent variable of gender and the dependent variable of delinquency. What we are saying here is that the difference between the male and female sample means is so large that "chances are" the samples came from different populations. In a sense, this means that the sample difference is "real" (a real population

Figure 9.1 Distribution of Male and Female Delinquent Offending With Different Population Means



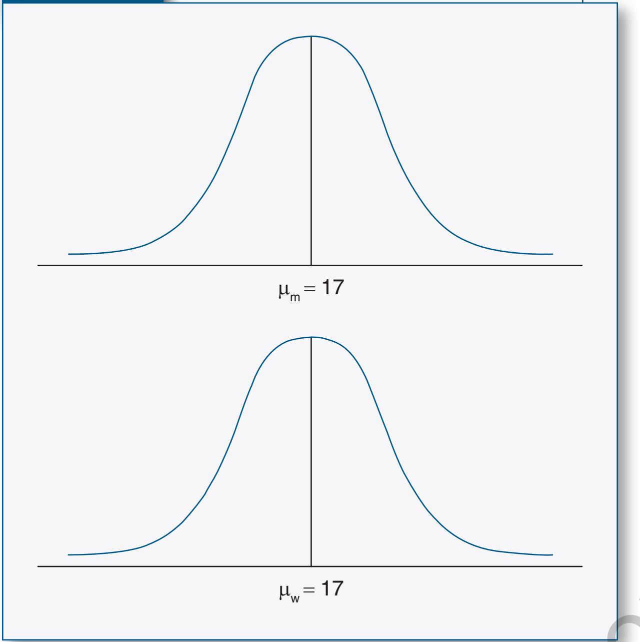
difference). On the other hand, if the second explanation is true, then we will conclude that the observed difference between the male and female means is no greater than what we would expect to observe by chance alone despite the sample means being different. In this case, the sample difference does not reflect a real difference in the population—it's due to luck or chance or random sampling variation, whatever you want to call it.

In sum, because we have sample data, not population data, any difference we actually observe in our sample means may be due to "real differences" between males and females in how frequently they commit delinquent acts or just due to chance/sampling variability. Enter probability theory! With the help of probability theory, we can determine which explanation is *more likely* to be true. In deciding which of these two possible explanations is more likely, we will proceed exactly as we did in the last two chapters when we conducted formal hypothesis tests.

We will begin by assuming that there is no difference between the two population means. That is, we will begin with the null hypothesis that assumes the populations from which both of the samples were drawn have equal means ($\mu_m = \mu_w$). With the use of probability theory and a new kind of sampling distribution, we will then ask, "Assuming that the population means are equal, how likely is it that we would have observed the difference between the two sample means that we actually observed?" If it is a likely event, where "likely" is defined by our alpha level, then we will conclude that our assumption of equal population means cannot be rejected. If, however, we find that the difference between our sample means is an unlikely or very rare event (say, an event with a probability of .05, .01, or .001), we will instead conclude that our assumption of equal population means is not likely true, and we will reject the null hypothesis.

In this chapter, we are interested in something called the *sampling distribution of sample mean differences*. We illustrate the process of hypothesis testing with two sample means in Figure 9.3.

Figure 9.2 Distribution of Male and Female Delinquent Offending With Equal Population Means



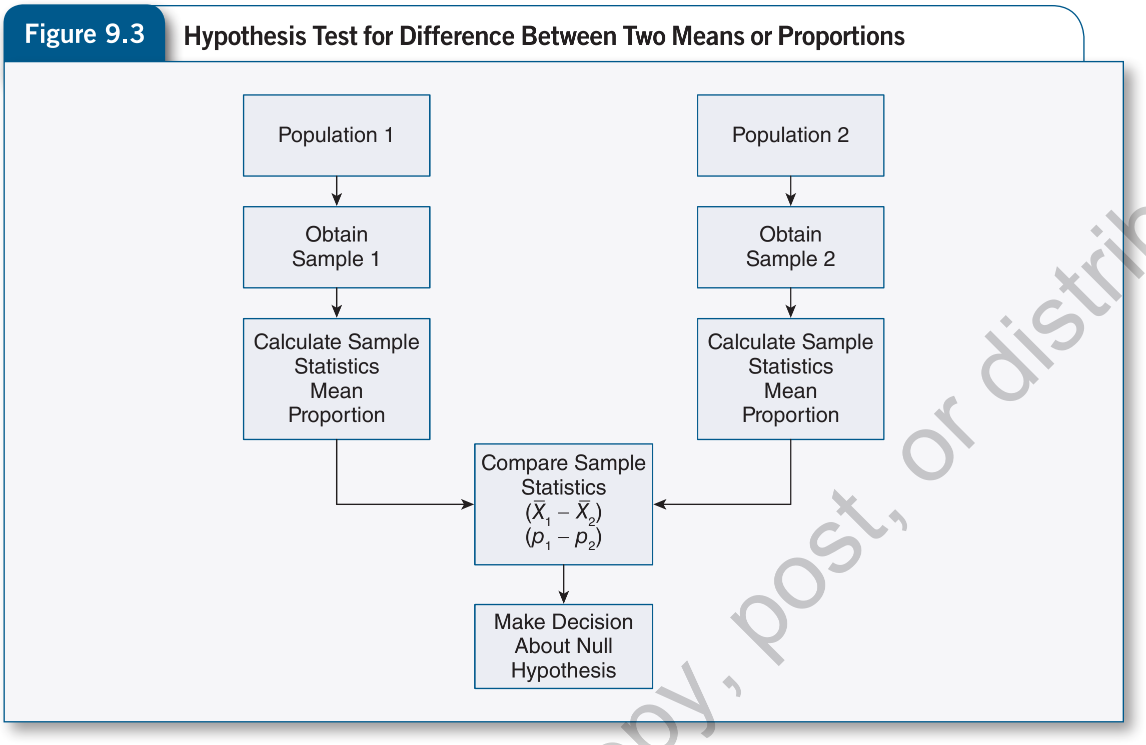
2 Sampling Distribution of Mean Differences

To understand what a sampling distribution of mean differences is, imagine that we take a sample of males and an equal-sized sample of females from their respective populations, compute a mean for each sample, and then calculate the difference between the two sample means $(\bar{X}_1 - \bar{X}_2)$. Imagine that we do this for 10,000 samples, calculating the mean for each group and the difference between the two means so that we now have 10,000 of these mean difference scores $(10{,}000\ \bar{X}_1 - \bar{X}_2)$. We can then create a frequency distribution of these 10,000 mean difference scores. This theoretical distribution of the difference between 10,000 sample means is our **sampling distribution of sample mean differences**. We illustrate what this distribution might look like in Figure 9.4 and provide a summary of the characteristics of this sampling distribution here:

Sampling distribution of sample mean differences: Theoretical distribution of the difference between an infinite number of sample means.

1. The mean of the sampling distribution of the difference between two means, $m_1 - m_2$, is equal to the difference between the population means.

Figure 9.3 **Hypothesis Test for Difference Between Two Means or Proportions**


Population 1
Population 2
Obtain Sample 1
Obtain Sample 2
Calculate Sample Statistics Mean Proportion
Calculate Sample Statistics Mean Proportion
Compare Sample Statistics $(\bar{X}_1 - \bar{X}_2)$ $(p_1 - p_2)$
Make Decision About Null Hypothesis

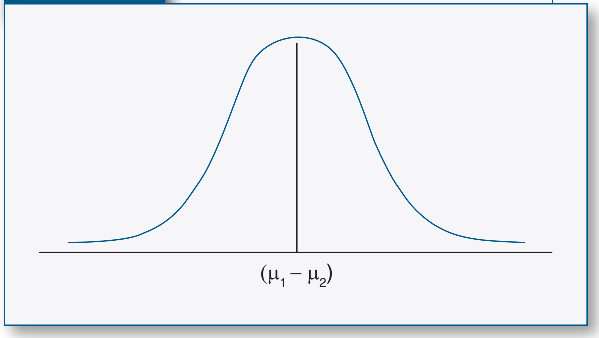
2. The standard deviation of the sampling distribution of the difference between two means $(\sigma_{\bar{X}_1-\bar{X}_2})$ is called the *standard error of the difference* between two means, and it reflects how much variation exists in the difference from sample to sample. In other words, it is the standard deviation of the large number of sample mean differences.

This sampling distribution of mean differences is analogous to the sampling distribution of the mean that we discussed in Chapter 8. What changes is that the sampling distribution in Figure 9.4 is composed of the *difference* between two sample means rather than the distribution of a single mean. In addition, the distribution of mean differences is centered about the difference between the two population means ($\mu_1 - \mu_2$), not around a single population mean (μ).

The mean of this distribution of mean differences is determined by the difference between the two population means. If the two population means are equal ($\mu_m = \mu_w$) as in Figure 9.2, the mean of the sampling distribution of mean differences will be 0 ($\mu_1 - \mu_2 = 0$). As we stated earlier, even if the means in the population are equal, not every sample mean difference is expected to be equal to zero. Sometimes the male sample mean will be greater than the female sample mean, sometimes the female sample mean will be greater than the male sample mean, and sometimes they will be equal. What will be true, however, is that with a very large number of samples, the mean of the distribution of sample differences (the mean difference of the infinite number of sample differences) will be zero.

If the two population means are different, however, as they are in Figure 9.1 with the population mean for men being greater than that for women ($\mu_m > \mu_w$), then most of the sample mean differences $(\bar{X}_1 - \bar{X}_2)$ will be positive. This will be true because in most of the sample comparisons the male mean will be greater than the female mean. In this case, the mean of the sampling distribution of differences will be greater than zero. More specifically, the mean of the sampling distribution will be equal to the difference between the two population means ($\mu_m - \mu_w$).

Figure 9.4 Sampling Distribution of the Difference Between Two Sample Means



Up to now, we have repeatedly stated that no matter what the value of the means for the two populations, when repeated random samples are taken, means are calculated, and differences between sample means are taken, not every mean difference will be exactly the same. There will, then, be dispersion about the mean of the sampling distribution of differences. You can see the spread about the mean of the sampling distribution of differences in Figure 9.4. This dispersion is measured by the standard deviation of sample mean differences, otherwise known as the standard error of the difference $(\sigma_{\bar{X}_1-\bar{X}_2})$, which is defined as

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \qquad (9\text{-}1)$$

where

σ_1 = the standard deviation of the first population

σ_2 = the standard deviation of the second population

Not only do we know the mean and standard deviation of the sampling distribution of differences, we also are in a position to know its shape. An important statistical theorem states the following: If two independent random samples of size n_1 and n_2 are drawn from normal populations, then the sampling distribution of the difference between the two sample means $(\bar{X}_1 - \bar{X}_2)$ will be normally distributed.

We now can use the central limit theorem to generalize this to include any population whenever the sample sizes are large. That is, no matter what the shape of the two populations, if independent random samples of size n_1 and n_2 are drawn, the sampling distribution of the difference between the two sample means will approximate a normal distribution as n_1 and n_2 become large (both sample sizes > 30). With normal populations or with large enough samples from any population, then, the sampling distribution of differences between sample means will approximate normality.

This should sound very familiar to you because it is similar to what we did in Chapter 7. An appropriate statistical test for the difference between two sample means is either a *z* test or *t* test. Therefore, an appropriate sampling distribution would be either a *z* distribution or a *t* distribution. The *z* test for two means is appropriate whenever the two population variances (s_1 and s_2) are known. If these values are unknown, the *t* test for two means is the appropriate statistical test. Since we are seldom in a position to know the value of the population variances, the *t* test is more frequently applied. Keep in mind, however, that when the sample size gets large, the *t* and *z* distributions start to look the same. Now, let's go through some examples of different types of hypothesis tests involving two population means.

2 Testing a Hypothesis About the Difference Between Two Means: Independent Samples

Independent random samples: Samples that are randomly and independently selected.

In this section, we will discuss the case of hypothesis tests for the difference between two *independent* sample means. In the case of **independent random samples**, we have two samples whose elements are randomly and independently selected. Random and independent selection occurs whenever there is a known probability of any element being selected into a sample, and the selection of an element into one sample has no effect on the selection of any element into the other sample. In other words, both samples are randomly selected and are independent of each other.

In our example, independence would occur if the selection of a male into one sample had no effect on the selection of a female into the other sample. The independence assumption is violated in the case of matched groups or dependent samples where an element is deliberately selected into a sample or when the same observations are found in both samples. We will review the special case of hypothesis testing presented by matched groups and dependent samples later in this chapter.

The statistical test we will conduct here is different from the *t* test we used in Chapter 7 in three ways: (1) Our sample statistic is not a single sample mean but rather the difference between two sample means $(\bar{X}_1 - \bar{X}_2)$; (2) the mean of the sampling distribution is not the population mean (μ) but rather the difference between two population means $(m_1 - m_2)$; and (3) the estimated standard deviation of the sampling distribution is the estimated standard deviation of the sampling distribution of the difference between sample means $(\hat{\sigma}_{\bar{X}_1 - \bar{X}_2})$. The general formula for the *t* test involving the difference between two sample means can be expressed as

$$t_{\text{obt}} = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} \tag{9-2}$$

This *t* test requires that the two samples be independent random samples and that the dependent variable be measured at the interval or ratio level.

As you can see from equation 9-2, the *t* statistic is obtained by dividing the difference between the two sample means by the estimated standard deviation of the sampling distribution (the standard error of the difference). There are, however, two versions of the *t* test between two means. In one test, we can assume that the unknown population standard deviations are equal $(s_1 = s_2)$; in the second case, we cannot make that assumption $(s_1 \neq s_2)$. The importance of this is that our estimate of the standard error of the difference $(\hat{\sigma}_{\bar{X}_1 - \bar{X}_2})$ is different for the two cases. We will examine the *t* test for both of these cases separately.

Model 1: Pooled Variance Estimate $(\sigma_1 = \sigma_2)$

If we can assume that the two unknown population standard deviations are equal $(\sigma_1 = \sigma_2)$, we estimate the standard error of the difference using what is called a *pooled variance estimate*. Because the population standard deviations are not known, the decision of whether they are equal is based on the equality of the sample standard deviations (s_1 and s_2). Something called an *F* test is the appropriate test for the significance of the difference between the two sample standard deviations. Without going into too much detail here, the *F* test tests the null hypothesis that $\sigma_1^2 = \sigma_2^2$. If we fail to reject this null hypothesis, we can assume that the population standard deviations are equal and that the *t* test we will discuss in this section for the difference between two population means is the right test. If, however, we are led to reject this null hypothesis, we cannot make the assumption that the two population standard deviations are equal $(\sigma_1^2 \neq \sigma_2^2)$, and we must estimate the standard error of the difference using what is called a *separate variance estimate,* which we will discuss later as Model 2: separate variance estimate. Since the *F* test has not yet been discussed, we will simply provide the information for you whether you can assume that the population standard deviations are equal (Model 1) or whether you cannot make that assumption (Model 2). We cover the *F* test in the next chapter.

In the pooled variance case, our estimate of the denominator of equation 9.2 above (the standard error of the difference) is

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

and the formula for the pooled variance *t* test is

$$t_{\text{obt}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1n_2}}} \tag{9-3}$$

As we have done with other hypothesis tests, once we have our obtained value of *t* (t_{obt}), we will compare it with our critical value (t_{crit}) taken from our probability distribution (the *t* distribution) and make a decision about the null hypothesis. The critical value of *t* is based on our chosen alpha level, whether we have a one- or two-tailed test, and our degrees of freedom and is obtained from the *t* table (Table B.3 of Appendix B).You will remember from Chapter 7 that before using the *t* table, we need to determine the appropriate degrees of freedom in addition to our selected alpha level. When we are testing the difference between two sample means, the degrees of freedom are equal to ($n_1 + n_2 - 2$) in the independent-samples two-sample case for the *t* test. Once we have determined our degrees of freedom, we can go to the *t* table with our chosen alpha level and a one- or two-tailed test and find our critical value.

Let's go through an example of a formal hypothesis test using the *t* test. In this example, we will assume that we have conducted our *F* test and have failed to reject the null hypothesis about equal population standard deviations. Because we can assume that the population standard deviations are equal, then, we can use the pooled variance estimate of the standard error of the difference (Model 1).

Case Study

State Prison Expenditures by Region

Suppose that we are interested in regional differences in the cost of housing state prison inmates. Table 9.2 displays data from state prisons for two regions in the United States (West and Northeast). The dependent variable of interest is the cost per inmate per day, which is measured at the interval or ratio level. Let's say we believe that the average annual cost to house an inmate in state prisons will differ between the West and the Northeast, but we cannot say in which region it will be more costly. In this scenario, region would be the two-category independent variable (West vs. Northeast), and cost would be the dependent variable. We have reviewed the steps necessary to conduct a hypothesis test in Table 9.3. We will go through each of these steps using this example.

Step 1: Since we have no real idea about the nature of the relationship between region of the country and prison costs, a non-directional (two-tailed) hypothesis test is appropriate. The null hypothesis (H_0) will state that the mean annual cost to house prison inmates in the West is equal to the mean cost in the Northeast. The alternative hypothesis (H_1) will represent our belief that the regional means are not equal to each other. These hypotheses are formally stated as follows:

H_0: There is no relationship between region and prison costs OR $\mu_{\text{West}} = \mu_{\text{Northeast}}$

H_1: There is a relationship between region and prison costs OR $\mu_{\text{West}} \neq \mu_{\text{Northeast}}$

Step 2: To determine the validity of this null hypothesis, we will rely on the *t* statistic along with its corresponding probability sampling distribution. Because we can assume that the unknown population standard deviations are equal, we can estimate the standard error of the difference using a pooled variance estimate (Model 1).

Step 3: Let's adopt an alpha level equal to .05 ($\alpha = .05$). With this level of alpha, using a non-directional test and 14 ($8 + 8 - 2 = 14$) degrees of freedom, the critical value of *t* is equal to ±2.145 ($t_{\text{crit}} = \pm 2.145$). Since we have a non-directional alternative hypothesis, the value of t we obtain from our statistical test must be equal to or greater than 2.145 or equal to or less than –2.145 in order to reject the null hypothesis of equal means. In other words, the obtained t value must be greater in absolute terms than 2.145 regardless of the sign. We show the two critical values and critical regions in Figure 9.5.

Step 4: Since we are assuming that the population standard deviations are equal, we can use a pooled variance estimate of the standard error of the difference. Please notice that in the preceding data, we have given you the sample standard deviation; to get the variance that the formula calls for, you will have to square the standard deviation. Sometimes we will provide the standard deviation, and sometimes we will provide the variances; you will have to be on your toes and alert as to which one you are given:

$$t_{\text{obt}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

$$t_{\text{obt}} = \frac{85.37-132.42}{\sqrt{\frac{(8-1)(29.33)^2+(8-1)(21.45)^2}{8+8-2}}\sqrt{\frac{8+8}{(8)(8)}}}$$

$$t_{\text{obt}} = \frac{-47.05}{\sqrt{\frac{(7)860.25+(7)460.10}{14}}\sqrt{\frac{16}{64}}}$$

$$t_{\text{obt}} = \frac{-47.05}{\sqrt{\frac{6,021.75+3,220.7}{14}}\sqrt{.25}}$$

$$t_{\text{obt}} = \frac{-47.05}{\sqrt{660.18}\sqrt{.25}}$$

$$t_{\text{obt}} = \frac{-47.05}{25.69(.50)}$$

$$t_{\text{obt}} = \frac{-47.05}{12.85}$$

$$t_{\text{obt}} = -3.66$$

Step 5: The obtained value of t (t_{obt} = –3.66) falls into the critical region since –3.66 < –2.145 (or is

Table 9.2 Prison Expenditures per Inmate per Day by State and Region, 2011

State	*Daily Mean State Prison Operating Expenditures per Inmate (in Dollars)*
West	
Nevada	56.59
Idaho	53.55
Arizona	67.96
Montana	82.81
Colorado	83.22
California	129.92
Washington	128.48
Utah	80.41
Sample Statistics for the West $\bar{X}_1 = 85.37$ $s_1 = 29.33$ $n_1 = 8$	
Northeast	
New Hampshire	93.37
Pennsylvania	116.00
New York	164.59
New Jersey	150.32
Vermont	135.62
Connecticut	137.70
Maine	127.13
Rhode Island	134.61
Sample Statistics for the Northeast $\bar{X}_2 = 132.42$ $s_2 = 21.45$ $n_2 = 8$	

Source: Adapted from *The Cost of Prisons: What Incarceration Costs Taxpayers* © 2012 from the Vera Institute of Justice.

Table 9.3 Steps Taken When Conducting a Hypothesis Test

Step 1: Formally state your null (H_0) and research (H_1) hypotheses.

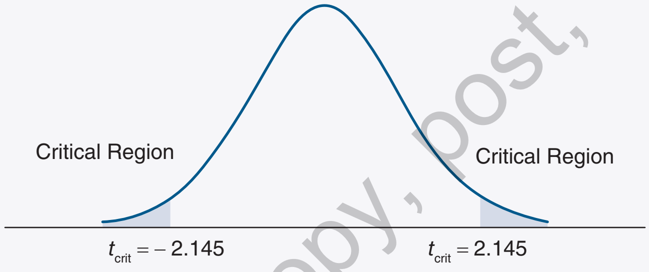
Step 2: Select an appropriate test statistic and the sampling distribution of that test statistic.

Step 3: Select a level of significance (alpha = α) and determine the critical value and rejection region of the test statistic based on the selected level of alpha and degrees of freedom.

Step 4: Conduct the test: Calculate the obtained value of the test statistic and compare it with the critical value.

Step 5: Make a decision about your null hypothesis and interpret this decision in a meaningful way based on the research question, sample, and population.

Figure 9.5 Critical *t* and Critical Region for Alpha = .05 (*df* = 14) and a Two-Tailed Test



greater in absolute terms of the 2.145 needed). The data obtained from our sample, then, provide enough evidence for us to reject the null hypothesis that the population means are actually equal. We conclude that there is a significant relationship between region and annual cost to house prison inmates. The results of our test indicate that region—at least the West versus the Northeast—does affect the cost of incarcerating offenders housed in state prisons. In addition, the direction of this difference seems to be that it is significantly more expensive to house prison inmates in the Northeast than in the West. Let's go through another example.

Case Study

Social Disorganization and Crime

Ever since the days of the Chicago School in the 1920s, criminologists have postulated that states of social disorganization within a residential community increase the likelihood of various kinds of social problems, including unemployment, mental illness, and criminal victimization. One indicator that has been used to measure social disorganization within communities is the extent to which people move in and out of the community. Communities where very few families move in and move out are considered more stable and more organized than those where there is a great deal of population "turnover" in the neighborhoods. This is because in communities with relatively little turnover, residents live in the same place for a long time and get to know their neighbors, and as a result a sense of community becomes established. It is hypothesized that this sense of community and the network of social relationships between community members is responsible for the lower crime rates in these kinds of stable neighborhoods. In this hypothesis, the population turnover in a community is the independent variable and the rate of crime is the dependent variable.

Suppose we wanted to investigate the relationship between social disorganization and household crime. To do so, we collect a random sample of residents in a neighborhood and ask them how many times something has been stolen from or around their home (including their automobile) within the last 6 months (household theft). In addition, we ask them how long they have lived at their current address. From this survey, we divide our sample into two groups according to the length of time they have resided at their address: (1) those who have resided at their current address for less than 1 year, whom we will term "transient," and (2) those who have resided there for more than 1 year, whom we will term "stable." These two categories are our independent variable. We then calculate the mean number of times each group experienced a household theft, which is our dependent variable. For this hypothetical example, we obtain the following sample statistics:

Less Than 1 Year	*More Than 1 Year*
$\bar{X}_1 = 22.4$ $s_1^2 = 4.3$ $n_1 = 49$	$\bar{X}_2 = 16.2$ $s_2^2 = 4.1$ $n_2 = 53$

Step 1: Because we have some idea about the nature of the relationship between residential stability and risk of household victimization, we adopt a directional (one-tailed) hypothesis test. Since we believe that those who have lived in an area less than 1 year (the "transients") will be more vulnerable to becoming victims of household crime than those who have lived in the neighborhood longer (the "stables"), our alternative hypothesis states that the mean number of household victimizations experienced by residents who have lived at their current addresses for less than 1 year will be greater than the mean for those who have lived in their residences for more than 1 year. The null and alternative hypotheses are formally stated as follows:

$$H_0: \mu_{\text{less than 1 year}} = \mu_{\text{more than 1 year}}$$

$$H_1: \mu_{\text{less than 1 year}} > \mu_{\text{more than 1 year}}$$

Step 2: To determine the validity of the null hypothesis, we will rely on the *t* statistic along with its corresponding sampling distribution. Because we can assume that the unknown population standard deviations are equal, we can estimate the standard error of the difference using a pooled variance estimate.

Step 3: For this test, let's select an alpha level of .01. With $\alpha = .01$, using a directional test and degrees of freedom equal to 100 $[(n_1 + n_2 - 2) = (49 + 53 - 2 = 100)]$, the critical value of *t* that defines the rejection region can be found in Table B.3 of Appendix B. By using the degrees of freedom of 120 listed in the table (since that is the closest value we have to 100), we see that the critical value of *t* that defines the lower limit of the rejection region is 2.358. Therefore, to reject the null hypothesis, we must obtain a *t* value equal to or greater than 2.358. We use a positive value of *t* in this case because our alternative hypothesis states that the value of the first mean (for the "transients") will be greater than the value of the mean for the second group (the "stables"); the obtained value of *t,* therefore, is predicted to be positive. If we obtain a negative value of *t,* no matter how large it is, we must fail to reject the null hypothesis. We show the critical value of *t* and the critical region for this problem in Figure 9.6.

Step 4: The next step of our hypothesis test is to convert the difference between our sample means into a *t* value. Notice that you have been given the sample variances in this problem, so there is no need to square the terms again! We're just making sure you are paying attention!

$$t_{\text{obt}} = \frac{22.4 - 16.2}{\sqrt{\frac{(49-1)4.3+(53-1)4.1}{49+53-2}}\sqrt{\frac{49+53}{(49)(53)}}}$$

$$t_{\text{obt}} = \frac{6.2}{\sqrt{\frac{(48)4.3+(52)4.1}{100}}\sqrt{\frac{102}{2{,}597}}}$$

$$t_{\text{obt}} = \frac{6.2}{\sqrt{\frac{206.4+213.2}{100}}\sqrt{.039}}$$

$$t_{\text{obt}} = \frac{6.2}{\sqrt{\frac{419.6}{100}}\sqrt{.039}}$$

$$t_{\text{obt}} = \frac{6.2}{\sqrt{4.2}\sqrt{.039}}$$

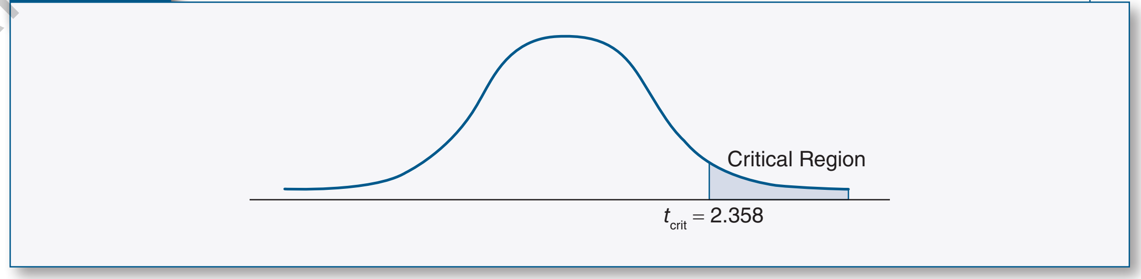
$$t_{\text{obt}} = \frac{6.2}{(2.049)(.198)}$$

$$t_{\text{obt}} = \frac{6.2}{.406}$$

$$t_{\text{obt}} = 15.27$$

Step 5: The *t* value we obtained of 15.27 is substantially greater than the critical value of *t* (2.358) that was needed in order to reject the null hypothesis—it falls into the critical region. Since $t_{\text{obt}} > t_{\text{crit}}$, we will reject the null hypothesis that the population means are equal. This suggests that the observed sample mean difference is too large to be attributed to chance or sampling variation; therefore, we can assume that the mean rate of household victimization experienced by those who have recently moved to a neighborhood is greater than the mean rate experienced by those who have lived in their places of residence for more than 1 year. The results of our statistical test lend support to one of the premises of social disorganization theory; we have found that individuals who have just recently been in a state of transiency (e.g., have moved within the last year) are more likely to become victims of household crime than are those who have been more residentially stable (e.g., have not moved within the last year).

Figure 9.6 Critical *t* and Critical Region for Alpha = .01 (*df* = 120) and a One-Tailed Test



Case Study

Boot Camps and Recidivism

When crime rates in the United States were high in the mid-1980s to the early 1990s, a correctional program called correctional boot camps (sometimes called shock incarceration programs) became very popular in both state and federal prison systems. Although they were rapidly put in place, it was not always clear whether they reduced recidivism any better than regular correctional facilities. Despite their increased popularity, there have been only a few really rigorous attempts to evaluate their efficacy in reducing recidivism. Perhaps some of the most ambitious evaluations of boot camps have been conducted by Doris MacKenzie and her colleagues, who have compared graduates from boot camps with individuals sentenced for the same crimes but sent to prison instead (for a review, see MacKenzie, 2013).

Suppose we want to conduct our own experiment on the issue. We get the help of a local judge and randomly select, from a group of young adult offenders convicted of felony offenses, those who will go to a military-style boot camp and those who will be sent to the state prison for regular correctional programming. After their release, individuals are followed for 2 years. To collect information on offending behavior, we conduct interviews with the individuals and obtain official arrest data. Mean levels of offending behavior (for all crimes, including violent, property, and drug offenses) are calculated for both groups as follows:

Boot Camp Group	*Prison Group*
$\bar{X}_1 = 15.2$ offenses $s_1^2 = 4.7$ $n_1 = 32$	$\bar{X}_2 = 15.9$ offenses $s_2^2 = 5.1$ $n_2 = 29$

In this example, the type of custody (boot camp vs. prison) is our independent variable and the number of offenses committed is our dependent variable. To determine whether there is a significant difference in the mean rates of offending between the boot camp graduates and those released from prison, we must conduct a formal hypothesis test about the difference between two means. To help you learn the steps of formal hypothesis testing, we ask that you check off each step as we go through them.

Because there has been so little research on the efficacy of boot camps, and because the research that does exist is inconsistent, it would be hard to predict in advance which type of programming is more effective in reducing recidivism, so let's select a non-directional alternative hypothesis. The formal hypothesis statements are as follows:

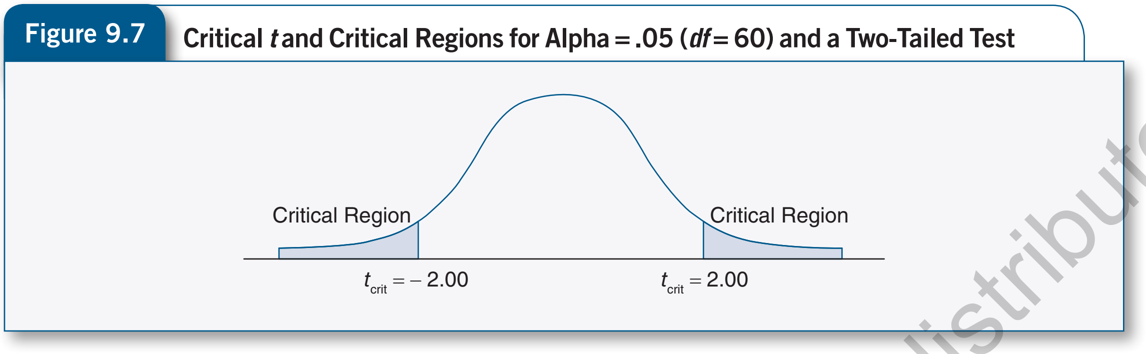
$$\text{H}_0\text{: } \mu_{\text{boot camps}} = \mu_{\text{prison}}$$

$$\text{H}_1\text{: } \mu_{\text{boot camps}} \neq \mu_{\text{prison}}$$

The *t* test, along with its corresponding sampling distribution, is an appropriate statistical test for our data. Let's select an alpha level of .05. With a non-directional hypothesis test, α = .05, and 59 degrees of freedom (32 + 29 – 2 = 59), we will use the critical value of *t* for 60 degrees of freedom because that is the closest value in Table B.3. We will reject the null hypothesis if our obtained *t* is either less than or equal to –2.00 or greater than or equal to +2.00. The critical values and corresponding critical regions are displayed in Figure 9.7. The obtained value of *t* is calculated as follows:

$$t_{\text{obt}} = \frac{15.2 - 15.9}{\sqrt{\dfrac{(32-1)4.7 + (29-1)5.1}{32+29-2}}\sqrt{\dfrac{32+29}{(32)(29)}}}$$

Figure 9.7 Critical *t* and Critical Regions for Alpha = .05 (*df* = 60) and a Two-Tailed Test



$$t_{\text{obt}} = \frac{-.7}{\sqrt{\frac{(31)4.7+(28)5.1}{59}}\sqrt{\frac{61}{928}}}$$

$$t_{\text{obt}} = \frac{-.7}{\sqrt{\frac{145.7+142.8}{59}}\sqrt{.066}}$$

$$t_{\text{obt}} = \frac{-.7}{\sqrt{\frac{288.5}{59}}\sqrt{.066}}$$

$$t_{\text{obt}} = \frac{-.7}{\sqrt{4.89}\sqrt{.066}}$$

$$t_{\text{obt}} = \frac{-.7}{(2.21)(.26)}$$

$$t_{\text{obt}} = \frac{-.7}{.57}$$

$$t_{\text{obt}} = -1.23$$

Our statistical test results in an obtained *t* value of –1.23. An obtained *t* of –1.23 does not lie within the critical region, so we must fail to reject the null hypothesis. Because we failed to reject the null hypothesis, we must conclude that the mean offending rates after release for boot camp and regular prison inmates are not significantly different from one another. This would be in line with much of the research to date on boot camps, which has shown that the core elements of boot camp programs—military-style discipline, hard labor, and physical training—do not reduce offender recidivism any better than regular prison (MacKenzie, 2013).

Model 2: Separate Variance Estimate ($\sigma_1 \neq \sigma_2$)

In the previous examples, we have assumed that the two population standard deviations were equal. Unfortunately, it will not always be possible for us to make this assumption about equal population standard deviations. In many instances, our *F* test will lead us to *reject* the null hypothesis that $s_1 = \sigma_2$ and we must conclude that the two population standard deviations are different. When this happens, we cannot use the pooled variance estimate of the standard error of the difference that we learned in the last section. Instead, we must estimate what is called a separate variance estimate of the standard error of the difference. The formula for this estimate is

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

With a separate variance estimate of the standard error of the difference, the formula for our *t* test now becomes

$$t_{\text{obt}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} \qquad (9\text{-}4)$$

The steps necessary to conduct a hypothesis test remain exactly the same as before except for determining the degrees of freedom. The correct degrees of freedom for the separate variance *t* test are not as easy as $n_1 + n_2 - 2$. In fact, the formula to calculate the degrees of freedom for a *t* test using the separate variance estimate is quite a bit more complicated. The following formula has been suggested to obtain the appropriate degrees of freedom for this test (Blalock, 1979; Hays, 1994):

$$df = \left[\frac{\left(\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1} \right)}{\left(\frac{s_1^2}{n_1 - 1} \right)^2 \left(\frac{1}{n_1 + 1} \right) + \left(\frac{s_2^2}{n_2 - 1} \right)^2 \left(\frac{1}{n_2 + 1} \right)} \right] - 2 \qquad (9\text{-}5)$$

Wow! And you thought the degrees of freedom were relatively unimportant! The results of this formula should be rounded to the nearest integer to obtain the approximate degrees of freedom. Let's go through two examples using the separate variance estimate approach for the *t* test.

Case Study

Formal Sanctions and Intimate Partner Assault

In 1981, the first large-scale experiment to test the deterrent effects of arrest on domestic batterers, called the Minneapolis Domestic Violence Experiment, was conducted by Lawrence Sherman and Richard Berk (1984a, 1984b). The theoretical impetus for this experiment was guided by notions of specific deterrence. The primary research question driving the study was as follows: "Does arresting a man who has assaulted his partner decrease the probability that he will assault her in the future compared with less punitive interventions that are typically used such as separating the parties?" From their study, the researchers concluded that arrest provided the strongest deterrent to future violence and consequently was the preferred police response to domestic violence. This led to many jurisdictions implementing mandatory arrest policies for intimate partner assault.

To test the validity of experimental findings, an important canon of science is *replication*. Accordingly, the National Institute of Justice funded replication experiments of the Minneapolis experiment in six other cities. Unlike the original Minneapolis experiment, the published findings from these replications, which became known as the Spouse Assault Replication Program (SARP), did not uniformly find that arrest is an effective deterrent in spouse assault cases. The effect of arrest on intimate partner assault was revisited recently by Lawrence Sherman and Heather Harris (2015), who examined death rates among the original victims of domestic violence 23 years after the first study in Minneapolis. They found that victims whose abusers were arrested were more likely to die prematurely than those victims whose abusers were simply warned. Clearly, this latest study calls into further question the effectiveness of mandatory arrest policies.

Let's say we attempted to conduct our own study on a much smaller scale about the effects of an arrest policy on future domestic violence. By working with a police department, we would randomly assign arrested suspects who had assaulted their intimate partners to either short-term (no more than 3 hours) or long-term (4 or more hours) detention in jail after their arrest. We would then follow these offenders and victims for a 120-day period and record the number of new victimizations the partners reported to interviewers along with the number of calls to police during that period. The independent variable is the type of detention and the dependent variable is the number of intimate partner assaults perpetrated post-release. The hypothetical mean numbers of post-detention assaults, along with other sample statistics for both groups, are as follows:

Short-Term Detention	*Long-Term Detention*
$\overline{X}_1 = 6.4$	$\overline{X}_2 = 8.1$
$s_1 = 2.2$	$s_2 = 3.9$
$n_1 = 14$	$n_2 = 42$

We would like to test the null hypothesis that the population means for the two groups are equal. In saying this, we are suggesting that the length of detention after an arrest has no effect on the frequency with which intimate partner assault is committed in the immediate future. Suppose also that, based on an *F* test, we rejected the null hypothesis that the population standard deviations are equal; therefore, we must assume they are significantly different and use the separate variance *t* test as our statistical test.

Step 1: Because the literature on the efficacy in deterring intimate partner assault with stiff penalties is unclear, we will conduct a non-directional (two-tailed) alternative hypothesis that states the two population means are simply different. Our null hypothesis states that the two population means are equal or, stated in words, that there is no relationship between the type of detention experienced by arrested suspects and rates of intimate partner assault post-release. The hypotheses are formally stated as follows:

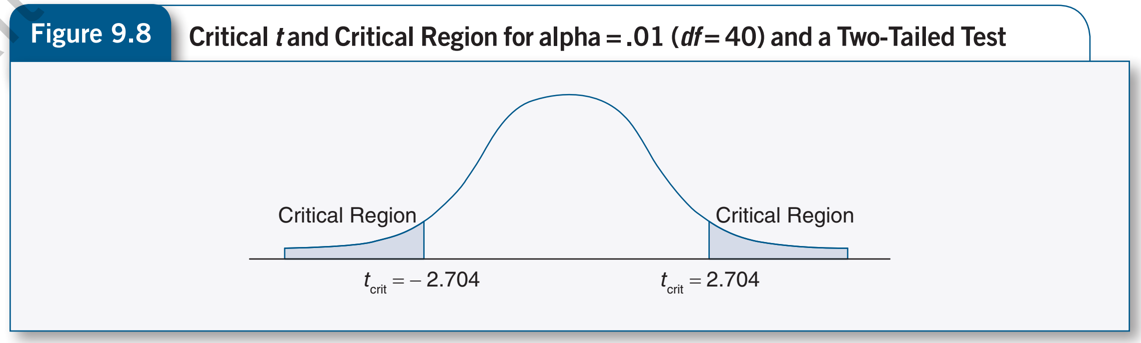
$$\text{H}_0: \mu_{\text{short detention}} = \mu_{\text{long detention}}$$

$$\text{H}_1: \mu_{\text{short detention}} \neq \mu_{\text{long detention}}$$

Step 2: As mentioned earlier, our statistical test will be the separate variance *t* test, and our sampling distribution will be the *t* distribution.

Step 3: We will select an alpha level of .01. To find our critical value of *t* and the critical region, we first need to determine the appropriate degrees of freedom. Based on formula 9-5, we can approximate our degrees of freedom as equal to

Figure 9.8 Critical *t* and Critical Region for alpha = .01 (*df* = 40) and a Two-Tailed Test



$$df=\left[\frac{\left(\frac{4.84}{14-1}+\frac{15.21}{42-1}\right)^2}{\left(\frac{4.84}{14-1}\right)^2\left(\frac{1}{14+1}\right)+\left(\frac{15.21}{42-1}\right)^2\left(\frac{1}{42+1}\right)}\right]-2$$

$$df=\left[\frac{\left(\frac{4.84}{13}+\frac{15.21}{41}\right)^2}{\left(\frac{4.84}{13}\right)^2\left(\frac{1}{15}\right)+\left(\frac{15.21}{41}\right)^2\left(\frac{1}{43}\right)}\right]-2$$

$$df=\frac{(.37+.37)^2}{(.37)^2(.07)+(.37)^2(.02)}-2$$

$$df=\frac{(.74)^2}{.010+.003}-2$$

$$df=\frac{.55}{.013}-2$$

$$df=42.3-2$$

$$df=40.3$$

$$df=40$$

With 40 degrees of freedom and an alpha of .01 for a two-tailed test, our critical values of t are ±2.704 (per Table B.3 in Appendix B). Because we are doing a two-tailed or non-directional test, our critical region will consist of any t_{obt} less than or equal to −2.704 or greater than or equal to 2.704. We show the critical values and critical region in Figure 9.8.

We now calculate our obtained value of t as shown here (notice that we have given you the sample standard deviations rather than the variances):

$$t_{\text{obt}}=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}}$$

$$t_{\text{obt}}=\frac{6.4-8.1}{\sqrt{\frac{(2.2)^2}{14-1}+\frac{(3.9)^2}{42-1}}}$$

$$t_{\text{obt}}=\frac{6.4-8.1}{\sqrt{\frac{4.84}{14-1}+\frac{15.21}{42-1}}}$$

$$t_{\text{obt}}=\frac{-1.7}{\sqrt{\frac{4.84}{13}+\frac{15.21}{41}}}$$

$$t_{\text{obt}}=\frac{-1.7}{\sqrt{.37+.37}}$$

$$t_{\text{obt}}=\frac{-1.7}{.86}$$

$$t_{\text{obt}}=-1.98$$

Our obtained *t* statistic is –1.98. Because this does not fall below the critical negative value of *t* (and, therefore, it does not fall into the critical region), we fail to reject the null hypothesis. Our conclusion, based on our sample results, is that there is no significant difference between the mean number of post-detention assaults for those who were given short-term detention and that for those who were given long-term detention in the population. Thus, it seems that there is no significant relationship between detention time and an arrested batterer's propensity to commit acts of violence in the future. Note that we also had to square the standard deviation that was provided to obtain the variance! Let's go through another quick example.

Case Study

Gender and Sentencing

An area that has received a great deal of research in criminology and criminal justice revolves around the idea of gender disparity in sentencing in both state and federal courts (Daly, 1987; Engen, Gainey, Crutchfield, & Weis, 2003; Starr, 2015; Steffensmeier, Kramer, & Streifel, 1993).

Controversy still exists over whether disparity in sentencing truly exists or whether observed gender differences in sentencing are due to legal characteristics of the offense or the offender. Some research has found that female defendants were sentenced for shorter prison terms than male defendants (Spohn & Spears, 1997; Starr, 2015), whereas other studies have found little or no evidence of gender disparity (Daly, 1994; Rapaport, 1991; Steffensmeier et al., 1993). Darrell Steffensmeier and his colleagues (1993), for example, went so far as to conclude, "When men and women appear in (contemporary) criminal court in similar circumstances and are charged with similar offenses, they receive similar treatment" (p. 411). You should have immediately recognized that gender in this scenario is the independent variable (a two-category, nominal-level variable) and the length of the sentence received is the continuous dependent variable.

Let's assume that we have a random sample of 50 male and 25 female defendants who were found guilty of burglary and sentenced to some time in prison. The mean sentence lengths received for male and female defendants, along with their respective standard deviations and sample sizes, are as follows:

Male Defendants	*Female Defendants*
$\overline{X}_1 = 12.02$	$\overline{X}_2 = 3.32$
$s_1 = 72.68$	$s_2 = 11.31$
$n_1 = 50$	$n_2 = 25$

Step 1: Let's say we believe that males will be sentenced more harshly than females. Accordingly, we state a directional (one-tailed) alternative hypothesis that the population mean sentence length is greater for male defendants than for female defendants. The null hypothesis is that the population means are equal:

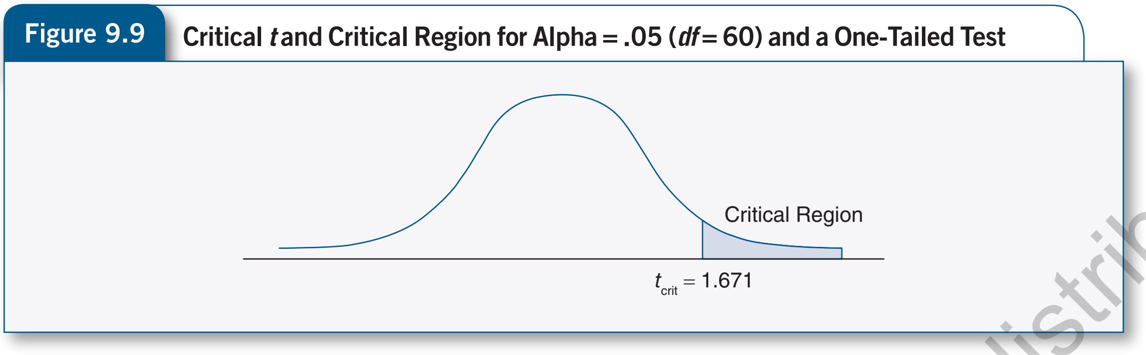
$$H_0: \mu_{\text{males}} = \mu_{\text{females}}$$

$$H_1: \mu_{\text{males}} > \mu_{\text{females}}$$

Step 2: Our test statistic is the separate variance *t* test, and our sampling distribution is the *t* distribution.

Step 3: We will choose an alpha level of .05 (α = .05). Based on formula 9-5, we determine that the approximate degrees of freedom is 56. (We will not show the work here, but it would be a good idea

Figure 9.9 Critical *t* and Critical Region for Alpha = .05 (*df* = 60) and a One-Tailed Test



to compute this for yourself just for practice.) With 56 degrees of freedom, α = .05, and a one-tailed test, we can see from the *t* table that our critical *t* value is 1.671 (actually, this *t* score corresponds to 60 degrees of freedom, but it is the closest value we have to 56 degrees of freedom in the table). Since we have predicted that the population mean for males will be greater than the population mean for females, we will reject the null hypothesis if *t*obt ≥ 1.671 and we will fail to reject the null hypothesis if *t*obt < 1.671. We show the critical value and the critical region in Figure 9.9.

Step 4: We will now calculate our obtained *t* value using the separate variance estimate, as shown:

$$t_{obt} = \frac{12.02 - 3.32}{\sqrt{\frac{(72.68)^2}{50-1} + \frac{(11.31)^2}{25-1}}}$$

$$t_{obt} = \frac{8.7}{\sqrt{107.80 + 5.33}}$$

$$t_{obt} = \frac{8.7}{\sqrt{113.13}}$$

$$t_{obt} = \frac{8.7}{10.64}$$

$$t_{obt} = .82$$

Note that we again had to square the standard deviation to obtain the variance! Our obtained *t* score of .82 is considerably less than the critical *t* of 1.671 and does not fall into the critical region. Our decision, then, will be to fail to reject the null hypothesis that there is no difference in the population means. We will conclude that, based on our sample data, there is not a significant relationship between sentence length received for burglary and gender of defendant in the population. Thus, the sentences handed down by judges for males convicted of robbery do not seem to be greater than the sentences received by females convicted of the same offense in the population. You may be thinking that the vast difference between the means should certainly have produced a significant result. However, remember that the test is also greatly influenced by the variation in each group—note the huge standard deviation around the mean sentence length for males!

So far, we have examined ways of comparing means across two independent samples or groups of cases. In the next section, we will examine a procedure called the matched-groups *t* test that is used to compare two means within the same or comparable group or sample. In this case, we cannot assume that we have two independent samples.

2 Matched-Groups or Dependent-Samples *t* Test

In our application of the *t* test for the difference between two means in the previous section, we assumed that the two samples were independent of each other. That is, we assumed that the selection of the elements of one sample had no effect on the selection of the elements of the other sample. There are times when this assumption is deliberately violated. One instance of this lack of independence occurs when we have a "treatment" group and a "control" group. To make sure that the two groups are comparable with one another in as many ways as possible, each observation in one group is sometimes "matched" with an observation in the other group on relevant characteristics. Matching of samples can be done whenever it is not practical or ethical to randomly assign the "treatment." Matching subjects is done so that the only thing that differentiates the two groups is that one group received a certain type of treatment or was exposed to some phenomenon and the other group did not or was not.

For example, one way to determine the effect of counseling on future delinquency would be to collect data from two samples that are very similar to each other with the exception that one has received counseling (treatment group) and the other has not (control group). In such a study, an 18-year-old White male who lives in an urban area and has no criminal history might be placed in a sample that is to receive treatment (counseling), while another 18-year-old White male who lives in an urban area and has no criminal history might be "matched" to this treatment male but placed in a sample that is to receive no treatment (no counseling). In this case, the two subjects are matched with respect to five characteristics: age (they are both 18 years old), ethnicity (both White), gender (both males), residence (both urban dwellers), and criminal history (both have no criminal history). If the members of the two groups are effectively matched on important characteristics that are related to the dependent variable (such as age, minority status, gender, location of residence, and criminal history), then any observed differences between the two groups on the dependent variable after treatment (future delinquency) are unlikely to be due to these demographic characteristics. Rather, they are more likely to be due to the treatment, which in this example is counseling. The important point here is that by matching someone in one sample with a counterpart in a second sample, we have violated our assumption that the two samples are independent and so cannot use either of our two independent-samples *t* tests.

A second common use of **matched or dependent samples** occurs with "before-and-after" research designs, more generally referred to as pre–post designs. In this type of study, there is only one sample, but measures of the dependent variable are taken at two different points in time, "before" some intervention or treatment and again "after." For example, suppose we have access to only one group of arrested delinquents and all of them are going to receive counseling. In this case, we would have access to the individuals before and after they received counseling. Here, we might use self-report or official arrest data before and after counseling to determine whether counseling actually decreased rates of offending. However, this type of sample also violates our assumption of independence since the same persons appear in both groups. The subjects before the intervention cannot be independent of those after the intervention since they are the same people.

It should be clear to you that the two previously described *t* tests would not be appropriate because we would not have independent samples given that the elements of each sample were deliberately selected to be alike or, in fact, are the same people. In both the matched-groups and the pre–post designs, the independent observation is actually a *pair of cases,* not two independent groups. If we now consider *each pair* as an independent observation, we can conduct a statistical test based on the difference between the scores for each pair. In other words, we will make a pair-by-pair comparison by obtaining a difference score for each pair. Unlike the *t* test for independent samples that tests for the difference between two sample means $(\bar{X}_2 - \bar{X}_1)$, the matched-groups or dependent-samples *t* test calculates the difference between the scores for each *pair* of subjects $(x_2 - x_1)$ where a pair is either a pair of subjects who have been matched on some characteristics or the same person measured at two different points in time. In this example, one population consists of one of the matched groups or the group before the treatment, whereas the second population consists of the other matched group or the group after the treatment.

Matched or dependent samples: Samples in which individuals are either dependent or matched on several characteristics (age, race, gender, etc.) or "before-and-after" samples of the same people.

In the null hypothesis of the *t* test for matched groups or dependent samples, we will assume that the two populations are equal, which implies that the treatment or intervention has no effect. Under the null hypothesis, the difference between each pair of observations is expected to be zero, and the mean of the differences is also expected to be zero. The null hypothesis, then, presumes that the population mean of group differences will be zero. We will symbolize the mean of the population of group differences as μ_D, with the subscript *D* indicating that this is the difference between the two populations. The statistical test in a dependent-samples *t* test, then, is really a single-sample test of the hypothesis that $\mu_D = 0$ (the sample $\bar{X}_D$ statistic is the mean of the difference between each pair of scores in the sample).

Our procedure will be to determine the difference between each pair of scores ($X_D = x_2 - x_1$) in the sample, calculate the mean of these differences ($\bar{X}_D$), and then test whether this sample mean difference is equal to the expected population mean difference (μ_D) of zero. If the null hypothesis is true, then most of these X_D differences will be close to zero, as will the mean of the differences $\bar{X}_D$. If, however, the null hypothesis is not true, then the two scores will tend to be different from each other and the mean difference score will be greater than or less than zero. The greater the difference between each pair of scores, the greater the mean difference will be and the more likely we will be to reject the null hypothesis.

The formula for the *t* test with dependent samples is

$$t = \frac{\bar{X}_D - \mu_D}{s_D / \sqrt{n}} \tag{9-6}$$

Remember that we have drawn an analogy between the *t* test for matched samples and a hypothesis test involving a single population mean. In the *t* test in Chapter 7, where we dealt with one-population problems, we subtracted the population mean from the sample mean and divided by the standard deviation of the sampling distribution. This is exactly what we do in the independent-samples or matched-groups *t* test in formula 9-6. We subtract the population mean (μ_D) from the sample mean ($\bar{X}_D$), where the sample mean is the mean of the differences between each pair of scores in the sample, and we divide by the estimated standard deviation of the sampling distribution, which is the standard deviation of the observed difference scores. Note that the dependent-samples *t* test is based solely on the difference scores X_D (where $X_D = x_2 - x_1$) and the standard deviation of the difference scores (s_D).

Since the null hypothesis assumes that the population mean is zero ($\mu_D = 0$), we can drop that term from the numerator and the formula for the dependent-samples *t* test can be reduced to

$$t_{\text{obt}} = \frac{\bar{X}_D}{s_D/\sqrt{n}} \tag{9-7}$$

where

$$s_D = \sqrt{\frac{\Sigma(x_D - \bar{X}_D)^2}{n-1}}$$

or

$$s_D = \sqrt{\frac{\left(\Sigma D^2 - \frac{(\Sigma D)^2}{n}\right)}{n-1}}$$

The term s_D in equation 9-7 is just our old friend the standard deviation, and we gave you two ways to calculate it that you should recognize from Chapter 4 as the definitional and computational formulas for the standard deviation respectively.

Once we have our obtained the *t* value, we do the same thing we have done with any *t* test discussed so far. We compare t_{obt} with t_{crit} and make a decision about our null hypothesis. We even go to the same *t* table as for independent-samples *t* tests (Table B.3 of Appendix B). The difference is that in the case of matched groups or dependent samples, since we have only *n* pairs of independent observations (rather than $n_1 + n_2$ observations as in the case of independent samples), we have $n - 1$ degrees of freedom where n is equal to the number of *pairs* of observations. If this sounds a bit confusing right now, no worries. A couple of examples will help to illustrate what is going on here. In each example, we will conduct a formal hypothesis test.

Case Study

Problem-Oriented Policing and Crime

Several recent studies have found that more than half of all crimes in a city are committed in only a few places. Some criminologists have called these places "hot spots" (Braga et al., 1999; Caplan, Kennedy, & Piza, 2013; Kennedy, Caplan, & Piza, 2013; Sherman, Gartin, & Buerger, 1989). Even within the most crime-ridden neighborhoods, it has been found that crime clusters at a few locations, while other areas remain relatively free of crime. The clustering of violent crime at particular locations suggests that there are important features of, or key dynamics at, these locations that give rise to frequent violence. Thus, focused crime prevention efforts have recently sought to modify these "criminogenic" conditions and reduce violence.

Problem-oriented policing strategies (similar to community policing) are increasingly used by urban jurisdictions to reduce crime in these high-activity or "hot spot" crime places. Problem-oriented policing challenges officers to identify and analyze the causes of problems behind a string of criminal incidents. Once the underlying conditions that give rise to crime problems are known, police officers can then develop and implement appropriate responses to reduce crime. For example, strategies include using community members as information sources to discuss the nature of the problems the community faces, the possible effectiveness of proposed responses, and the assessment of implemented responses. Other strategies target the social disorder problems inherent in these neighborhoods such as cleaning up the environment of the place and making physical improvements, securing vacant lots, and removing trash from the street.

Suppose we are interested in the efficacy of these policing strategies in reducing acts of violence in neighborhoods plagued by high rates of crime. We target 20 neighborhoods within a city and send out teams of community police officers to implement problem-oriented policing strategies in these neighborhoods. Before the program begins, we obtain the number of arrests for violent offenses that were made in each neighborhood within the 60 days prior to program implementation. After the program has been in place, we again obtain the number of arrests for violent offenses that were made in each neighborhood for a 60-day period. In this case, the program of having problem-oriented policing in the community is the independent variable and the number of violent offenses is the dependent variable. Notice that we have the same neighborhoods here; we have the number of crimes before and after the introduction of problem-oriented policing. We want to know whether the average number of arrests for violent offenses increased or decreased after the policing program was implemented. The hypothetical numbers of arrests for each time point are reported in the second and third columns of Table 9.4. We now are ready to conduct our hypothesis test.

Step 1: First, we state our null and research hypotheses. Our null hypothesis is that the mean difference score in the population is equal to zero. This implies that the problem-oriented policing had no effect on the number of arrests for violent offenses within neighborhoods. Since we are unsure what the exact effect of our problem-oriented policing strategy will be (maybe it will make things better, but maybe with more police it will make things worse or more crime will simply be seen), we will opt for a non-directional alternative hypothesis stating our belief that, on average, the number of arrests in neighborhoods after the new policing strategy was implemented will be different from the number of arrests before problem-oriented policing was implemented. The null and research hypotheses are formally stated as follows:

H_0: $\mu_D = 0$. This implies that the same number of crimes were committed before and after the policing program was put in place

H_1: $\mu_D \neq 0$. This implies that there is some effect of the policing program on crime, but we cannot state in advance if it decreases or increases crime

Step 2: The next step is to state our test statistic and the sampling distribution of that test statistic. Because we have dependent samples (the same community is used before and after the policing program was introduced), we use the dependent-samples *t* test as the statistical test and use the *t* distribution as our sampling distribution.

Step 3: The third step is to select our alpha level and determine the critical value and region. Let's select an alpha level of .01 ($\alpha = .01$) for this example. Because we have 20 pairs of observations ($n = 20$), we have 19 ($20 - 1 = 19$) degrees of freedom. We go to Table B.3 of Appendix B and find that for a two-tailed hypothesis test, $\alpha = .01$, and 19 degrees of freedom, the critical value of *t* is ±2.861. Therefore, we will reject the null hypothesis if t_{obt} is less than or equal to –2.861 or if t_{obt} is greater than or equal to 2.861. We illustrate this for you in Figure 9.10.

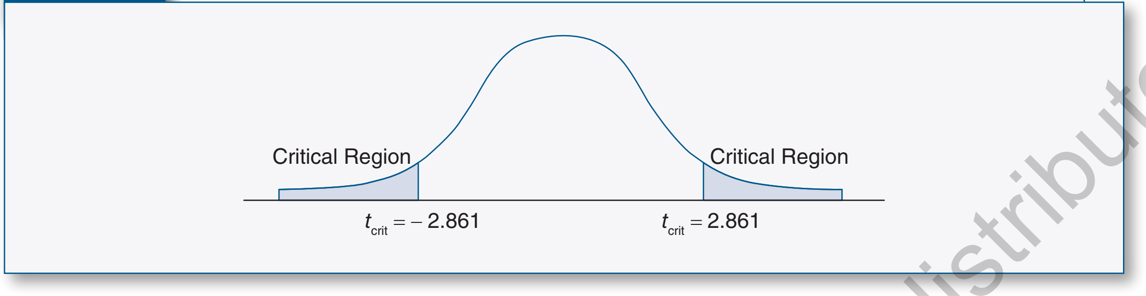
Step 4: The fourth step of our hypothesis-testing procedure is to calculate the test statistic and compare it with our critical value.

For our first example of a matched-groups *t* test, we illustrate the calculations in detail. From equation 9-6, we see that we need to determine the mean of the difference scores and the estimated standard deviation of the difference scores. In Table 9.4, we report that the sum of the difference scores is equal to –48 (ΣX_D or $\Sigma (x_2 - x_1) = -48$). Note how these difference scores are created. For each neighborhood, we subtract the first score (before the policing program was implemented) from the second score (after the policing program was implemented). For example, the first pair of cases

Table 9.4 Number of Arrests for Violent Offenses in Neighborhoods Before (First Score) and After (Second Score) Implementation of Problem-Oriented Policing

Pair Number	First Score x_1	Second Score x_2	$x_2 - x_1$	$(x_2 - x_1)^2$
1	25	21	–4.00	16
2	29	25	–4.00	16
3	32	32	0.00	0
4	42	39	–3.00	9
5	21	25	4.00	16
6	29	25	–4.00	16
7	33	29	–4.00	16
8	35	36	1.00	1
9	32	29	–3.00	9
10	36	35	– 1.00	1
11	39	40	1.00	1
12	25	21	–4.00	16
13	27	25	–2.00	4
14	41	35	–6.00	36
15	36	35	– 1.00	1
16	21	23	2.00	4
17	38	31	–7.00	49
18	25	21	–4.00	16
19	29	25	–4.00	16
20	25	20	– 5.00	25
			$\Sigma = -48$	$\Sigma = 268$
			$\overline{X}_D = -2.40$	

Figure 9.10 Critical *t* and Critical Regions for Alpha = .01 (*df* = 19) and a Two-Tailed Test



had 21 arrests after the problem-oriented policing strategy was implemented and 25 arrests before. The difference, then, is 21 – 25 = –4, or a reduction by 4 crimes. We do this for each neighborhood (each pair), sum across the pairs, and then divide by the number of pairs to obtain a mean difference score. All scores are added in calculating this mean difference score, including zeros and scores with negative signs. With 20 pairs of scores, the mean difference score for these data is –48 / 20, or –2.40 ($\bar{X}_D = -2.40$).

We now calculate the estimated standard deviation of the difference scores. This is just like calculating the standard deviation for any other group of scores except that the raw data are the difference scores and the mean is the mean of the difference scores. First, we will use the definitional formula for the standard deviation. We subtract the mean difference score from each difference score, square this difference, sum these squared differences, divide by the number of pairs minus 1, and then take the square root. This is equal to the standard deviation of the difference scores. To get the standard error or the standard deviation of the sampling distribution, divide this standard deviation by the square root of the sample size. The calculations necessary to find this are shown in Table 9.5.

We can place this into our definitional formula for the standard deviation (Chapter 4):

$$s_D = \sqrt{\frac{\Sigma(X_D - \bar{X}_D)^2}{n-1}}$$

The standard deviation of the difference scores is symbolized as s_D and, for this example, is calculated using the earlier standard deviation formula:

$$s_D = \sqrt{\frac{152.80}{19}}$$

$$s_D = \sqrt{8.042}$$

$$s_D = 2.836$$

Now that we have the standard deviation of the difference scores, we can calculate our test statistic:

$$t_{\text{obt}} = \frac{\bar{X}_D}{s_D/\sqrt{n}}$$

$$t_{\text{obt}} = \frac{-2.40}{2.836/\sqrt{20}}$$

$$t_{\text{obt}} = \frac{-2.40}{2.836 / 4.472}$$

$$t_{\text{obt}} = \frac{-2.40}{.634}$$

$$t_{\text{obt}} = -3.785$$

Table 9.5 **Standard Deviations of the Sampling Distribution for the Number of Neighborhood Arrests for Violent Offenses Before (First Score) and After (Second Score) Problem-Oriented Policing Implementation**

Pair	$x_D - \overline{X}_D$	$(x_D - \overline{X}_D)^2$
1	$-4 - (-2.4) = -1.60$	2.56
2	$-4 - (-2.4) = -1.60$	2.56
3	$0 - (-2.4) = 2.40$	5.76
4	$-3 - (-2.4) = -0.60$	0.36
5	$4 - (-2.4) = 6.40$	40.96
6	$-4 - (-2.4) = -1.60$	2.56
7	$-4 - (-2.4) = -1.60$	2.56
8	$1 - (-2.4) = 3.40$	11.56
9	$-3 - (-2.4) = -0.60$	0.36
10	$- 1 - (-2.4) = 1.40$	1.96
11	$1- (-2.4) = 3.40$	11.56
12	$-4 - (-2.4) = -1.60$	2.56
13	$-2 - (-2.4) = 0.40$	0.16
14	$-6 - (-2.4) = -3.60$	12.96
15	$- 1 - (-2.4) = 1.40$	1.96
16	$2 - (-2.4) = 4.40$	19.36
17	$-7 - (-2.4) = -4.60$	21.16
18	$-4 - (-2.4) = -1.60$	2.56
19	$-4 - (-2.4) = -1.60$	2.56
20	$-5 - (-2.4) = -2.60$	6.76
$n = 20$		$\Sigma(X_D - \overline{X}_{D2}) = 152.80$

Step 5: Finally, we compare our obtained value of t (–3.785) with our critical value (±2.861) and the critical region. Since t_{obt} falls within the critical region ($-3.785 < -2.861$), we can reject the null hypothesis that the mean of the differences is equal to zero. We will conclude that the number of post-arrests for violence is significantly different from the number of pre-arrests. The implementation of problem-oriented policing within our sample of neighborhoods seems to have had a significant impact in reducing the number of arrests for violent offenses made within neighborhoods in the population.

Case Study

Siblings and Delinquency

One of the most comprehensive studies ever undertaken on the causes of delinquent behavior was reported more than 60 years ago by Sheldon and Eleanor Glueck (1950). The Gluecks compared 500 institutionalized chronic delinquents with a matched group of 500 non-delinquents. Among their findings, the Gluecks reported that members of the delinquent group were more likely than the non-delinquents to come from broken homes and economically disadvantaged families, to have friends who were also delinquents, and to have parents who were cruel and erratic in their discipline.

Table 9.6 **Number of Delinquent Siblings for 15 Delinquent Youths and a Matched Group of 15 Non-delinquent Youths and the Calculations Necessary for a Matched-Group *t* Test**

Pair	*Non-delinquent Score* x_1	*Delinquent Score* x_2	x_D $x_2 - x_1$	x_D^2 $(x_2 - x_1)^2$	$x_D - \bar{X}_D$	$(x_D - \bar{X}_D)^2$
1	1	3	2	4	2 – 1.40 = 0.60	0.36
2	0	2	2	4	2 – 1.40 = 0.60	0.36
3	0	1	1	1	1 – 1.40 = –0.40	0.16
4	1	4	3	9	3 – 1.40 = 1.60	2.56
5	2	1	–1	1	–1 – 1.40 = –2.40	5.76
6	0	3	3	9	3 – 1.40 = 1.60	2.56
7	2	2	0	0	0 – 1.40 = –1.40	1.96
8	1	4	3	9	3 – 1.40 = 1.60	2.56
9	0	1	1	1	1 – 1.40 = –0.40	0.16
10	0	2	2	4	2 – 1.40 = 0.60	0.36
11	0	0	0	0	0 – 1.40 = –1.40	1.96
12	1	2	1	1	1 – 1.40 = –0.40	0.16
13	0	2	2	4	2 – 1.40 = 0.60	0.36
14	1	3	2	4	2 – 1.40 = 0.60	0.36
15	0	0	0	0	0 – 1.40 = –1.40	1.96
$n = 15$			$\Sigma x_D = 21$ $\bar{X}_D = 21/15 = 1.40$ $\Sigma x_D^2 = 51$			$\Sigma(x_D - \bar{X}_D)^2 = 21.60$ $s_D = \sqrt{\dfrac{21.60}{15-1}} = 1.24$

Let's presume that, like the Gluecks, we have a group of 15 non-delinquents and a group of 15 delinquents who are matched with respect to social class, gender, age, race, and whether both parents are in the home. For each youth, we also have the number of siblings who he or she reports have been arrested for a crime. What we want to know is whether the delinquent youths have more delinquent siblings than the non-delinquent youths. In this scenario, whether a youth has been arrested for delinquency is the independent variable and the number of delinquent siblings is the dependent variable. The data from the two groups are reported in the second and third columns of Table 9.6.

Step 1: Our null hypothesis is that the number of delinquent siblings is not different between the two matched groups. In other words, we are assuming that the population mean for the difference between the pair of scores is zero. Based on our knowledge of the delinquency literature, we will make the directional (one-tailed) alternative hypothesis that the delinquent group will have more siblings who have violated the law than the non-delinquent group. Our prediction, therefore, is that if the number of law-violating siblings for the non-delinquent group is the first score and the number of law-violating siblings for the delinquent group is the second score, then the difference scores ($x_2 - x_1$) will generally be positive and the population mean for the differences will be greater than zero. The hypotheses are formally stated as follows:

$H_0: \mu_D = 0$

$H_1: \mu_D > 0$

Step 2: Our test statistic will be the dependent-samples *t* test, and the sampling distribution will be the *t* distribution.

Step 3: For our hypothesis test, we will choose an alpha level of .05. Since our alternative hypothesis stated that the true population mean was greater than zero, our critical region will lie in the right tail of the sampling distribution. With $n - 1$ or 14 degrees of freedom, $\alpha = .05$, and a one-tailed test, we can find in the *t* table (Table B.3) that $t_{\text{crit}} = 1.761$. The critical region consists of all obtained *t* scores that are greater than or equal to 1.761. Therefore, we will fail to reject the null hypothesis if $t_{\text{obt}} < 1.761$. We show the critical *t* value and critical region in Figure 9.11.

Step 4: The second and third columns of Table 9.6 show the calculations necessary to determine both the mean and the standard deviation of the difference scores. We use the definitional formula for the standard deviation of the difference scores, but you would have obtained the same value with the computational formula! The value of t_{obt} is calculated as follows:

$$t_{\text{obt}} = \frac{\bar{X}_D}{\sqrt{\dfrac{\Sigma(D-\bar{X}_D)^2}{n-1}} \,/\, \sqrt{n}}$$

$$t_{\text{obt}} = \frac{1.40}{\sqrt{\dfrac{21.60}{14}} \,/\, \sqrt{15}}$$

$$t_{\text{obt}} = \frac{1.40}{\sqrt{1.54} \,/\, \sqrt{15}}$$

$$t_{\text{obt}} = \frac{1.40}{1.24 \,/\, 3.87}$$

$$t_{\text{obt}} = \frac{1.40}{.32}$$

$$t_{\text{obt}} = 4.375$$

Figure 9.11 Critical *t* and Critical Regions for alpha = .05 (*df* = 14) and a One-Tailed Test

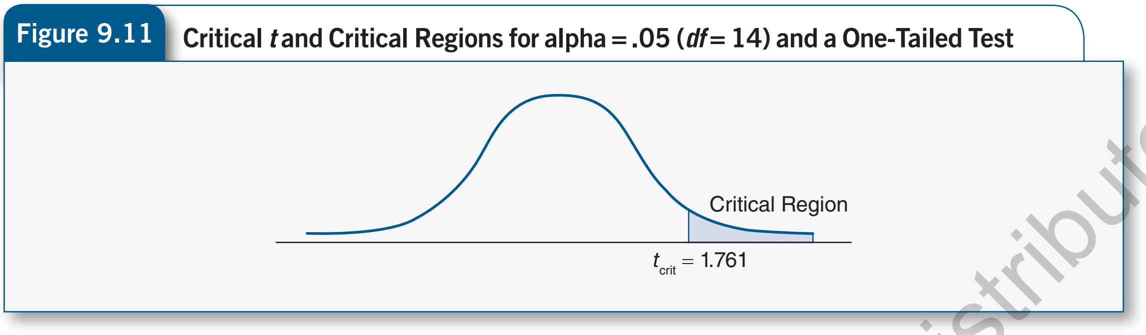
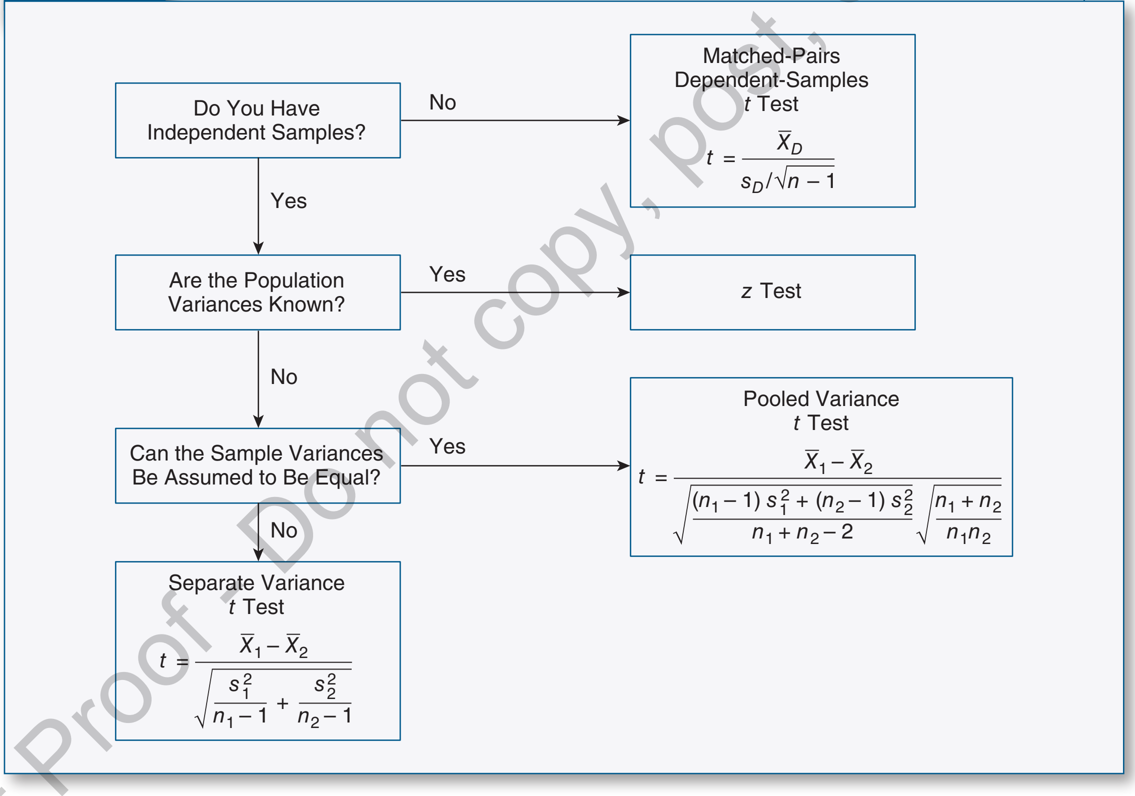


Figure 9.12 Decision Chart for Using the Appropriate Statistical Test for Two-Sample Mean Problems


Do You Have Independent Samples?
No
Matched-Pairs Dependent-Samples t Test
$t = \frac{\bar{X}_D}{s_D/\sqrt{n-1}}$
Yes
Are the Population Variances Known?
Yes
z Test
No
Can the Sample Variances Be Assumed to Be Equal?
Yes
Pooled Variance t Test
$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$
No
Separate Variance t Test
$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}}}$

Step 5: The obtained value of our test statistic is 4.375. Because $t_{obt} > t_{crit}$, we can reject the null hypothesis that the mean population difference is zero. We conclude that there is a significant relationship in the population between delinquency and the number of delinquent siblings a youth has. In the population, we can assume delinquents have significantly more siblings who have violated the law than non-delinquents.

In this and the previous two sections of this chapter, we have examined several different types of statistical tests to test a hypothesis about two population means. This must present a somewhat bewildering picture, and we admit that it might seem a bit overwhelming right now. In selecting the appropriate test statistic for the two-sample mean problem, however, a good deal of confusion can be eliminated if you remember that you need to answer only a few fundamental questions before deciding which test is appropriate for your problem. Figure 9.12 summarizes these decisions. Think of this figure as a road map in deciding which statistical test you should use for two-sample mean problems. In the next section, we will examine hypothesis tests about the difference between two sample proportions.

2 Hypothesis Tests for the Difference Between Two Proportions: Large Samples

In this section, we will examine a statistical test for the significance of the difference between two population *proportions* (P_1 and P_2). Think of the difference of proportions test as a special case of the difference of means test. There are many cases in our discipline where this test is applicable.

Let's say we have a random sample of 100 persons and we asked each of them whether they favor the death penalty for those who commit first-degree murder. We arbitrarily assign a score of "0" for those who say no and a score of "1" for those who say yes. Let's assume that 89 of the 100 persons said they approved of the death penalty under that circumstance and 11 said they did not. Since there are only two values (0 = no and 1 = yes), we can treat this variable as being measured at the interval level. We can determine the mean of this variable by counting the number of "1" scores (or "0" scores) and dividing by n. Since we have 89 "1" scores, the mean would be 89 / 100, or .89. The mean for a binary variable (a variable with only two values coded as "0" and "1"), then, is the proportion of "1" scores—in this case, the proportion of our sample that was in favor of the death penalty. The mean, therefore, is actually the proportion of "1" scores. Even though the population is dichotomous (it is made up of 0s and 1s), we know from the central limit theorem that with a large enough sample, the distribution of sample means and the difference between two sample means will be approximately normal. Hence, we can use a z test and the z distribution to test hypotheses about the difference between two proportions.

In this section, we will consider only tests appropriate for data obtained from large independent samples. If $n\hat{p} \geq 5$ *and* $n\hat{q} \geq 5$ for each of the two samples (where $\hat{p}$ is the sample proportion and $\hat{q} = 1 - \hat{p}$), the sampling distribution of the difference between proportions will be approximately normal and we can use a z test as our test statistic.

In calculating the test statistic for the t test for two sample means, we subtracted one sample mean from the other and divided by the standard error of the difference between means. We will conduct the same procedure in our test for the difference between two proportions. In our z test for two proportions, we will subtract the two sample proportions ($\hat{p}_1 - \hat{p}_2$) and divide by our estimate of the standard deviation of the sampling distribution of the difference between proportions $(\hat{\sigma}_{\hat{p}_1 - \hat{p}_2})$. This estimated standard deviation is also referred to as the standard error of the difference between proportions. The z test for the difference between proportions is

$$z_{\text{obt}} = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,\hat{q}}\sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}} \qquad (9\text{-}8)$$

where

$\hat{p}_1$ = the sample proportion for the first sample

$\hat{p}_2$ = the sample proportion for the second sample

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$$\hat{q} = 1 - \hat{p}$$

Again, notice that in formula 9-8 the p terms in the numerator have a subscript because they come from our two samples, whereas the p and q terms under the first radical in the denominator do not. Just as with all of the preceding hypothesis tests, once we have obtained the test statistic, we compare our z_{obt} with z_{crit} and make a decision about the null hypothesis. Let's go through an example.

Case Study

Education and Recidivism

One of the primary questions in the correctional literature is to determine what programs within the correctional setting decrease inmates' rates of recidivism once they are released. Recently, Ryang Kim and David Clark (2013) examined recidivism rates between inmates who participated in prison-based college programs and those who did not. Not surprisingly to some, they found that the recidivism rate for the inmates who had participated in the college program was lower than the recidivism rate for those who did not participate.

Let's say we have an independent random sample of 120 inmates from a correctional institution where 60 inmates in this group have received either their associate or baccalaureate degree while in prison and the remaining 60 inmates have received no educational curriculum whatsoever. Of the 60 who had received an education, 18% ($\hat{P}_1 = .18$) were rearrested within 1 year of their release from prison. Of the 60 who had not received any education, 38% ($\hat{P}_2 = .38$) were rearrested within the same time period. We wonder whether there is a significant difference between the percentage of released inmates who were rearrested (our measure of recidivism) for those who received an education in prison compared with those who did not. To answer this question, we need to conduct an explicit hypothesis test.

Step 1: Our null hypothesis is that the two samples came from populations with the same proportion of inmates who were rearrested after release. In other words, receiving an education while in prison had no effect on the likelihood of recidivating during the year after release. To be on the safe side, we test a non-directional (two-tailed) alternative hypothesis that the two proportions are simply different from each other. These hypotheses are stated as follows:

$H_0: P_1 = P_2$

$H_1: P_1 \neq P_2$

Step 2: To test these hypotheses, we select as our test statistic the z test for a difference of proportions. Since we have a large sample size, the z distribution will be our sampling distribution.

Step 3: We will select an alpha level of .01. For a two-tailed test, the critical level of z at $\alpha = .01$ is $z_{crit} = \pm 2.58$ (see Table B.2 in Appendix B or Table 7.1 in Chapter 7 for the critical values of z for common levels of alpha). Since this is a two-tailed test, the critical region lies in both tails of the z distribution and consists of all obtained z scores less than or equal to –2.58 or greater than or equal to 2.58. We will reject the null hypothesis if z_{obt} is less than or equal to –2.58 or greater than or equal to +2.58. Figure 9.13 shows the two critical regions and the critical z values.

Step 4: To make the calculations more manageable, we will find our obtained value of z in a series of steps.

Step 4a: We find the estimated value of the pooled population proportions:

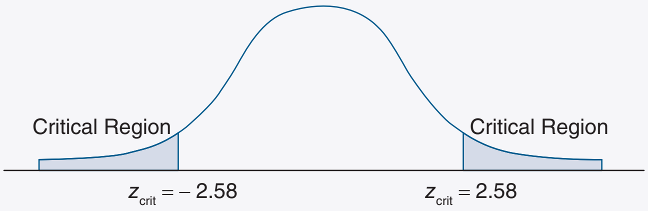
$$\hat{p} = \frac{(60)(.18)+(60)(.38)}{60+60}$$

$$\hat{p} = \frac{10.8+22.8}{120}$$

$$\hat{p} = \frac{33.6}{120}$$

$$\hat{p} = .28$$

Figure 9.13 Critical z and Critical Regions for Alpha = .01 and a Two-Tailed Test



Step 4b: We find the standard error estimate of the difference between population proportions:

$$\hat\sigma_{p_1-p_2} = \sqrt{\hat p\ \hat q}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

$$\hat\sigma_{p_1-p_2} = \sqrt{(.28).72}\sqrt{\frac{60+60}{60(60)}}$$

$$\hat\sigma_{p_1-p_2} = \sqrt{.20}\sqrt{\frac{120}{3{,}600}}$$

$$\hat\sigma_{p_1-p_2} = (.45)\sqrt{.033}$$

$$\hat\sigma_{p_1-p_2} = (.45)(.18)$$

$$\hat\sigma_{p_1-p_2} = .081$$

Step 4c: Finally, plugging our sample proportions into the numerator and this standard error estimate into the denominator of formula 9-8, we calculate the value of our obtained z test:

$$z_{\text{obt}} = \frac{\hat p_1 - \hat p_2}{\sqrt{\hat p\ \hat q}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

$$z_{\text{obt}} = \frac{.18-.38}{.081}$$

$$z_{\text{obt}} = \frac{-.2}{.081}$$

$$z_{\text{obt}} = -2.47$$

Step 5: Our obtained z statistic is –2.47. This value of z_{obt} just misses falling into our rejection region. Since it does not lie within the critical region, we must fail to reject the null hypothesis. We cannot conclude, based on our sample data, that in the population the proportion of inmates who recidivate is significantly different between those inmates who receive an education in prison and those who do not. To test yourself, conduct the same null hypothesis using an alpha of .05 ($\alpha = .05$). What do you conclude?

2Summary

In this chapter, we have examined techniques used to perform hypothesis tests to determine the difference between two means and two proportions. With unknown population variances, the statistical test for the difference between two means is conducted with a *t* test. If the test involves two independent random samples, we can choose from two different kinds of *t* tests. The first type is called a pooled variance *t* test. This test for two-sample means is appropriate when we can assume that the population standard deviations are equal. When we cannot maintain that assumption, the correct *t* test to use is the separate variance *t* test.

In addition to these tests for independent samples, we also examined a *t* test for matched groups or dependent samples. In this kind of *t* test, we are less interested in the difference between two means than in testing whether the difference between two sets of scores is equal to zero.

Finally, we learned how to test for the significance of the difference between two proportions and discovered that it was a special instance of the two-sample mean test.

Key Terms

➤ Review key terms with eFlashcards. SAGE edge™

independent random samples 226
matched or dependent samples 240
sampling distribution of sample mean differences 224

Key Formulas

Pooled variance *t* test (equation 9-3):

$$t_{\text{obt}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

Separate variance *t* test (equation 9-4):

$$t_{\text{obt}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}}$$

Degrees of freedom for separate variance *t* test (equation 9-5):

$$df = \left[\frac{\left(\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}\right)^2}{\left(\frac{s_1^2}{n_1-1}\right)^2\left(\frac{1}{n_1+1}\right)+\left(\frac{s_2^2}{n_2-1}\right)^2\left(\frac{1}{n_2+1}\right)}\right]-2$$

Dependent-samples *t* test (equation 9-7):

$$t_{\text{obt}} = \frac{\bar{X}_D}{s_D/\sqrt{n}}$$

Difference between proportions *z* test (equation 9-8):

$$z_{\text{obt}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,\hat{q}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

Practice Problems

➤ Test your understanding of chapter content.
Take the practice quiz. SAGE edge™

1. Explain the difference between independent and dependent variables. If you think that low self-control affects crime, which is the independent variable and which is the dependent variable?

2. When is it appropriate to use an independent-samples *t* test, and when is it appropriate to use a *t* test for dependent samples or matched groups?

3. John Worrall and colleagues found that the fear of losing the good opinion of one's family and peers kept people from driving home drunk (Worrall, Els, Piquero, & TenEyck, 2014). Let's say we have two independent random samples of people: those who think that their peers would disapprove of them for driving drunk and those who think that their peers either would not care or would approve of their driving drunk. We ask each person in each group to self-report the number of times that he or she has driven drunk during the past 12 months. Here are the results:

Would Not Approve of Driving Drunk	*Would Approve of Driving Drunk*
$n_1 = 40$	$n_2 = 25$
$x_1 = 2.1$	$x_2 = 8.2$
$s_1 = 1.8$	$s_2 = 1.9$

Test the null hypothesis that the two population means are equal against the alternative hypothesis that the group whose peers would not approve of driving drunk has a lower mean rate of driving drunk. In your hypothesis test, assume that the unknown population standard deviations are equal and use an alpha level of .01.

4. The use of monetary fines as a criminal sanction is being considered as one possible solution to the problem of prison overcrowding. Supporters of the use of fines contend that it would be both an effective deterrent to crime and a way to punish even moderately severe crimes without imprisonment. Critics argue that giving criminal offenders fines only increases their motivation to commit more crimes in order to pay their fines. You want to test the effect of fines versus incarceration on criminal behavior. You take a random sample of 150 convicted offenders who have been given a fine as punishment and follow them up for 3 years. You take a second independent random sample of 110 offenders recently released from prison and follow them up for 3 years. At the end of the 3-year follow-up period, you find that 33% of those given a fine had been rearrested and 38% of those given a prison sentence had been rearrested. Test the null hypothesis that the proportions rearrested in the two groups are equal against the alternative hypothesis that they are different. Use an alpha level of .05.

5. Jason Ingram and William Terrill (2014) conducted some research on the perceptions that female and male police officers have of their roles, the public, and their departments. They concluded that female and male police officers do not view their jobs very differently. Let's say that we wanted to continue their work and were interested in how female and male police officers view one component of police work: the handling of domestic disputes. To do this research, we have created a scale that measures how important settling domestic disputes is and whether it is perceived as part of "police work." Those who score high on this scale think that the fair settling of domestic disturbances is of high priority and that it should be an important part of a police officer's duties. We have then taken two random samples. One is a sample of 50 male police officers, and the other is an independent random sample of 25 female police officers. We give each officer a questionnaire that includes our domestic dispute scale. The mean score for female officers is 21.3 with a standard deviation of 3.0. The mean score for male officers is 18.8 with a standard deviation of 4.5. Test the null hypothesis that the two population means are equal against the alternative hypothesis that the male mean is lower than the female mean. In your hypothesis test, *do not* presume that the population standard deviations are equal and use an alpha level of .05.

6. Capital punishment law is among the most complex bodies of law in our legal system. As a result, judges make frequent errors in capital cases in terms of their rulings regarding a change of venue, the decision to sequester jurors, questions of voir dire, suppression of evidence, and so on. When these errors are made, cases are often won on appeal and have to be retried or have a second penalty phase hearing. The Trial Judges Association thinks that only judges who have received special training should sit on capital cases because these judges would commit fewer errors and there would be fewer cases lost on appeal. You decide to test this hypothesis. You take a random sample of 15 judges who have received extensive training in capital punishment law. You match these judges with 15 other judges who have not received such training but are matched in terms of their number of years on the bench, experience as trial lawyers, gender, and age. You want the

two groups of judges to be alike in every way except the experience of capital punishment law training. The data on your matched groups of judges are as follows:

Number of Cases Lost on Appeal

Judge	Untrained	Trained
1	3	0
2	1	3
3	2	4
4	7	4
5	5	2
6	4	5
7	6	1
8	2	1
9	7	0
10	5	6
11	3	4
12	4	2
13	5	5
14	6	3
15	2	1

Test the null hypothesis that the mean difference in the number of cases lost on appeal for the two groups of judges is zero against the alternative hypothesis that the untrained judges lose more cases on appeal. Use an alpha level of .01.

7. Adrian Raine (1994) discussed some research in biological criminology suggesting that children with criminal parents are more likely to be criminals themselves than are children with noncriminal parents. Suppose you conduct a study on a random sample of 100 delinquent youths confined in a correctional institution and a random sample of 75 non-delinquent youths. You find that 43% of the delinquent youths have at least one criminal parent but that only 17% of the non-delinquent youths have a criminal parent. Test the null hypothesis that the two population proportions are equal against the alternative hypothesis that the delinquent group has a greater proportion of criminal parents. Use an alpha level of .01.

8. It is common wisdom to believe that dropping out of high school leads to delinquency. For example, Travis Hirschi's (1969) control theory might predict that those with little or no commitment to education are delinquent more often than those with strong educational commitments. In his general strain theory, however, Robert Agnew (1992) might predict that dropping out of school would lower one's involvement in delinquency because it would get youths out of an aversive and painful environment. You want to examine the relationship between dropping out of high school and delinquency. You have a random sample of 11 students. You have the number of delinquent offenses that each student reported committing during the year before dropping out of school and the number of offenses that each reported committing during the year after dropping out of school. Here are those data:

Number of Delinquent Acts

Person	Before	After
1	5	7
2	9	5
3	2	3
4	7	7
5	8	11
6	11	13
7	8	4
8	8	10
9	5	7
10	2	1
11	9	3

Test the null hypothesis that the mean difference between the two sets of scores is zero against the alternative hypothesis that it is different from zero. Use an alpha level of .05.