similar to lambda and does not offer many advantages in this context. A formula for this statistic can be found in R. A. Cooper and A. J. Weeks, *Data, Models, and Statistics Analysis* (Oxford: Philip Allan, 1983), and other general statistics books.

10. The McNemar test statistic is defined as $\chi^2_{McNemar} = (|f_{0,1} - f_{1,0}|)^2/(f_{0,1} + f_{1,0})$.
11. The chi-square statistic with Yates' continuity correction is defined as

$$\sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

12. The following measures may also be considered. Phi (ϕ) is defined as $\sqrt{\chi^2/n}$, and ranges from zero to one for two-by-k tables ($k \geq 2$). Phi-squared (ϕ^2) has a "variance-explained" interpretation; for example, a ϕ^2 value of 0.35 (or $\phi = 0.59$) means that 35 percent of the variance in one variable is explained by the other. Yule's Q is a measure of association with a PRE interpretation but without a test of statistical significance. Yule's Q is defined as follows. Assume the following two-by-two table:

A	B
C	D

Then, Yule's Q is $\frac{(AD) - (BC)}{(AD) + (BC)}$.

13. With the sample sizes of Table 10.7, the discontinuation rate of the alternative treatment would have to drop further to 13.3 percent (2 of 15 clients). Then, Goodman and Kruskal's tau is .180 (p = .022), chi-square with continuity correction is 3.750 (p = .53), and the Fisher exact test shows p = .50.
14. Another approach is to calculate power (see Box 9.1). Using the above continuation rates and $n = 60$ for each group, we find power to be 71.8 percent, which is indeed close to 80 percent.
15. The formula for H is

$$\frac{12}{n(n+1)}\left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots\right) - 3(n+1),$$

where T_i is the sum of ranks in Group 1, and so on.

16. The Friedman test is quite sensitive to the number of items; it is best to have at least 10 rows.

The T-Test

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Test whether two or more groups have different means of a continuous variable
- Assess whether the mean is consistent with a specified value
- Evaluate whether variables meet test assumptions
- Understand the role of variable transformations
- Identify t-test alternatives

When analysts need to compare the means of a continuous variable across different groups, they have a valuable tool at their disposal: the t-test. T-tests are used for testing whether two groups have different means of a continuous variable, such as when we want to know whether mean incomes vary between men and women. They could also be used to compare program performance between two periods, when performance in each period is measured as a continuous variable.

The examples in this chapter differ from those in Chapters 9 and 10 in that in this chapter's examples one of the variables is continuous and the

other is categorical. Many variables are continuous, such as income, age, height, case loads, service calls, and counts of fish in a pond. Moreover, when ordinal-level variables are used for constructing index variables (see Chapter 3), the resulting index variables typically are continuous as well. When variables are continuous, we should not recode them as categorical variables just to use the techniques of the previous chapters. Continuous variables provide valuable information about distances between categories and often have a broader range of values than ordinal variables. Recoding continuous variables as categorical variables is discouraged because it results in a loss of information; we should use tests such as the t-test.

Statistics involving continuous variables usually require more test assumptions. Many of these tests are referred to as *parametric statistics*; this term refers to the fact that they make assumptions about the distribution of data and also that they are used to make inferences about population parameters. Formally, the term "parametric" means that a test makes assumptions about the distribution of the underlying population. Parametric tests have more test assumptions than nonparametric tests, and most typically that the variable is continuous and normally distributed (see Chapter 7). These and other test assumptions are also part of t-tests.

This chapter focuses on three common t-tests: for independent samples, for dependent (paired) samples, and the one-sample t-test. For each, we provide examples and discuss test assumptions.

This chapter also discusses nonparametric alternatives to t-tests, which analysts will want to consider when t-test assumptions cannot be met for their variables. As a general rule, a bias exists towards using parametric tests because they are more powerful than nonparametric tests. Nonparametric alternatives to parametric tests often transform continuous testing variables into other types of variables, such as rankings, which reduces information about them. While nonparametric statistics are easier to use because they have fewer assumptions, parametric tests are more likely to find statistical evidence that two variables are associated; their tests often have lower p values than nonparametric statistics.[1]

T-TESTS FOR INDEPENDENT SAMPLES

T-tests are used to test whether the means of a continuous variable differ across two different groups. For example, do men and women differ in their levels of income, when measured as a continuous variable? Does crime vary between two parts of town? Do rich people live longer than poor people? Do high-performing students commit fewer acts of violence than do low-performing students? The t-test approach is shown graphically in Figure 11.1, which illustrates the incomes of men and women as boxplots (the lines in the middle of the boxes indicate the means rather than the medians).[2]

Figure 11.1 The T-Test: Mean Incomes by Gender

When the two groups are independent samples, the t-test is called the ***independent-samples t-test.*** Sometimes the continuous variable is called a "test variable" and the dichotomous variable is called a "grouping variable." The t-test tests whether the *difference of the means* ($\Delta\bar{x}$, or $\bar{x}_1 - \bar{x}_2$) *is significantly different from zero,* that is, whether men and women have different incomes. The following hypotheses are posited:

H_0: Men and women do not have different mean incomes (in the population).
H_A: Men and women do have different mean incomes (in the population).

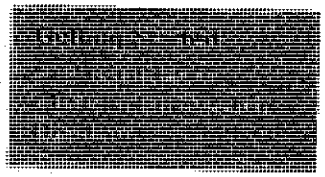
Alternatively, using the Greek letter μ to refer to differences in the population, H_0: $\mu_m = \mu_f$, and H_A: $\mu_m \neq \mu_f$. The formula for calculating the t-test test statistic (a tongue twister?) is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

As always, the computer calculates the test statistic and reports at what level it is significant. Such calculations are seldom done by hand. To further conceptual understanding of this formula, it is useful to relate it to the discussion of hypothesis testing in Chapter 9. First, note that the difference of means, $\bar{x}_1 - \bar{x}_2$, appears in the numerator: the larger the difference of means, the larger the t-test test statistic, and the more likely we might reject the null hypothesis. Second, s_p is the pooled variance of the two groups, that is, the weighted average of the variances of each group.[3] Increases in the standard deviation decrease the test statistic. Thus, it is easier to reject the null hypotheses when two populations are clustered narrowly around their means than when they are spread widely around them. Finally, more observations (that is, increased information or larger n_1 and n_2) increase the size of the test statistic, making it easier to reject the null hypothesis.

The test statistics of a t-test can be positive or negative, although this depends merely on which group has the larger mean; the sign of the test statistic has no substantive interpretation. *Critical values* (see Chapter 9) of the t-test are shown in Appendix C as ***(Student's) t-distribution.***[4] For this test, the *degrees of freedom* are defined as $n - 1$, where n is the total number of observations for both groups. The critical value decreases as the number of observations increases, making it easier to reject the null hypothesis.

The t-distribution shows one- and two-tailed tests. ***Two-tailed t-tests*** should be used when analysts do not have prior knowledge about which group has a larger mean; ***one-tailed t-tests*** are used when analysts do have such prior knowledge. This choice is dictated by the research situation, not by any statistical criterion. In practice, two-tailed tests are used most often, unless compelling a priori knowledge exists or it is known that one group cannot have a larger mean than the other. Two-tailed testing is more conservative than one-tailed testing because the critical values of two-tailed tests are larger, thus requiring larger t-test test statistics in order to reject the null hypothesis.[5] Many statistical software packages provide only two-tailed testing. The above null hypothesis (men and women do not have different mean incomes in the population) requires a two-tailed test because we do not know, a priori, which gender has the larger income.[6] Finally, note that the t-test distribution approximates the normal distribution for large samples: the critical values of 1.96 (5 percent significance) and 2.58 (1 percent significance), for large degrees of freedom (∞), are identical to those of the normal distribution.



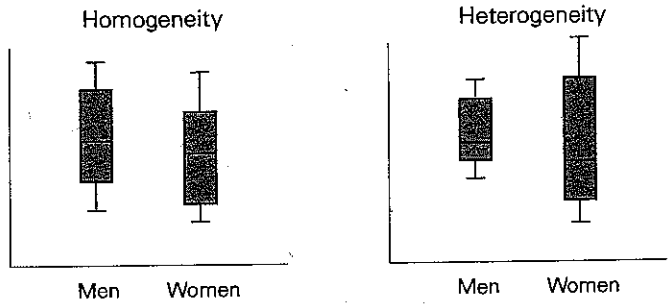
T-Test Assumptions

Like other tests, the t-test has various *test assumptions* that must be met to ensure test validity. Statistical testing always begins by determining whether test assumptions are met before examining the main research hypotheses. This section discusses these tests, as well as ways in which tests and variables are adapted to meet test assumptions. ***Four t-test test assumptions*** must be met to ensure test validity:

- One variable is continuous, and the other variable is dichotomous.
- The two distributions have equal variances.
- The observations are independent.
- The two distributions are normally distributed.

The *first assumption*, that one variable is continuous and the other dichotomous, usually does not present much of a problem. Some analysts use t-tests with ordinal rather than continuous data for the testing variable. This approach is theoretically controversial because the distances among ordinal categories are undefined. This situation is avoided easily by using nonparametric alternatives (discussed later in this chapter). Also, when the grouping variable is not dichotomous, analysts need to make it so in order to perform a t-test. Many statistical software packages allow dichotomous variables to be created from other types of variables, such as by grouping or recoding ordinal or continuous variables.

Figure 11.2 Equal and Unequal Variances



The *second assumption* is that the variances of the two distributions are equal. This is called ***homogeneity of variances.*** The use of pooled variances in the earlier formula is justified only when the variances of the two groups are equal. When variances are unequal (called ***heterogeneity of variances***), revised formulas are used to calculate t-test test statistics and degrees of freedom.[7] The difference between homogeneity and heterogeneity is shown graphically in Figure 11.2. Although we needn't be concerned with the precise differences in these calculation methods, all t-tests *first* test whether variances are equal in order to know which t-test test statistic is to be used for *subsequent* hypothesis testing. Thus, every t-test involves a (somewhat tricky) two-step procedure. A common test for the equality of variances is the ***Levene's test.*** The null hypothesis of this test is that variances are equal. Many statistical software programs provide the Levene's test along with the t-test, so that users know which t-test to use—the t-test for equal variances or that for unequal variances. The Levene's test is performed first, so that the correct t-test can be chosen.

The term ***robust*** is used, generally, to describe the extent to which test conclusions are unaffected by departures from test assumptions. T-tests are relatively robust for (hence, unaffected by) departures from assumptions of homogeneity and normality (see below) when groups are of approximately equal size. When groups are of about equal size, test conclusions about any difference between their means will be unaffected by heterogeneity.

The *third assumption* is that observations are independent. (Quasi-) experimental research designs violate this assumption, as discussed in Chapter 10. The formula for the t-test test statistic, then, is modified to test

whether the *difference* between before and after measurements is zero. This is called a ***paired t-test***, which is discussed later in this chapter.

The *fourth assumption* is that the distributions are normally distributed. Although normality is an important test assumption, a key reason for the popularity of the t-test is that t-test conclusions often are robust against considerable violations of normality assumptions that are not caused by highly skewed distributions. We provide some detail about tests for normality and how to address departures thereof. Remember, when nonnormality cannot be resolved adequately, analysts consider nonparametric alternatives to the t-test discussed at the end of this chapter. Box 11.1 provides a bit more discussion about the reason for this assumption.

A combination of visual inspection and statistical tests is always used to determine the normality of variables. Two tests of normality are the ***Kolmogorov-Smirnov test*** (also known as the K-S test) for samples with more than 50 observations and the ***Shapiro-Wilk test*** for samples with up to 50 observations. The ***null hypothesis of normality*** is that the variable *is* normally distributed: thus, we *do not* want to reject the null hypothesis. A problem with statistical tests of normality is that they are *very sensitive* to small samples and minor deviations from normality. The extreme sensitivity of these tests implies the following: whereas failure to reject the null hypothesis indicates normal distribution of a variable, rejecting the null hypothesis does not indicate that the variable is not normally distributed. It is acceptable to consider variables as being normally distributed when they visually appear to be so, even when the null hypothesis of normality is rejected by normality tests. Of course, variables are preferred that are supported by both visual inspection and normality tests.

Remedies exist for correcting substantial departures from normality, but these remedies may make matters worse when departures from normality are minimal. The *first* course of action is to identify and remove any outliers that may affect the mean and standard deviation. The *second* course of action is ***variable transformation***, which involves transforming the variable, often by taking $\log(x)$, $\sqrt{x}$ or x^2 of each observation, and then testing the transformed variable for normality. Variable transformation may address excessive skewness by adjusting the measurement scale, thereby helping variables to better approximate normality.[8] Substantively, we strongly prefer to make conclusions that satisfy test assumptions, regardless of which measurement scale is chosen.[9] Keep in mind that when variables are transformed, the units in which results are expressed are transformed, as well. Examples of variable transformation are provided below.

Typically, analysts have different ways to address test violations. Examination of the causes of assumption violations often helps analysts to better understand their data. Different approaches may be successful for addressing

In Greater Depth...

Box 11.1 Why Normality?

[illegible]

test assumptions. Analysts should not merely go by the result of one approach that supports their case, ignoring others that perhaps do not. Rather, analysts should rely on the weight of robust, converging results to support their final test conclusions.

Working Example 1

Earlier we discussed efforts to reduce high school violence by enrolling violence-prone students into classes that address anger management. Now, after some time, administrators and managers want to know whether the program is effective. As part of this assessment, students are asked to report their perception of safety at school. An index variable is constructed from different items measuring safety (see Chapter 3). Each item is measured on a seven-point Likert scale (1 = Strongly Disagree to 7 = Strongly Agree), and the index is constructed such that a high value indicates that students feel safe.[10] The survey was initially administered at the beginning of the program. Now, almost a year later, the survey is implemented again.[11]

Administrators want to know whether students who did not participate in the anger management program feel that the climate is now safer. The analysis included here focuses on 10th graders. For practical purposes, the samples of 10th graders at the beginning of the program and one year later are regarded as independent samples; the subjects are not matched. Descriptive analysis shows that the mean perception of safety at the beginning of the program was 4.40 (standard deviation, $SD = 1.00$), and one year later 4.80 ($SD = 0.94$). The mean safety score increased among 10th graders, but is the increase statistically significant? Among other concerns is that the standard deviations are considerable for both samples.

As part of the analysis, we conduct a t-test to answer the question of whether the means of these two distributions are significantly different. First, we examine whether test assumptions are met. The samples are independent, and the variables meet the requirement that one is continuous (the index variable) and the other dichotomous. The assumption of equality of variances is answered as part of conducting the t-test, and so the remaining question is whether the variables are normally distributed. The distributions are shown in the histograms in Figure 11.3:[12]

Are these normal distributions? Visually, they are not the textbook ideal—real-life data seldom are. The Kolmogorov-Smirnov tests for both distributions are insignificant (both $p > .05$). Hence, we conclude that the two distributions can be considered normal. Having satisfied these t-test assumptions, we next conduct the t-test for two independent samples. Table 11.1 shows the t-test results.

The top part of Table 11.1 shows the descriptive statistics, and the bottom part reports the test statistics. Recall that the t-test is a two-step test.

Figure 11.3 Perception of High School Safety among 10th Graders

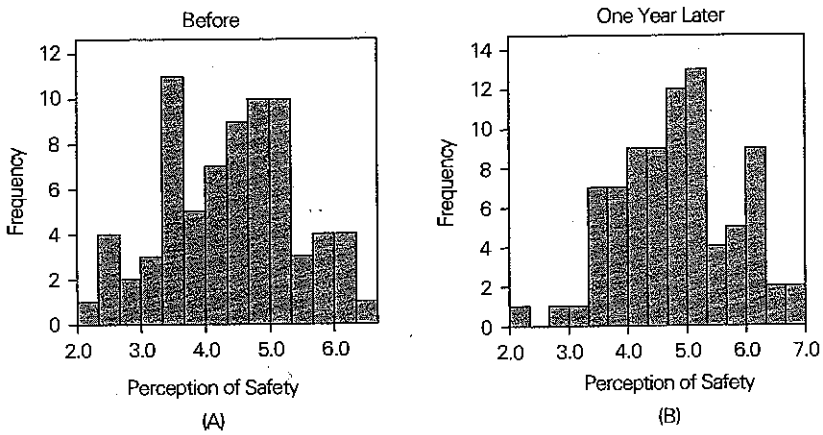


Table 11.1 Independent-Samples T-Test: Output

Group Statistics

Group	N	Mean	SD
One year later	82	4.805	0.962
Before	74	4.399	1.008

High school safety	0.177	0.675	Equal variances assumed	2.576	154	0.011
			Equal variances not assumed	2.570	150.57	0.011

Note: SD = standard deviation.

We first test whether variances are equal. This is shown as the "Levene's test for the equality of variances." The null hypothesis of the Levene's test is that variances are equal; this is rejected when the p-value of this Levene's test statistic is less than .05. The Levene's test uses an F-test statistic (discussed in Chapters 13 and 16), which, other than its p-value, need not concern us

here. In Table 11.1, the level of significance is .675, which exceeds .05. Hence, we accept the null hypothesis—the variances of the two distributions shown in Figures 11.3 are equal.

Now we go to the second step, the main purpose. Are the two means (4.40 and 4.80) significantly different? Because the variances are equal, we read the t-test statistics from the top line, which states "equal variances assumed." (If variances had been unequal, then we would read the test statistics from the second line, "equal variances not assumed."). The t-test statistic for equal variances for this test is 2.576, which is significant at p = .011.[13] Thus, we conclude that the means are significantly different; the 10th graders report feeling safer one year after the anger management program was implemented.

Working Example 2

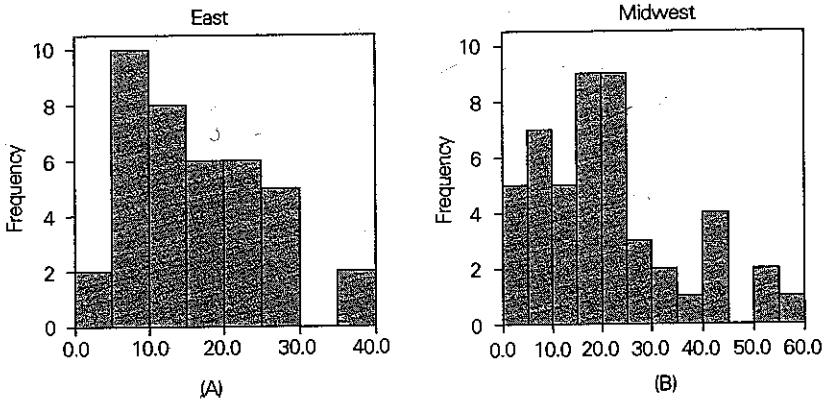
In the preceding example, the variables were both normally distributed, but this is not always the case. Many variables are highly skewed and not normally distributed. Consider another example. The U.S. Environmental Protection Agency (EPA) collects information about the water quality of watersheds, including information about the sources and nature of pollution. One such measure is the percentage of samples that exceed pollution limits for ammonia, dissolved oxygen, phosphorus, and pH.[14] A manager wants to know whether watersheds in the East have higher levels of pollution than those in the Midwest.

An index variable of such pollution is constructed. The index variable is called "Pollution," and the first step is to examine it for test assumptions. Analysis indicates that the range of this variable has a low value of 0.00 percent and a high value of 59.17 percent. These are plausible values (any value above 100.00 percent is implausible). A boxplot (not shown) demonstrates that the variable has two values greater than 50.00 percent that are indicated as outliers for the Midwest region. However, the histograms shown in Figure 11.4 do not suggest that these values are unusually large; rather, the peak in both histograms is located off to the left. The distributions are heavily skewed.[15]

Because the samples each have fewer than 50 observations, the Shapiro-Wilk test for normality is used. The respective test statistics for East and Midwest are .969 (p = .355) and .931 (p = .007). Visual inspection confirms that the Midwest distribution is indeed nonnormal. The Shapiro-Wilk test statistics are given only for completeness; they have no substantive interpretation.

We must now either transform the variable so that it becomes normal for purposes of testing, or use a nonparametric alternative. The second

Figure 11.4 Untransformed Variable: Watershed Pollution



option is discussed later in this chapter. We also show the consequences of ignoring the problem.

To transform the variable, we try the recommended transformations, $\log(x)$, $\sqrt{x}$, or x^2, and then examine the transformed variable for normality. If none of these transformations work, we might modify them, such as using $x^{1/3}$ instead of $x^{1/2}$ (recall that the latter is $\sqrt{x}$).[16] Thus, some experimentation is required. In our case, we find that the $x^{1/2}$ works. The new Shapiro-Wilk test statistics for East and Midwest are, respectively, .969 (p = .361) and .987 (p = .883). Visual inspection of Figure 11.5 shows these two distributions to be quite normal, indeed.

The results of the t-test for the transformed variable are shown in Table 11.2. The transformed variable has equal variances across the two groups (Levene's test, p = .119), and the t-test statistic is –1.308 (df = 85, p = .194). Thus, the differences in pollution between watersheds in the East and Midwest are not significant. (The negative sign of the t-test statistic, –1.308, merely reflects the order of the groups for calculating the difference: the testing variable has a larger value in the Midwest than in the East. Reversing the order of the groups results in a positive sign.)

For comparison, results for the untransformed variable are shown as well. The untransformed variable has unequal variances across the two groups (Levene's test, p = .036), and the t-test statistic is –1.801 (df = 80.6, p =.075). Although this result also shows that differences are insignificant, the level of significance is higher; there are instances in which using nonnormal variables could lead to rejecting the null hypothesis. While our finding

Figure 11.5 ——〰—— Transformed Variable: Watershed Pollution

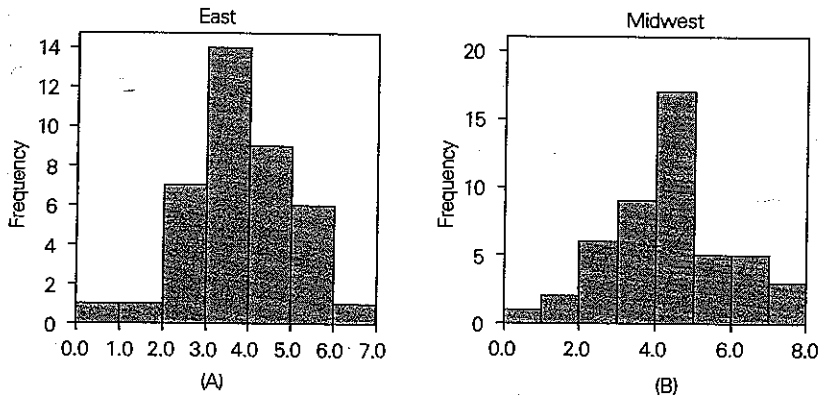


Table 11.2 ——〰—— Independent-Samples T-Test: Output

Variable: watershed pollution

Transformed	2.479	0.119	Equal variances assumed	−1.308	85	0.194
			Equal variances not assumed	−1.347	80.6	0.182
Untransformed	4.537	0.036	Equal variances assumed	−1.723	85	0.089
			Equal variances not assumed	−1.801	80.6	0.075

of insignificant differences is indeed *robust*, analysts cannot know this in advance. Thus, analysts will need to deal with nonnormality.

Variable transformation is one approach to the problem of nonnormality, but transforming variables can be a time-intensive and somewhat artful activity. The search for alternatives has led many analysts to consider nonparametric methods.

TWO T-TEST VARIATIONS

Paired-Samples T-Test

The paired t-test often is used when using before-and-after tests to assess student or client progress. Paired t-tests are used when analysts have a *dependent* rather than an independent sample (see the third t-test assumption, described earlier in this chapter). The ***paired-samples t-test*** tests the null hypothesis that the mean difference between the before and after test scores is zero. Consider the following data from Table 11.3:

Table 11.3 ——〰—— Paired-Samples Data

Before	After	Difference
3.2	4.3	1.1
3.2	4.3	1.1
4.0	3.8	−0.2
2.4	3.5	1.1
3.0	3.3	0.3
4.0	4.4	0.4
4.3	4.2	−0.1
3.8	3.3	−0.5
2.9	3.9	1.0
3.8	4.2	0.4
2.5	3.8	1.3

The mean "before" score is 3.39, and the mean "after" score is 3.87; the mean difference is 0.48. The paired t-test tests the null hypothesis by testing whether the means of the difference variable ("Difference") is zero. The paired t-test test statistic is calculated as

$$t = \frac{\overline{D}}{s_D/\sqrt{n}},$$

where D = the difference between before and after measurements, and s_D is the standard deviation of these differences. Regarding t-test assumptions, the variables are continuous, and the issue of heterogeneity (unequal variances) is moot because this test involves only one variable, D; no Levene's test statistics are produced. We do test the normality of D and find that it is normally distributed (Shapiro-Wilk = .925, p = .402). Thus, the assumptions are satisfied.

We proceed with testing whether the difference between before and after scores is statistically significant. We find that the paired t-test yields a t-test statistic of 2.43, which is significant at the 5 percent level ($df = 9$, $p = .038 < .05$).[17] Hence, we conclude that the increase between the before and after scores is significant at the 5 percent level.[18]

One-Sample T-Test

Finally, the ***one-sample t-test*** tests whether the mean of a single variable is different from a prespecified value (norm). For example, suppose we want to know whether the mean of the before group in Table 11.3 is different from the value of, say, 3.5? Testing against a norm is akin to the purpose of the chi-square goodness-of-fit test described in Chapter 10, but here we are dealing with a continuous variable rather than a categorical one, and we are testing the mean rather than its distribution.

The one-sample t-test assumes that the single variable is continuous and normally distributed. As with the paired t-test, the issue of heterogeneity is moot because there is only one variable. The Shapiro-Wilk test shows that the variable "Before" is normal (.917, $p = .336$). The one-sample t-test statistic for testing against the test value of 3.5 is –0.515 ($df = 9$, $p = .619 > .05$). Hence, the mean of 3.39 is *not* significantly different from 3.5. However, it is different from larger values, such as 4.0 ($t = 2.89$, $df = 9$, $p = .019$).

Finally, note that the one-sample t-test is identical to the paired-samples t-test for testing whether the mean $D = 0$. Indeed, the one-sample t-test for $D = 0$ produces the same results ($t = 2.43$, $df = 9$, $p = .038$).

NONPARAMETRIC ALTERNATIVES TO T-TESTS

The tests described in the preceding sections have nonparametric alternatives. The chief advantage of these tests is that they do not require continuous variables to be normally distributed. The chief disadvantage is that they are less likely to reject the null hypothesis. A further, minor disadvantage is that these tests do not provide descriptive information about variable means; separate analysis is required for that.

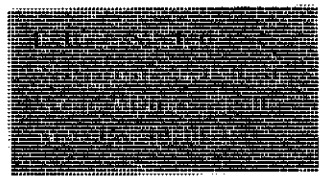
Nonparametric alternatives to the independent samples test are the ***Mann-Whitney*** and ***Wilcoxon tests.*** The Mann-Whitney and Wilcoxon tests are equivalent and are thus discussed jointly. Both are simplifications of the more general Kruskal-Wallis' H test, discussed in Chapter 10.[19] The Mann-Whitney and Wilcoxon tests assign ranks to the testing variable in the exact manner shown in Table 11.4. The sum of the ranks of each group is computed, shown in the table. Then a test is performed of the statistical significance of the difference between the sums, 22.5 and 32.5. Although the Mann-Whitney U and Wilcoxon W test statistics are calculated differently, they both have the same level of statistical significance: $p = .295$. Technically,

Table 11.4 Rankings of Two Groups

Group	Rating	Rank	Group	Rating	Rank
1	2.5	3	2	3.4	7
1	2.9	4	2	3.3	6
1	4.0	9.5	2	4.0	9.5
1	3.2	5	2	3.9	8
1	1.2	1	2	2.1	2
Sum		22.5	Sum		32.5

this is not a test of different means but of different distributions; the lack of significance implies that Groups 1 and 2 can be regarded as coming from the same population.[20]

For comparison, we use the Mann-Whitney test to compare the two samples of 10th graders discussed earlier in this chapter. The sum of ranks for the "before" group is 69.55, and for the "one year later group," 86.57. The test statistic is significant at $p = .019$, yielding the same conclusion as the independent-samples t-test, $p = .011$. This comparison also shows that nonparametric tests do have higher levels of significance. As mentioned earlier, the Mann-Whitney test (as a nonparametric test) does not calculate the group means; separate, descriptive analysis needs to be undertaken for that information.



A nonparametric alternative to the paired-samples t-test is the ***Wilcoxon signed rank test.*** This test assigns ranks based on the absolute values of these differences (Table 11.5). The signs of the differences are retained (thus, some values are positive and others are negative). For the data in Table 11.5, the are seven positive ranks (with mean rank = 6.57) and three negative ranks (with mean rank = 3.00). The Wilcoxon signed rank test statistic is normally distributed. The Wilcoxon signed rank test statistic, Z, for a difference

Table 11.5 Wilcoxon Signed Rank Test

Before	After	Difference	Signed rank
3.2	4.3	1.1	8.5
4.0	3.8	–0.2	–2.0
2.4	3.5	1.1	8.5
3.0	3.3	0.3	3.0
4.0	4.4	0.4	4.5
4.3	4.2	–0.1	–1.0
3.8	3.3	–0.5	–6.0
2.9	3.9	1.0	7.0
3.8	4.2	0.4	4.5
2.5	3.8	1.3	10.0

between these values is 1.89 ($p = .059 > .05$). Hence, according to this test, the differences between the before and after scores are not significant.

Again, nonparametric tests result in larger p-values. The paired-samples t-test finds that $p = .038 < .05$, providing sufficient statistical evidence to conclude that the differences are significant. It might also be noted that a doubling of the data in Table 11.5 results in finding a significant difference between the before and after scores with the Wilcoxon signed rank test, $Z = 2.694$, $p = .007$.

The Wilcoxon signed rank test can also be adapted as a nonparametric alternative to the one-sample t-test. In that case, analysts create a second variable that, for each observation, is the test value. For example, if in Table 11.5 we wish to test whether the mean of variable "Before" is different from, say, 4.0, we create a second variable with 10 observations for which each value is, say, 4.0. Then using the Wilcoxon signed rank test for the "Before" variable and this new, second variable, we find that $Z = 2.103$, $p = .035$. This value is larger than that obtained by the parametric test, $p = .019$.[21]

SUMMARY

When analysts need to determine whether two continuous variables differ in their means, the t-test is the tool of choice. This situations arises, for example, when analysts compare measurements at two points in time or the responses of two different groups. There are three common t-tests, for independent samples, for dependent (paired) samples, and the one-sample t-test.

T-tests are parametric tests, which means that variables in these tests must meet certain assumptions, notably that they are normally distributed. The requirement of normally distributed variables follows from how parametric tests make inferences. Specifically, t-tests have four assumptions:

- One variable is continuous, and the other variable is dichotomous.
- The two distributions have equal variances.
- The observations are independent.
- The two distributions are normally distributed.

The assumption of homogeneous variances does not apply to dependent-samples and one-sample t-tests because both are based on only a single variable for testing significance. When assumptions of normality are not met, variable transformation may be used. The search for alternative ways for dealing with normality problems may lead analysts to consider nonparametric alternatives.

The chief advantage of nonparametric tests is that they do not require continuous variables to be normally distributed. The chief disadvantage is that they yield higher levels of statistical significance, making it less likely that the null hypothesis may be rejected. A nonparametric alternative for the independent-samples t-test is the Mann-Whitney test, and the nonparametric alternative for the dependent-samples t-test is the Wilcoxon signed rank test.

T-tests and their nonparametric alternatives provide information about whether two group means are significantly different. Analysts will need to further assess the magnitude of these differences, and to determine whether they are practically significant. Chapter 16 discusses analysis of variance, or ANOVA, which can be used when means are compared across three or more groups, rather than the two groups of a dichotomous variable.

KEY TERMS

Central limit theorem (p. 185)
Four t-test test assumptions (p. 182)
Heterogeneity of variances (p. 183)
Homogeneity of variances (p. 183)
Independent-samples t-test (p. 181)
Kolmogorov-Smirnov test (p. 184)
Levene's test (p. 183)
Mann-Whitney test (p. 192)
Null hypothesis of normality (p. 184)
One-sample t-test (p. 192)
One-tailed t-tests (p. 182)
Paired-samples t-test (p. 191)
Paired t-test (p. 184)
Robust (p. 183)
Shapiro-Wilk test (p. 184)
Student's t-distribution (p. 182)
T-tests (p. 180)
Two-tailed t-tests (p. 182)
Variable transformation (p. 184)
Wilcoxon signed rank test (p. 193)
Wilcoxon test (p. 192)

Notes

1. Some research suggests that nonparametric tests may not be as robust as thought when variances of groups of rankings are substantially unequal.
2. Boxplots are shown for ease of presentation. It is more appropriate, theoretically, to show two normal distributions, but that clutters the presentation. In any event, continuous data can be presented in boxplots.
3. The formula for the pooled variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

When $s_1 = s_2$, the value of s_p is affected by the relative number of observations in each group, that is, n_1 and n_2. The computer calculates the pooled variance, of course. For more on this topic, see David Howell, *Statistical Methods for Psychology*, 3d ed. (Belmont, Calif.: Duxbury Press, 1992), 181–187.
4. The name *Student's t* is derived from W. S. Gossett, who used "Student" as a pseudonym in the early twentieth century to protect his identity. Legend has it that Gossett was concerned that his employer, an

agro-industrial company, might want to protect the formula as a trade secret because of competitive advantages: the t-test enables very efficient testing of samples.

5. See Box 9.1. The decision to require a higher critical value is not without cost; it could increase Type II errors. However, many analysts prefer to err on the side of caution.
6. Even though studies have shown that men typically have higher incomes than women, this need not always be the case. In any specific setting, in any specific industry, at any point in time, women could have higher incomes.
7. The revised formula for calculating the t-test when variances are unequal is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}.$$

See Howell, *Statistical Methods for Psychology*, for the revised formula for calculating degrees of freedom.

8. Students often want to know how they can transform variables. In most software packages it is simply a matter of specifying something like: newvar = sqrt(oldvar) or newvar = lg10(oldvar). Students also ask what transformation works best. This is largely unknown. It is a matter of trial and error.
9. Some students initially consider variable transformation to be "playing with the data." However, we need to consider that the ancient development of the common measurement scale (1, 2, 3, 4, 5. . .) is as arbitrary as any other scale that might have been chosen (such as 1, 4, 9, 16, 25. . .). The fact that the common measurement scale is frequently useful from the perspective of satisfying test assumptions should not lead us to assign supreme considerations to it or to be reluctant to try other measurement scales that work better in other situations. It is far more important to ensure that the variables are normally distributed for the purpose of test validity.
10. With a Cronbach alpha measure of 0.79, the analyst concludes that the index measure has adequate reliability (see Chapter 3).
11. The data in this example are real, but the reported scenario is fictitious.
12. SPSS readily produces these plots as part of the Analyze → Descriptive Statistics → Explore routine.
13. Software output may also include the 95 percent confidence interval for estimates of the difference. When t-tests are insignificant, the interval will include the value zero, indicating that no difference between the means can be ruled out. When t-tests are significant, the interval will not include the value zero.

14. For more information about this measure, visit www.epa.gov/iwi. See also the Watershed dataset on the CD accompanying the workbook *Exercising Essential Statistics* to replicate the results given here. The index variable is called "conpolut" in the dataset.
15. This conclusion is further indicated by the measures of skewness: East (.519) and Midwest (.912). Based on the test described in Chapter 7, skewness/se(skewness) for the two regions is, respectively, [.559/.378 =] 1.48 and [.912/.343 =] 2.65, which confirms the considerable departure from zero for Midwest. The measures of kurtosis are –.113 and .406.
16. This conclusion is consistent for a wide range of root variable transformations that result in a normal distribution (for example, using $x^{.35}$, not shown, rather than the root variable $\sqrt{x}$.
17. In paired tests, degrees of freedom are defined as $n - 1$ (where n is the number of *pairs* or, equivalently, difference scores).
18. In many t-tests the output includes a 95 percent confidence interval of the difference. This is the range within which we can be 95 percent certain that the population difference lies. For this test, the range is between .032 and .927. Although this is a considerable range, it excludes the value zero, or, no difference of the means.
19. The formula for calculating the Mann-Whitney U test statistic is

$$n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1,$$

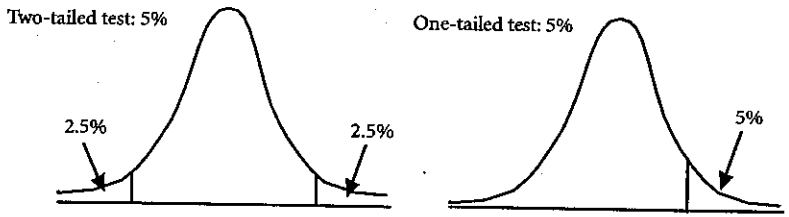
where T_1 = the sum of ranks for Group 1, n_1 = the number of observations in Sample 1, and n_2 = the number of observations in Sample 2. The relationship between U and the Wilcoxon W test statistic is

$$U + W = \frac{m(m+2n+1)}{2}.$$

where m = the number of observations in the group that has the smaller number of observations, and n = the number of observations in the group that has the larger number of observations.

20. By contrast, the p-value for comparing Groups 1 and 3 in Table 10.8 is .016. We may note that using the Kruskal-Wallis' H test for these two groups yields the exact same level of significance.
21. Another, less powerful alternative is the sign test. It is conducted in the same manner as described in the text, but it compares only the number of positive and negative signs rather than the differences of the mean ranks. It is a very crude test that is generally not preferred. In the example in the text, the sign test finds p = .070, indicating that the mean "Before" score is not significantly different from 4.0.

Appendix C
T-Test Distribution



	Alpha Level for One-Tailed Test					
	.10	.05	.025	.01	.005	.0025
Degree of Freedom (df)	Alpha Level for Two-Tailed Test					
	.20	.10	.05	.02	.01	.005
1	3.078	6.314	12.706	31.821	63.657	127.32
2	1.886	2.920	4.303	6.965	9.925	14.089
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.598
5	1.476	2.015	2.571	3.365	4.032	4.773
6	1.440	1.943	2.447	3.143	3.707	4.317
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.869	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
11	1.363	1.796	2.201	2.718	3.106	3.497
12	1.356	1.782	2.179	2.681	3.055	3.428
13	1.350	1.771	2.160	2.650	3.012	3.372
14	1.345	1.761	2.145	2.624	2.977	3.326
15	1.341	1.753	2.131	2.602	2.947	3.286
16	1.337	1.746	2.120	2.583	2.921	3.252
17	1.333	1.740	2.110	2.567	2.898	3.222
18	1.330	1.734	2.101	2.552	2.878	3.197
19	1.328	1.729	2.093	2.539	2.861	3.174
20	1.325	1.725	2.086	2.528	2.845	3.153
21	1.323	1.721	2.080	2.518	2.831	3.135
22	1.321	1.717	2.074	2.508	2.819	3.119
23	1.319	1.714	2.069	2.500	2.807	3.104
24	1.318	1.711	2.064	2.492	2.797	3.091
25	1.316	1.708	2.060	2.485	2.787	3.078

(continued)

	Alpha Level for One-Tailed Test					
	.10	.05	.025	.01	.005	.0025
Degree of Freedom (df)	Alpha Level for Two-Tailed Test					
	.20	.10	.05	.02	.01	.005
26	1.315	1.706	2.056	2.479	2.779	3.067
27	1.314	1.703	2.052	2.473	2.771	3.057
28	1.313	1.701	2.048	2.467	2.763	3.047
29	1.311	1.699	2.045	2.462	2.756	3.038
30	1.310	1.697	2.042	2.457	2.750	3.030
40	1.303	1.684	2.021	2.423	2.704	2.971
60	1.296	1.671	2.000	2.390	2.660	2.915
120	1.289	1.658	1.980	2.358	2.617	2.860
∞	1.282	1.645	1.960	2.326	2.576	2.807

Source: Adapted from Table III of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th edition, Longman Group, Ltd., London, 1974. (Previously published by Oliver & Boyd, Ltd., Edinburgh). Used with permission of the authors and publishers.