CHAPTER 7

Statistical Inference

Benis Arapovic/Shutterstock.com

Learning Objectives

After studying this chapter, you will be able to:

- Explain the purpose of hypothesis testing.
- Explain the difference between the null and alternative hypotheses.
- List the steps in the hypothesis-testing procedure.
- State the proper forms of hypotheses for one-sample hypothesis tests.
- Correctly formulate hypotheses.
- List the four possible outcome results from a hypothesis test.
- Explain the difference between Type I and Type II errors.
- State how to increase the power of a test.
- Choose the proper test statistic for hypothesis tests involving means and proportions.
- Explain how to draw a conclusion for one- and two-tailed hypothesis tests.
- Use *p*-values to draw conclusions about hypothesis tests.
- State the proper forms of hypotheses for two-sample hypothesis tests.
- Select and use Excel *Analysis Toolpak* procedures for two-sample hypothesis tests.
- Explain the purpose of analysis of variance.
- Use the Excel *ANOVA* tool to conduct an analysis of variance test.
- List the assumptions of ANOVA.
- Conduct and interpret the results of a chi-square test for independence.

Managers need to know if the decisions they have made or are planning to make are effective. For example, they might want to answer questions like the following: Did an advertising campaign increase sales? Will product placement in a grocery store make a difference? Did a new assembly method improve productivity or quality in a factory? Many applications of business analytics involve seeking statistical evidence that decisions or process changes have met their objectives. **Statistical inference** focuses on drawing conclusions about populations from samples. Statistical inference includes estimation of population parameters and hypothesis testing, which involves drawing conclusions about the value of the parameters of one or more populations based on sample data. The fundamental statistical approach for doing this is called **hypothesis testing**. Hypothesis testing is a technique that allows you to draw valid statistical conclusions about the value of population parameters or differences among them.

Hypothesis Testing

Hypothesis testing involves drawing inferences about two contrasting propositions (each called a **hypothesis**) relating to the value of one or more population parameters, such as the mean, proportion, standard deviation, or variance. One of these propositions (called the **null hypothesis**) describes the existing theory or a belief that is accepted as valid unless strong statistical evidence exists to the contrary. The second proposition (called the **alternative hypothesis**) is the complement of the null hypothesis; it must be true if the null hypothesis is false. The null hypothesis is denoted by H_0, and the alternative hypothesis is denoted by H_1. Using sample data, we either

1. *reject* the null hypothesis and conclude that the sample data provide sufficient statistical evidence to support the alternative hypothesis, or
2. *fail to reject* the null hypothesis and conclude that the sample data does not support the alternative hypothesis.

If we fail to reject the null hypothesis, then we can only accept as valid the existing theory or belief, but we can never prove it.

EXAMPLE 7.1 A Legal Analogy for Hypothesis Testing

A good analogy for hypothesis testing is the U.S. legal system. In our system of justice, a defendant is innocent until proven guilty. The null hypothesis—our belief in the absence of any contradictory evidence—is not guilty, whereas the alternative hypothesis is guilty. If the evidence (sample data) strongly indicates that the defendant is guilty, then we reject the assumption of innocence. If the evidence is not sufficient to indicate guilt, then we cannot reject the not guilty hypothesis; however, we haven't *proven* that the defendant is innocent. In reality, you can only conclude that a defendant is guilty from the evidence; you still have not proven it!

Hypothesis-Testing Procedure

Conducting a hypothesis test involves several steps:

1. Identifying the population parameter of interest and formulating the hypotheses to test
2. Selecting a *level of significance*, which defines the risk of drawing an incorrect conclusion when the assumed hypothesis is actually true
3. Determining a decision rule on which to base a conclusion
4. Collecting data and calculating a test statistic
5. Applying the decision rule to the test statistic and drawing a conclusion

We apply this procedure to two different types of hypothesis tests; the first involving a single population (called one-sample tests) and, later, tests involving more than one population (multiple-sample tests).

One-Sample Hypothesis Tests

A **one-sample hypothesis test** is one that involves a single population parameter, such as the mean, proportion, standard deviation, and so on. To conduct the test, we use a single sample of data from the population. We may conduct three types of one-sample hypothesis tests:

H_0: population parameter $\geq$ constant vs. H_1: population parameter $<$ constant

H_0: population parameter $\leq$ constant vs. H_1: population parameter $>$ constant

H_0: population parameter $=$ constant vs. H_1: population parameter $\neq$ constant

Notice that one-sample tests always compare a population parameter to some constant. For one-sample tests, the statements of the null hypotheses are expressed as either $\geq$, $\leq$, or $=$. It is *not correct* to formulate a null hypothesis using $>$, $<$, or $\neq$.

How do we determine the proper form of the null and alternative hypotheses? Hypothesis testing always *assumes* that H_0 is true and uses sample data to determine whether H_1 is more likely to be true. Statistically, we cannot "prove" that H_0 is true; we can only *fail to reject* it. Thus, if we cannot reject the null hypothesis, we have shown only that there is insufficient evidence to conclude that the alternative hypothesis is true. However, rejecting the null hypothesis provides strong evidence (in a statistical sense) that the null hypothesis is not true and that the alternative hypothesis is true. Therefore, what we wish to provide evidence for statistically should be identified as the alternative hypothesis.

EXAMPLE 7.2 Formulating a One-Sample Test of Hypothesis

CadSoft, a producer of computer-aided design software for the aerospace industry receives numerous calls for technical support. In the past, the average response time has been at least 25 minutes. The company has upgraded its information systems and believes that this will help reduce response time. As a result, it believes that the average response time can be reduced to less than 25 minutes. The company collected a sample of 44 response times in the Excel file *CadSoft Technical Support Response Times* (see Figure 7.1).

Figure 7.1

Portion of Technical Support Response-Time Data

	A	B	C	D	E
1	CadSoft Technical Support Response Times				
2					
3	Customer	Time (min)			
4	1	20			
5	2	12			
6	3	15			
7	4	11			
8	5	22			
9	6	6			
10	7	39			

If the new information system makes a difference, then data should be able to confirm that the mean response time is less than 25 minutes; this defines the alternative hypothesis, H_1.

Therefore, the proper statements of the null and alternative hypotheses are:

H_0: population mean response time $\geq$ 25 minutes

H_1: population mean response time $<$ 25 minutes

We would typically write this using the proper symbol for the population parameter. In this case, letting μ be the mean response time, we would write:

$$H_0: \mu \geq 25$$
$$H_1: \mu < 25$$

Understanding Potential Errors in Hypothesis Testing

We already know that sample data can show considerable variation; therefore, conclusions based on sample data may be wrong. Hypothesis testing can result in one of four different outcomes:

1. The null hypothesis is actually *true*, and the test *correctly fails to reject it*.
2. The null hypothesis is actually *false*, and the hypothesis test *correctly reaches this conclusion*.
3. The null hypothesis is actually *true*, but the hypothesis test *incorrectly rejects it* (called **Type I error**).
4. The null hypothesis is actually *false*, but the hypothesis test *incorrectly fails to reject it* (called **Type II error**).

The probability of making a Type I error, that is, $P(\text{rejecting } H_0 | H_0 \text{ is true})$, is denoted by α and is called the **level of significance**. This defines the likelihood that you are willing to take in making the incorrect conclusion that the alternative hypothesis is true when, in fact, the null hypothesis is true. The value of α can be controlled by the decision maker and is selected before the test is conducted. Commonly used levels for α are 0.10, 0.05, and 0.01.

The probability of *correctly failing to reject* the null hypothesis, or $P(\text{not rejecting } H_0 | H_0 \text{ is true})$, is called the **confidence coefficient** and is calculated as $1 - \alpha$. For a confidence coefficient of 0.95, we mean that we expect 95 out of 100 samples to support the null hypothesis rather than the alternate hypothesis when H_0 is actually true.

Unfortunately, we cannot control the probability of a Type II error, $P(\text{not rejecting } H_0 | H_0 \text{ is false})$, which is denoted by β. Unlike α, β cannot be specified in advance but depends on the true value of the (unknown) population parameter.

EXAMPLE 7.3 How β Depends on the True Population Mean

Consider the hypotheses in the CadSoft example:

H_0: mean response time $\geq$ 25 minutes

H_1: mean response time $<$ 25 minutes

If the true mean response from which the sample is drawn is, say, 15 minutes, we would expect to have a much smaller probability of incorrectly concluding that the null hypothesis is true than when the true mean response is 24 minutes, for example. If the true mean were 15 minutes, the sample mean would very likely be much less than 25, leading us to reject H_0. If the true mean were 24 minutes, even though it is less than 25, we would have a much higher probability of failing to reject H_0 because a higher likelihood exists that the sample mean would be greater than 25 due to sampling error. Thus, the farther away the true mean response time is from the hypothesized value, the smaller is β. Generally, as α decreases, β increases, so the decision maker must consider the trade-offs of these risks. So, if you choose a level of significance of 0.01 instead of 0.05 and keep the sample size constant, you would reduce the probability of a Type I error but increase the probability of a Type II error.

The value $1 - \beta$ is called the **power of the test** and represents the probability of *correctly rejecting* the null hypothesis when it is indeed false, or $P(\text{rejecting } H_0 | H_0 \text{ is false})$. We would like the power of the test to be high (equivalently, we would like the probability of a Type II error to be low) to allow us to make a valid conclusion. The power of the test is sensitive to the sample size; small sample sizes generally result in a low value of $1 - \beta$. The power of the test can be increased by taking larger samples, which enable us to detect small differences between the sample statistics and population parameters with more accuracy. However, a larger sample size incurs higher costs, giving new meaning to the adage, there is no such thing as a free lunch. This suggests that if you choose a small level of significance, you should try to compensate by having a large sample size when you conduct the test.

Selecting the Test Statistic

The next step is to collect sample data and use the data to draw a conclusion. The decision to reject or fail to reject a null hypothesis is based on computing a *test statistic* from the sample data. The test statistic used depends on the type of hypothesis test. Different types of hypothesis tests use different test statistics, and it is important to use the correct one. The proper test statistic often depends on certain assumptions about the population—for example, whether or not the standard deviation is known. The following formulas show two types of one-sample hypothesis tests for means and their associated test statistics. The value of μ_0 is the hypothesized value of the population mean; that is, the "constant" in the hypothesis formulation.

Type of Test	Test Statistic	
One-sample test for mean, σ known	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	(7.1)
One-sample test for mean, σ unknown	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	(7.2)

EXAMPLE 7.4 Computing the Test Statistic

For the CadSoft example, the average response time for the sample of 44 customers is $\bar{x} = 21.91$ minutes and the sample standard deviation is $s = 19.49$. The hypothesized mean is $\mu_0 = 25$. You might wonder why we even have to test the hypothesis statistically when the sample average of 21.91 is clearly less than 25. The reason is because of sampling error. It is quite possible that the population mean truly is 25 or more and that we were just lucky to draw a sample whose mean was smaller. Because of potential sampling error, it would be dangerous to conclude that the company was meeting its goal just by looking at the sample mean without better statistical evidence.

Because we don't know the value of the population standard deviation, the proper test statistic to use is formula (7.2):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Therefore, the value of the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.91 - 25}{19.49/\sqrt{44}} = \frac{-3.09}{2.938} = -1.05$$

Observe that the numerator is the distance between the sample mean (21.91) and the hypothesized value (25). By dividing by the standard error, the value of t represents the number of standard errors the sample mean is from the hypothesized value. In this case, the sample mean is 1.05 standard errors below the hypothesized value of 25. This notion provides the fundamental basis for the hypothesis test—if the sample mean is "too far" away from the hypothesized value, then the null hypothesis should be rejected.

Drawing a Conclusion

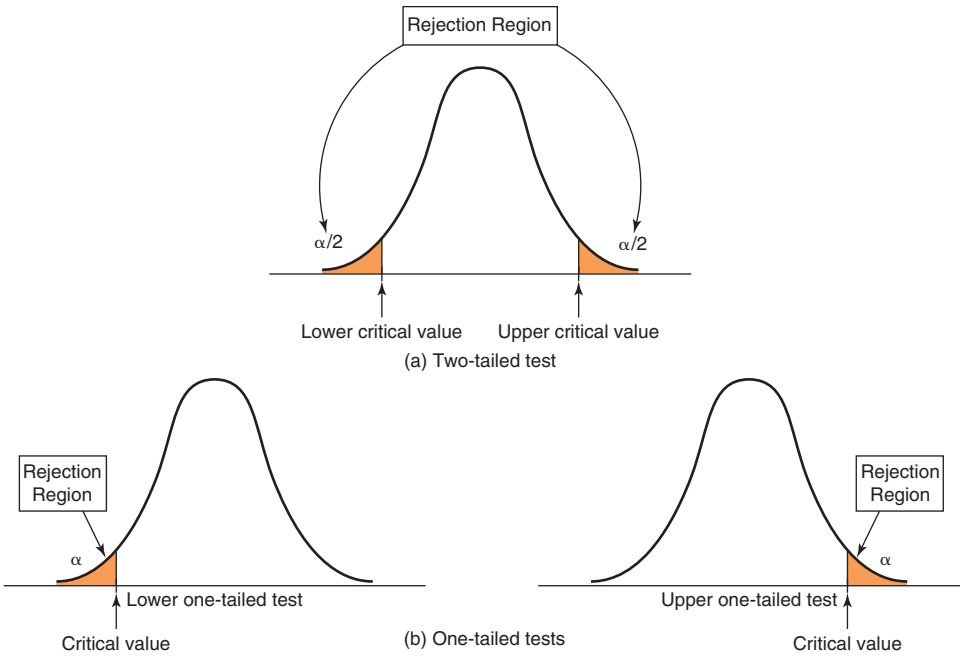
The conclusion to reject or fail to reject H_0 is based on comparing the value of the test statistic to a "critical value" from the sampling distribution of the test statistic when the null hypothesis is true and the chosen level of significance, α. The sampling distribution of the test statistic is usually the normal distribution, t-distribution, or some other well-known distribution. For example, the sampling distribution of the z-test statistic in formula (7.1) is a standard normal distribution; the t-test statistic in formula (7.2) has a t-distribution with $n - 1$ degrees of freedom. For a one-tailed test, the critical value is the number of standard errors away from the hypothesized value for which the probability of exceeding the critical value is α. If $\alpha = 0.05$, for example, then we are saying that there is only a 5% chance that a sample mean will be that far away from the hypothesized value purely because of sampling error and should this occur, it suggests that the true population mean is different from what was hypothesized.

The critical value divides the sampling distribution into two parts, a *rejection region* and a *nonrejection region*. If the null hypothesis is false, it is more likely that the test statistic will fall into the rejection region. If it does, we reject the null hypothesis; otherwise, we fail to reject it. The rejection region is chosen so that the probability of the test statistic falling into it if H_0 is true is the probability of a Type I error, α.

The rejection region occurs in the tails of the sampling distribution of the test statistic and depends on the structure of the hypothesis test, as shown in Figure 7.2. If the null hypothesis is structured as $=$ and the alternative hypothesis as $\neq$, then we would reject H_0 if the test statistic is *either* significantly high or low. In this case, the rejection region will occur in *both* the upper and lower tail of the distribution [see Figure 7.2(a)]. This is called a **two-tailed test of hypothesis**. Because the probability that the test statistic falls into the rejection region, given that H_0 is true, the combined area of both tails must be α; each tail has an area of $\alpha/2$.

Figure 7.2

Illustration of Rejection Regions in Hypothesis Testing



The other types of hypothesis tests, which specify a direction of relationship (where H_0 is either $\geq$ or $\leq$), are called **one-tailed tests of hypothesis**. In this case, the rejection region occurs only in one tail of the distribution [see Figure 7.2(b)]. Determining the correct tail of the distribution to use as the rejection region for a one-tailed test is easy. If H_1 is stated as $<$, the rejection region is in the lower tail; if H_1 is stated as $>$, the rejection region is in the upper tail (just think of the inequality as an arrow pointing to the proper tail direction).

Two-tailed tests have both upper and lower critical values, whereas one-tailed tests have either a lower or upper critical value. For standard normal and *t*-distributions, which have a mean of zero, lower-tail critical values are negative; upper-tail critical values are positive.

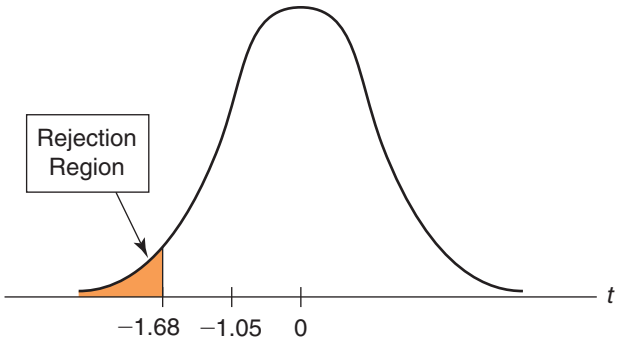
Critical values make it easy to determine whether or not the test statistic falls in the rejection region of the proper sampling distribution. For example, for an upper one-tailed test, if the test statistic is greater than the critical value, the decision would be to reject the null hypothesis. Similarly, for a lower one-tailed test, if the test statistic is less than the critical value, we would reject the null hypothesis. For a two-tailed test, if the test statistic is *either* greater than the upper critical value or less than the lower critical value, the decision would be to reject the null hypothesis.

EXAMPLE 7.5 Finding the Critical Value and Drawing a Conclusion

For the CadSoft example, if the level of significance is 0.05, then the critical value for a one-tail test is the value of the *t*-distribution with $n - 1$ degrees of freedom that provides a tail area of 0.05, that is, $t_{\alpha,n-1}$. We may find *t*-values in Table A.2 in Appendix A at the end of the book or by using the Excel function T.INV($1 - \alpha$, $n - 1$). Thus, the critical value is $t_{0.05,43}$ = T.INV(0.95,43) = 1.68. Because the *t*-distribution is symmetric with a mean of 0 and this is a lower-tail test, we use the negative of this number (– 1.68) as the critical value.

Figure 7.3

t-Test for Mean Response Time



By comparing the value of the *t*-test statistic with this critical value, we see that the test statistic does not fall below the critical value (i.e., $-1.05 > -1.68$) and is not in the rejection region. Therefore, we cannot reject H_0 and cannot conclude that the mean response time has improved to less than 25 minutes. Figure 7.3 illustrates the conclusion we reached. Even though the sample mean is less than 25, we cannot conclude that the population mean response time is less than 25 because of the large amount of sampling error.

Two-Tailed Test of Hypothesis for the Mean

Basically, all hypothesis tests are similar; you just have to ensure that you select the correct test statistic, critical value, and rejection region, depending on the type of hypothesis. The following example illustrates a two-tailed test of hypothesis for the mean.

EXAMPLE 7.6 Conducting a Two-Tailed Hypothesis Test for the Mean

Figure 7.4 shows a portion of data collected in a survey of 34 respondents by a travel agency (provided in the Excel file *Vacation Survey*). Suppose that the travel agency wanted to target individuals who were approximately 35 years old. Thus, we wish to test whether the average age of respondents is equal to 35. The hypothesis to test is

$$H_0: \text{mean age} = 35$$

$$H_1: \text{mean age} \neq 35$$

The sample mean is computed to be 38.677, and the sample standard deviation is 7.858.

We use the *t*-test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{38.677 - 35}{7.858/\sqrt{34}} = 2.73$$

In this case, the sample mean is 2.73 standard errors above the hypothesized mean of 35. However, because this is a two-tailed test, the rejection region and decision rule are different. For a level of significance α, we reject H_0 if the *t*-test statistic falls either below the negative critical value, $-t_{\alpha/2,n-1}$, or above the positive critical value, $t_{\alpha/2,n-1}$. Using either Table A.2 in Appendix A at the back of this book or the Excel function T.INV.2T(.05,33) to calculate $t_{0.025,33}$, we obtain 2.0345. Thus, the critical values are ±2.0345. Because the *t*-test statistic does *not* fall between these values, we must reject the null hypothesis that the average age is 35 (see Figure 7.5).

p-Values

An alternative approach to comparing a test statistic to a critical value in hypothesis testing is to find the probability of obtaining a test statistic value equal to or more extreme than that obtained from the sample data when the null hypothesis is true. This probability

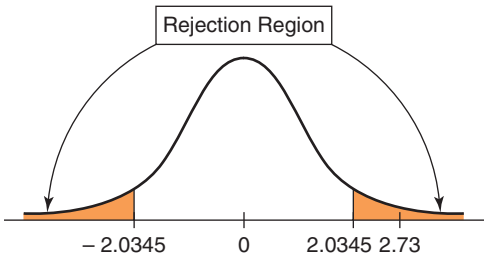
Figure 7.4

Portion of *Vacation Survey* Data

	A	B	C	D	E
1	Vacation Survey				
2					
3	Age	Gender	Relationship Status	Vacations per Year	Number of Children
4	24	Male	Married	2	0
5	26	Female	Married	4	0
6	28	Male	Married	2	2
7	33	Male	Married	4	0
8	45	Male	Married	2	0
9	49	Male	Married	1	2
10	29	Male	Married	4	0

Figure 7.5

Illustration of a Two-Tailed Test for Example 7.6



is commonly called a ***p*-value**, or **observed significance level**. To draw a conclusion, compare the *p*-value to the chosen level of significance α; whenever $p < \alpha$, reject the null hypothesis and otherwise fail to reject it. *p*-Values make it easy to draw conclusions about hypothesis tests. For a lower one-tailed test, the *p*-value is the probability to the left of the test statistic t in the t-distribution, and is found by T.DIST(t, $n - 1$, TRUE). For an upper one-tailed test, the *p*-value is the probability to the right of the test statistic t, and is found by $1 -$ T.DIST(t, $n - 1$, TRUE). For a two-tailed test, the *p*-value is found by T.DIST.2T (t, $n - 1$), if $t > 0$; if $t < 0$, use T.DIST.2T($-t$, $n - 1$).

EXAMPLE 7.7 Using *p*-Values

For the CadSoft example, the *t*-test statistic for the hypothesis test in the response-time example is -1.05. If the true mean is really 25, then the *p*-value is the probability of obtaining a test statistic of -1.05 or less (the area to the left of -1.05 in Figure 7.3). We can calculate the *p*-value using the Excel function T.DIST(−1.05,43,TRUE) = 0.1498. Because $p = 0.1498$ is not less than $\alpha = 0.05$, we do not reject H_0. In other words, there is about a 15% chance that the test statistic would be -1.05 or smaller if the null hypothesis were true. This is a fairly high probability, so it would be difficult to conclude that the true mean is less than 25 and we could attribute the fact that the test statistic is less than the hypothesized value to sampling error alone and not reject the null hypothesis.

For the Vacation Survey two-tailed hypothesis test in Example 7.6, the *p*-value for this test is 0.010, which can also be computed by the Excel function T.DIST.2T(2.73,33); therefore, since $0.010 < 0.05$, we reject H_0.

One-Sample Tests for Proportions

Many important business measures, such as market share or the fraction of deliveries received on time, are expressed as proportions. We may conduct a test of hypothesis about a population proportion in a similar fashion as we did for means. The test statistic for a one-sample test for proportions is

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} \tag{7.3}$$

where π_0 is the hypothesized value and $\hat{p}$ is the sample proportion. Similar to the test statistic for means, the z-test statistic shows the number of standard errors that the sample proportion is from the hypothesized value. The sampling distribution of this test statistic has a standard normal distribution.

EXAMPLE 7.8 A One-Sample Test for the Proportion

CadSoft also sampled 44 customers and asked them to rate the overall quality of the company's software product using a scale of

0—very poor
1—poor
2—good
3—very good
4—excellent

These data can be found in the Excel File *CadSoft Product Satisfaction Survey*. The firm tracks customer satisfaction of quality by measuring the proportion of responses in the top two categories. Over the past, this proportion has averaged about 75%. For these data, 35 of the 44 responses, or 79.5%, are in the top two categories. Is there sufficient evidence to conclude that this satisfaction measure has significantly exceeded 75% using a significance level of 0.05? Answering this question involves testing the hypotheses about the population proportion π:

$$H_0: \pi \leq 0.75$$
$$H_1: \pi > 0.75$$

This is an upper-tailed, one-tailed test. The test statistic is computed using formula (7.3):

$$z = \frac{0.795 - 0.75}{\sqrt{0.75(1 - 0.75)/44}} = 0.69$$

In this case, the sample proportion of 0.795 is 0.69 standard error above the hypothesized value of 0.75. Because this is an upper-tailed test, we reject H_0 if the value of the test statistic is larger than the critical value. Because the sampling distribution of z is a standard normal, the critical value of z for a level of significance of 0.05 is found by the Excel function NORM.S.INV(0.95) = 1.645. Because the test statistic does not exceed the critical value, we cannot reject the null hypothesis that the proportion is no greater than 0.75. Thus, even though the sample proportion exceeds 0.75, we cannot conclude statistically that the customer satisfaction ratings have significantly improved. We could attribute this to sampling error and the relatively small sample size. The p-value can be found by computing the area to the right of the test statistic in the standard normal distribution: 1 – NORM.S.DIST(0.69,TRUE) = 0.24. Note that the p-value is greater than the significance level of 0.05, leading to the same conclusion of not rejecting the null hypothesis.

For a lower-tailed test, the p-value would be computed by the area to the left of the test statistic; that is, NORM.S.DIST(z, TRUE). If we had a two-tailed test, the p-value is 2*NORM.S.DIST(z, TRUE) if $z < 0$; otherwise, the p-value is 2*(1−NORM.S.DIST($-z$, TRUE)) if $z > 0$.

Confidence Intervals and Hypothesis Tests

A close relationship exists between confidence intervals and hypothesis tests. For example, suppose we construct a 95% confidence interval for the mean. If we wish to test the hypotheses

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

at a 5% level of significance, we simply check whether the hypothesized value μ_0 falls within the confidence interval. If it does not, then we reject H_0; if it does, then we cannot reject H_0.

For one-tailed tests, we need to examine on which side of the hypothesized value the confidence interval falls. For a lower-tailed test, if the confidence interval falls entirely below the hypothesized value, we reject the null hypothesis. For an upper-tailed test, if the confidence interval falls entirely above the hypothesized value, we also reject the null hypothesis.

Two-Sample Hypothesis Tests

Many practical applications of hypothesis testing involve comparing two populations for differences in means, proportions, or other population parameters. Such tests can confirm differences between suppliers, performance at two different factory locations, new and old work methods or reward and recognition programs, and many other situations. Similar to one-sample tests, two-sample hypothesis tests for differences in population parameters have one of the following forms:

1. *Lower-tailed test* H_0: population parameter (1) − population parameter (2) $\geq D_0$ vs. H_1: population parameter (1) − population parameter (2) $< D_0$. This test seeks evidence that the difference between population parameter (1) and population parameter (2) is less than some value, D_0. When $D_0 = 0$, the test simply seeks to conclude whether population parameter (1) is smaller than population parameter (2).
2. *Upper-tailed test* H_0: population parameter (1) − population parameter (2) $\leq D_0$ vs. H_1: population parameter (1) − population parameter (2) $> D_0$. This test seeks evidence that the difference between population parameter (1) and population parameter (2) is greater than some value, D_0. When $D_0 = 0$, the test simply seeks to conclude whether population parameter (1) is larger than population parameter (2).
3. *Two-tailed test* H_0: population parameter (1) − population parameter (2) $= D_0$ vs. H_1: population parameter (1) − population parameter (2) $\neq D_0$. This test seeks evidence that the difference between the population parameters is equal to D_0. When $D_0 = 0$, we are seeking evidence that population parameter (1) differs from parameter (2).

In most applications $D_0 = 0$, and we are simply seeking to compare the population parameters. However, there are situations when we might want to determine if the parameters differ by some non-zero amount; for example, "job classification A makes at least $5,000 more than job classification B."

The hypothesis-testing procedures are similar to those previously discussed in the sense of computing a test statistic and comparing it to a critical value. However, the test statistics for two-sample tests are more complicated than for one-sample tests and we will not delve into the mathematical details. Fortunately, Excel provides several tools for conducting two-sample tests, and we will use these in our examples. Table 7.1 summarizes the Excel *Analysis Toolpak* procedures that we will use.

Two-Sample Tests for Differences in Means

In a two-sample test for differences in means, we always test hypotheses of the form

$$H_0: \mu_1 - \mu_2 \{\geq, \leq, \text{or} =\} 0$$
$$H_1: \mu_1 - \mu_2 \{<, >, \text{or} \neq\} 0 \qquad \textbf{(7.4)}$$

Table 7.1

Excel *Analysis Toolpak* Procedures for Two-Sample Hypothesis Tests

Type of Test	Excel Procedure
Two-sample test for means, σ^2 known	Excel *z*-test: Two-sample for means
Two-sample test for means, σ^2 unknown, assumed unequal	Excel *t*-test: Two-sample assuming unequal variances
Two-sample test for means, σ^2 unknown, assumed equal	Excel *t*-test: Two-sample assuming equal variances
Paired two-sample test for means	Excel *t*-test: Paired two-sample for means
Two-sample test for equality of variances	Excel *F*-test Two-sample for variances

EXAMPLE 7.9 Comparing Supplier Performance

The last two columns in the *Purchase Orders* data file provide the order date and arrival date of all orders placed with each supplier. The time between placement of an order and its arrival is commonly called the lead time. We may compute the lead time by subtracting the Excel date function values from each other (Arrival Date − Order Date), as shown in Figure 7.6.

Figure 7.7 shows a pivot table for the average lead time for each supplier. Purchasing managers have noted that they order many of the same types of items from Alum Sheeting and Durrable Products and are considering dropping Alum Sheeting from its supplier base if its lead time is significantly longer than that of Durrable Products. Therefore, they would like to test the hypothesis

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

where μ_1 = mean lead time for Alum Sheeting and μ_2 = mean lead time for Durrable Products.

Rejecting the null hypothesis suggests that the average lead time for Alum Sheeting is statistically longer than Durrable Products. However, if we cannot reject the null hypothesis, then even though the mean lead time for Alum Sheeting is longer, the difference would most likely be due to sampling error, and we could not conclude that there is a statistically significant difference.

Selection of the proper test statistic and Excel procedure for a two-sample test for means depends on whether the population standard deviations are known, and if not, whether they are assumed to be equal.

1. *Population variance is known.* In Excel, choose *z-Test: Two-Sample for Means* from the *Data Analysis* menu. This test uses a test statistic that is based on the standard normal distribution.
2. *Population variance is unknown and assumed unequal.* From the *Data Analysis* menu, choose *t-test: Two-Sample Assuming Unequal Variances*. The test statistic for this case has a *t*-distribution.

Figure 7.6

Portion of *Purchase Orders* Database with Lead Time Calculations

	A	B	C	D	E	F	G	H	I	J	K
1	Purchase Orders										
2											
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date	Lead Time
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11	8
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11	6
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11	5
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11	7
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11	11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11	6
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11	7

Figure 7.7

Pivot Table for Average Supplier Lead Time

	A	B
1		
2		
3	Row Labels	Average of Lead Time
4	Alum Sheeting	7.00
5	Durrable Products	4.92
6	Fast-Tie Aerospace	8.47
7	Hulkey Fasteners	6.47
8	Manley Valve	6.45
9	Pylon Accessories	8.00
10	Spacetime Technologies	15.25
11	Steelpin Inc.	10.20
12	Grand Total	8.41

3. *Population variance unknown but assumed equal.* In Excel, choose *t-test: Two-Sample Assuming Equal Variances.* The test statistic also has a *t*-distribution, but it is different from the unequal variance case.

These tools calculate the test statistic, the *p*-value for both a one-tail and two-tail test, and the critical values for one-tail and two-tail tests. For the *z*-test with known population variances, these are called *z*, $P(Z \leq z)$ *one-tail* or $P(Z \leq z)$ *two-tail*, and *z Critical one-tail* or *z Critical two-tail*, respectively. For the *t*-tests, these are called *t Stat*, $P(T \leq t)$ *one-tail* or $P(T \leq t)$ *two-tail*, and *t Critical one-tail* or *t Critical two-tail*, respectively.

Caution: You must be *very careful* in interpreting the output information from these Excel tools and apply the following rules:

1. If the test statistic is negative, the one-tailed *p*-value is the correct *p*-value for a lower-tail test; however, for an upper-tail test, you must subtract this number from 1.0 to get the correct *p*-value.
2. If the test statistic is nonnegative (positive or zero), then the *p*-value in the output is the correct *p*-value for an upper-tail test; but for a lower-tail test, you must subtract this number from 1.0 to get the correct *p*-value.
3. For a lower-tail test, you must change the sign of the one-tailed critical value.

Only rarely are the population variances known; also, it is often difficult to justify the assumption that the variances of each population are equal. Therefore, in most practical situations, we use the *t-test: Two-Sample Assuming Unequal Variances*. This procedure also works well with small sample sizes if the populations are approximately normal. It is recommended that the size of each sample be approximately the same and total 20 or more. If the populations are highly skewed, then larger sample sizes are recommended.

EXAMPLE 7.10 Testing the Hypotheses for Supplier Lead-Time Performance

To conduct the hypothesis test for comparing the lead times for Alum Sheeting and Durrable Products, first sort the data by supplier and then select *t-test: Two-Sample Assuming Unequal Variances* from the *Data Analysis* menu. The dialog is shown in Figure 7.8. The dialog prompts you for the range of the data for each variable, hypothesized mean difference, whether the ranges have labels, and the level of significance α. If you leave the box *Hypothesized Mean Difference* blank or enter zero, the test is for equality of means. However, the tool allows you to specify a value D_0 to test the hypothesis H_0: $\mu_1 - \mu_2 = D_0$ if you want to test whether the population means have a certain distance between them. In this example, the *Variable 1* range defines the lead times for Alum Sheeting, and the *Variable 2* range for Durrable Products.

Figure 7.9 shows the results from the tool. The tool provides information for both one-tailed and two-tailed tests. Because this is a one-tailed test, we use the

highlighted information in Figure 7.9 to draw our conclusions. For this example, *t Stat* is positive and we have an upper-tailed test; therefore using the rules stated earlier, the *p*-value is 0.00166. Based on this alone, we reject the null hypothesis and must conclude that Alum Sheeting has a statistically longer average lead time than Durrable Products. We may draw the same conclusion by comparing the value of *t Stat* with the critical value *t Critical one-tail*. Being an upper-tail test, the value of *t Critical one-tail* is 1.812. Comparing this with the value of *t Stat*, we would reject H_0 only if *t Stat* > *t Critical one-tail*. Since *t Stat* is greater than *t Critical one-tail*, we reject the null hypothesis.

Two-Sample Test for Means with Paired Samples

In the previous example for testing differences in the mean supplier lead times, we used independent samples; that is, the orders in each supplier's sample were not related to each other. In many situations, data from two samples are naturally paired or matched. For example, suppose that a sample of assembly line workers perform a task using two different types of work methods, and the plant manager wants to determine if any differences exist between the two methods. In collecting the data, each worker will have performed the task using each method. Had we used independent samples, we would have randomly selected two different groups of employees and assigned one work method to one group and the alternative method to the second group. Each worker would have performed the task using only one of the methods. As another example, suppose that we wish to compare retail prices of grocery items between two competing grocery stores. It makes little sense to compare different samples of items from each store. Instead, we would select a sample of grocery items and

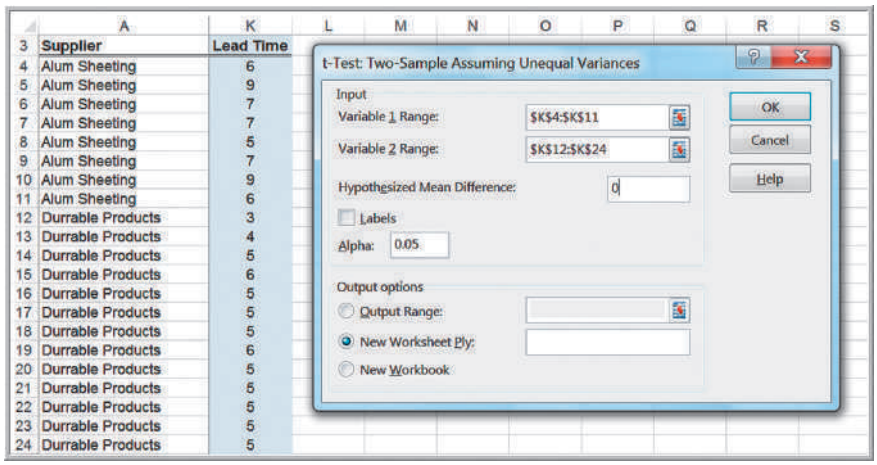


Figure 7.8

Dialog for Two-Sample *t*-Test, Sigma Unknown

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2		Alum Sheeting	Durrable Products
3		*Variable 1*	*Variable 2*
4	Mean	7	4.923076923
5	Variance	2	0.576923077
6	Observations	8	13
7	Hypothesized Mean Difference	0	
8	df	10	
9	t Stat	3.827958507	
10	P(T<=t) one-tail	0.001664976	
11	t Critical one-tail	1.812461123	
12	P(T<=t) two-tail	0.003329952	
13	t Critical two-tail	2.228138852	

Figure 7.9

Results for Two-Sample Test for Lead-Time Performance

find the price charged for the same items by each store. In this case, the samples are paired because each item would have a price from each of the two stores.

When paired samples are used, a paired *t*-test is more accurate than assuming that the data come from independent populations. The null hypothesis we test revolves around the mean difference (μ_D) between the paired samples; that is

$$H_0: \mu_D \{\geq, \leq, \text{or } =\} 0$$

$$H_1: \mu_D \{<, >, \text{or } \neq\} 0.$$

The test uses the average difference between the paired data and the standard deviation of the differences similar to a one-sample test.

Excel has a *Data Analysis* tool, *t-Test: Paired Two-Sample for Means* for conducting this type of test. In the dialog, you need to enter only the variable ranges and hypothesized mean difference.

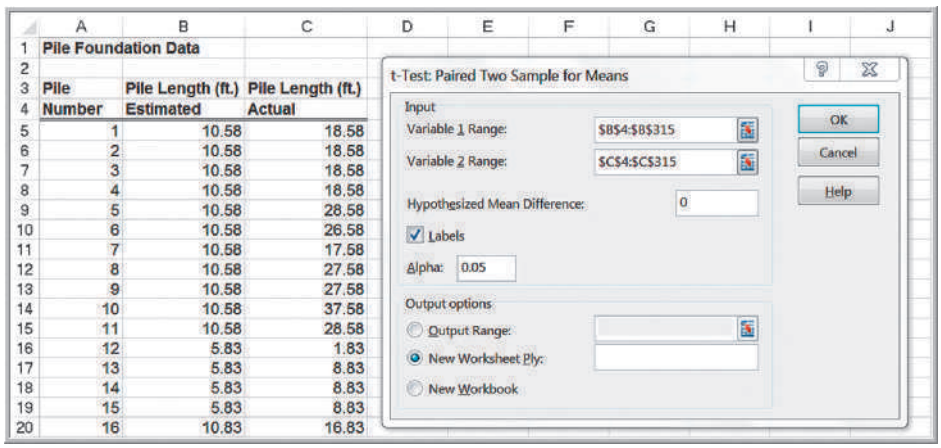
EXAMPLE 7.11 Using the Paired Two-Sample Test for Means

The Excel file *Pile Foundation* contains the estimates used in a bid and actual auger-cast pile lengths that engineers ultimately had to use for a foundation-engineering project. The contractor's past experience suggested that the bid information was generally accurate, so the average of the paired differences between the actual pile lengths and estimated lengths should be close to zero. After this project was completed, the contractor found that the average difference between the actual lengths and the estimated lengths was 6.38. Could the contractor conclude that the bid information was poor?

Figure 7.10 shows a portion of the data and the Excel dialog for the paired two-sample test. Figure 7.11 shows the output from the Excel tool using a significance level of 0.05, where *Variable 1* is the estimated lengths, and *Variable 2* is the actual lengths. This is a two-tailed test, so in Figure 7.11 we interpret the results using only the two-tail information that is highlighted. The critical values are ±1.968, and because *t Stat* is much smaller than the lower critical value, we must reject the null hypothesis and conclude that the mean of the differences between the estimates and the actual pile lengths is statistically significant. Note that the *p*-value is essentially zero, verifying this conclusion.

Test for Equality of Variances

Understanding variation in business processes is very important, as we have stated before. For instance, does one location or group of employees show higher variability than others? We can test for equality of variances between two samples using a new type of test,



	A	B	C
1	Pile Foundation Data		
2			
3	Pile	Pile Length (ft.)	Pile Length (ft.)
4	Number	Estimated	Actual
5	1	10.58	18.58
6	2	10.58	18.58
7	3	10.58	18.58
8	4	10.58	18.58
9	5	10.58	28.58
10	6	10.58	26.58
11	7	10.58	17.58
12	8	10.58	27.58
13	9	10.58	27.58
14	10	10.58	37.58
15	11	10.58	28.58
16	12	5.83	1.83
17	13	5.83	8.83
18	14	5.83	8.83
19	15	5.83	8.83
20	16	10.83	16.83

Figure 7.10

Portion of Excel File *Pile Foundation*

Figure 7.11

Excel Output for Paired Two-Sample Test for Means

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*Estimated*	*Actual*
4	Mean	28.17755627	34.55623794
5	Variance	255.8100385	267.0113061
6	Observations	311	311
7	Pearson Correlation	0.79692836	
8	Hypothesized Mean Difference	0	
9	df	310	
10	t Stat	-10.91225025	
11	P(T<=t) one-tail	5.59435E-24	
12	t Critical one-tail	1.649783823	
13	P(T<=t) two-tail	1.11887E-23	
14	t Critical two-tail	1.967645929	

the F-test. To use this test, we must assume that both samples are drawn from normal populations. The hypotheses we test are

$$H_0: \sigma_1^2 - \sigma_2^2 = 0$$
$$H_1: \sigma_1^2 - \sigma_2^2 \neq 0 \qquad \textbf{(7.5)}$$

To test these hypotheses, we collect samples of n_1 observations from population 1 and n_2 observations from population 2. The test uses an F-test statistic, which is the ratio of the variances of the two samples:

$$F = \frac{s_1^2}{s_2^2} \qquad \textbf{(7.6)}$$

The sampling distribution of this statistic is called the F-distribution. Similar to the t-distribution, it is characterized by degrees of freedom; however, the F-distribution has *two* degrees of freedom, one associated with the numerator of the F-statistic, $n_1 - 1$, and one associated with the denominator of the F-statistic, $n_2 - 1$. Table A.4 in Appendix A at the end of the book provides only upper-tail critical values, and the distribution is *not* symmetric, as is the standard normal or the t-distribution. Therefore, although the hypothesis test is really a two-tailed test, we will simplify it as a one-tailed test to make it easy to use tables of the F-distribution and interpret the results of the Excel tool that we will use. We do this by ensuring that when we compute F, we take the ratio of the larger sample variance to the smaller sample variance.

If the variances differ significantly from each other, we would expect F to be much larger than 1; the closer F is to 1, the more likely it is that the variances are the same. Therefore, we need only to compare F to the upper-tail critical value. Hence, for a level of significance α, we find the critical value $F_{\alpha/2,df1,df2}$ of the F-distribution, and then we reject the null hypothesis if the F-test statistic exceeds the critical value. Note that we are using $\alpha/2$ to find the critical value, not α. This is because we are using only the upper tail information on which to base our conclusion.

EXAMPLE 7.12 Applying the *F*-Test for Equality of Variances

To illustrate the *F*-test, suppose that we wish to determine whether the variance of lead times is the same for Alum Sheeting and Durrable Products in the *Purchase Orders* data. The *F*-test can be applied using the Excel *Data Analysis* tool *F-test for Equality of Variances*. The dialog prompts you to enter the range of the sample data for each variable. As we noted, you should ensure that the first variable has the larger variance; this might require you to

Figure 7.12

Results for Two-Sample *F*-Test for Equality of Variances

	A	B	C
1	F-Test Two-Sample for Variances		
2		Alum Sheeting	Durrable Products
3		*Variable 1*	*Variable 2*
4	Mean	7	4.923076923
5	Variance	2	0.576923077
6	Observations	8	13
7	df	7	12
8	F	3.466666667	
9	P(F<=f) one-tail	0.028595441	
10	F Critical one-tail	3.606514642	

calculate the variances *before* you use the tool. In this case, the variance of the lead times for Alum Sheeting is larger than the variance for Durrable Products (see Figure 7.9), so this is assigned to *Variable 1*. Note also that if we choose $\alpha = 0.05$, we must enter 0.025 for the level of significance in the Excel dialog. The results are shown in Figure 7.12.

The value of the *F*-statistic, *F*, is 3.467. We compare this with the upper-tail critical value, *F Critical one-tail*, which is 3.607. Because $F < F\ Critical\ one\text{-}tail$, we cannot reject the null hypothesis and conclude that the variances are not significantly different from each other. Note that the *p*-value is $P(F<=f)\ one\ tail = 0.0286$. Although the level of significance is 0.05, remember that we must compare this to $\alpha/2 = 0.025$ because we are using only upper-tail information.

The *F*-test for equality of variances is often used before testing for the difference in means so that the proper test (population variance is unknown and assumed unequal or population variance is unknown and assumed equal, which we discussed earlier in this chapter) is selected.

Analysis of Variance (ANOVA)

To this point, we have discussed hypothesis tests that compare a population parameter to a constant value or that compare the means of two different populations. Often, we would like to compare the means of several different groups to determine if all are equal or if any are significantly different from the rest.

EXAMPLE 7.13 Differences in *Insurance Survey* Data

In the Excel data file *Insurance Survey*, we might be interested in whether any significant differences exist in satisfaction among individuals with different levels of education. We could sort the data by educational level and then create a table similar to the following.

	College Graduate	Graduate Degree	Some College
	5	3	4
	3	4	1
	5	5	4
	3	5	2
	3	5	3
	3	4	4
	3	5	4
	4	5	
	2		
Average	3.444	4.500	3.143
Count	9	8	7

Although the average satisfaction for each group is somewhat different and it appears that the mean satisfaction of individuals with a graduate degree is higher, we cannot tell conclusively whether or not these differences are significant because of sampling error.

In statistical terminology, the variable of interest is called a **factor**. In this example, the factor is the educational level, and we have three categorical levels of this factor, college graduate, graduate degree, and some college. Thus, it would appear that we will have to perform three different pairwise tests to establish whether any significant differences exist among them. As the number of factor levels increases, you can easily see that the number of pairwise tests grows large very quickly.

Fortunately, other statistical tools exist that eliminate the need for such a tedious approach. **Analysis of variance (ANOVA)** is one of them. The null hypothesis for ANOVA is that the population means of all groups are equal; the alternative hypothesis is that at least one mean differs from the rest:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$$

H_1: at least one mean is different from the others

ANOVA derives its name from the fact that we are analyzing variances in the data; essentially, ANOVA computes a measure of the variance between the means of each group and a measure of the variance within the groups and examines a test statistic that is the ratio of these measures. This test statistic can be shown to have an F-distribution (similar to the test for equality of variances). If the F-statistic is large enough based on the level of significance chosen and exceeds a critical value, we would reject the null hypothesis. Excel provides a *Data Analysis* tool, *ANOVA: Single Factor* to conduct analysis of variance.

EXAMPLE 7.14 Applying the Excel ANOVA Tool

To test the null hypothesis that the mean satisfaction for all educational levels in the Excel file *Insurance Survey* are equal against the alternative hypothesis that at least one mean is different, select *ANOVA: Single Factor* from the *Data Analysis* options. First, you must set up the worksheet so that the data you wish to use are displayed in contiguous columns as shown in Example 7.13. In the dialog shown in Figure 7.13, specify the input range of the data (which must be in contiguous columns) and whether it is stored in rows or columns (i.e., whether each factor level or group is a row or column in the range). The sample size for each factor level need not be the same, but the input range must be a rectangular region that contains all data. You must also specify the level of significance (α).

The results for this example are given in Figure 7.14. The output report begins with a summary report of basic statistics for each group. The ANOVA section reports the details of the hypothesis test. You needn't worry about all the mathematical details. The important information to interpret the test is given in the columns labeled *F* (the *F*-test statistic), *P-value* (the *p*-value for the test), and *F crit* (the critical value from the *F*-distribution). In this example, $F = 3.92$, and the critical value from the *F*-distribution is 3.4668. Here $F > F\ crit$; therefore, we must reject the null hypothesis and conclude that there are significant differences in the means of the groups; that is, the mean satisfaction is not the same among the three educational levels. Alternatively, we see that the *p*-value is smaller than the chosen level of significance, 0.05, leading to the same conclusion.

Figure 7.13

ANOVA Single Factor Dialog

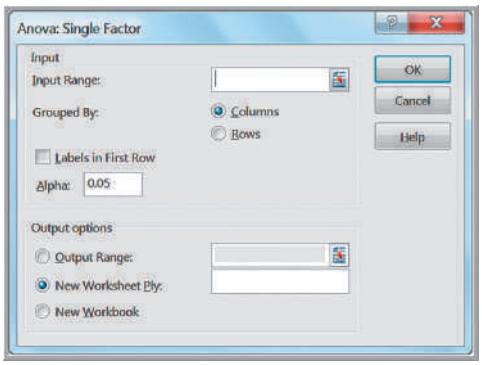


Figure 7.14

ANOVA Results for *Insurance Survey* Data

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	College graduate	9	31	3.444444444	1.027777778		
6	Graduate degree	8	36	4.5	0.571428571		
7	Some college	7	22	3.142857143	1.476190476		
8							
9							
10	ANOVA						
11	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between Groups	7.878968254	2	3.939484127	3.924651732	0.035635398	3.466800112
13	Within Groups	21.07936508	21	1.003779289			
14							
15	Total	28.95833333	23				

Although ANOVA can identify a difference among the means of multiple populations, it cannot determine which means are different from the rest. To do this, we may use the Tukey-Kramer multiple comparison procedure. Unfortunately, Excel does not provide this tool, but it may be found in other statistical software.

Assumptions of ANOVA

ANOVA requires assumptions that the m groups or factor levels being studied represent populations whose outcome measures

1. are randomly and independently obtained,
2. are normally distributed, and
3. have equal variances.

If these assumptions are violated, then the level of significance and the power of the test can be affected. Usually, the first assumption is easily validated when random samples are chosen for the data. ANOVA is fairly robust to departures from normality, so in most cases this isn't a serious issue. If sample sizes are equal, violation of the third assumption does not have serious effects on the statistical conclusions; however, with unequal sample sizes, it can.

When the assumptions underlying ANOVA are violated, you may use a *nonparametric test* that does not require these assumptions; we refer you to more comprehensive texts on statistics for further information and examples.

Finally, we wish to point out that students often use ANOVA to compare the equality of means of exactly two populations. It is important to realize that by doing this, you are making the assumption that the populations *have equal variances* (assumption 3). Thus, you will find that the *p*-values for both ANOVA and the *t*-Test: *Two-Sample Assuming Equal Variances* will be the same and lead to the same conclusion. However, if the variances are unequal as is generally the case with sample data, ANOVA may lead to an erroneous conclusion. We recommend that you do not use ANOVA for comparing the means of two populations, but instead use the appropriate *t*-test that assumes unequal variances.

Chi-Square Test for Independence

A common problem in business is to determine whether two categorical variables are independent. We introduced the concept of independent events in Chapter 5. In the energy drink survey example (Example 5.9), we used conditional probabilities to determine whether brand preference was independent of gender. However, with sample data, sampling error can make it difficult to properly assess the independence of categorical variables. We would never expect the joint probabilities to be exactly the same as the product of the marginal probabilities because of sampling error even if the two variables are statistically independent. Testing for independence is important in marketing applications.

EXAMPLE 7.15 Independence and Marketing Strategy

Figure 7.15 shows a portion of the sample data used in Chapter 5 for brand preferences of energy drinks (Excel file *Energy Drink Survey*) and the cross-tabulation of the results. A key marketing question is whether the proportion of males who prefer a particular brand is no different from the proportion of females. For instance, of the 63 male students, 25 (40%) prefer brand 1. If gender and brand preference are indeed independent, we would expect that about the same proportion of the sample of female students would also prefer brand 1. In actuality, only 9 of 37 (24%) prefer brand 1. However, we do not know whether this is simply due to sampling error or represents a significant difference. Knowing whether gender and brand preference are independent can help marketing personnel better target advertising campaigns. If they are not independent, then advertising should be targeted differently to males and females, whereas if they are independent, it would not matter.

We can test for independence by using a hypothesis test called the *chi-square test for independence.* The chi-square test for independence tests the following hypotheses:

H_0: the two categorical variables are independent

H_1: the two categorical variables are dependent

The chi-square test is an example of a *nonparametric test*; that is, one that does not depend on restrictive statistical assumptions, as ANOVA does. This makes it a widely applicable and popular tool for understanding relationships among categorical data. The first step in the procedure is to compute the expected frequency in each cell of the cross-tabulation if the two variables are independent. This is easily done using the following:

$$\text{expected frequency in row } i \text{ and column } j = \frac{(\text{grand total row } i)(\text{grand total column } j)}{\text{total number of observations}} \tag{7.7}$$

Figure 7.15

Portion of *Energy Drink Survey* and Cross-Tabulation

	A	B	C	D	E	F	G	H	I
1	Energy Drink Survey								
2									
3	Respondent	Gender	Brand Preference						
4	1	Male	Brand 3		Count of Respondent	Column Labels			
5	2	Female	Brand 3		Row Labels	Brand 1	Brand 2	Brand 3	Grand Total
6	3	Male	Brand 3		Female	9	6	22	37
7	4	Male	Brand 1		Male	25	17	21	63
8	5	Male	Brand 1		Grand Total	34	23	43	100
9	6	Female	Brand 2						
10	7	Male	Brand 2						

Figure 7.16

Expected Frequencies for the Chi-Square Test

	E	F	G	H	I	J	K
1	Chi-Square Test						
2							
3	Count of Respondent	Column Labels					
4	Row Labels	Brand 1	Brand 2	Brand 3	Grand Total		
5	Female	9	6	22	37		
6	Male	25	17	21	63		
7	Grand Total	34	23	43	100		
8							
9							
10	Expected Frequency	Brand 1	Brand 2	Brand 3	Grand Total		
11	Female	12.58	8.51	15.91	37		
12	Male	21.42	14.49	27.09	63		
13	Grand Total	34	23	43	100		

Expected frequency of Female and Brand 1 = 37*34/100

EXAMPLE 7.16 Computing Expected Frequencies

For the *Energy Drink Survey* data, we may compute the expected frequencies using the data from the cross-tabulation and formula (7.7). For example, the expected frequency of females who prefer brand 1 is (37)(34)/100 = 12.58. This can easily be implemented in Excel. Figure 7.16 shows the results (see the Excel file *Chi-Square Test*). The formula in cell F11, for example, is =$I5*F$7/I7, which can be copied to the other cells to complete the calculations.

Next, we compute a test statistic, called a **chi-square statistic**, which is the sum of the squares of the differences between observed frequency, f_o, and expected frequency, f_e, divided by the expected frequency in each cell:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \tag{7.8}$$

The closer the observed frequencies are to the expected frequencies, the smaller will be the value of the chi-square statistic. The sampling distribution of χ^2 is a special distribution called the **chi-square** (χ^2) **distribution**. The chi-square distribution is characterized by degrees of freedom, similar to the *t*-distribution. Table A.3 in Appendix A in the back of this book provides critical values of the chi-square distribution for selected values of α. We compare the chi-square statistic for a specified level of significance α to the critical value from a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom, where r and c are the number of rows and columns in the cross-tabulation table, respectively. The Excel function CHISQ.INV.RT(*probability, deg_ freedom*) returns the value of χ^2 that has a right-tail area equal to *probability* for a specified degree of freedom. By setting *probability* equal to the level of significance, we can obtain the critical value for the hypothesis test. If the test statistic exceeds the critical value for a specified level of significance, we reject H_0. The Excel function CHISQ.TEST(*actual_range, expected_range*) computes the *p*-value for the chi-square test.

Figure 7.17

Excel Implementation of Chi-Square Test

	E	F	G	H	I
1	Chi-Square Test				
2					
3	Count of Respondent	Column Labels			
4	Row Labels	Brand 1	Brand 2	Brand 3	Grand Total
5	Female	9	6	22	37
6	Male	25	17	21	63
7	Grand Total	34	23	43	100
8					
9					
10	Expected Frequency	Brand 1	Brand 2	Brand 3	Grand Total
11	Female	12.58	8.51	15.91	37
12	Male	21.42	14.49	27.09	63
13	Grand Total	34	23	43	100
14					
15					
16	Chi Square Statistic	Brand 1	Brand 2	Brand 3	Grand Total
17	Female	1.02	0.74	2.33	4.09
18	Male	0.60	0.43	1.37	2.40
19	Grand Total	1.62	1.18	3.70	6.49
20					
21		Chi-square critical value			5.99146455
22		p-value			0.03892134

EXAMPLE 7.17 Conducting the Chi-Square Test

For the *Energy Drink Survey* data, Figure 7.17 shows the calculations of the chi-square statistic using formula (7.8). For example, the formula in cell F17 is $=(F5 - F11)^2/F11$, which can be copied to the other cells. The grand total in the lower right cell is the value of χ^2. In this case, the chi-square test statistic is 6.4924. Since the cross-tabulation has $r = 2$ rows and $c = 3$ columns, we have $(2 - 1)(3 - 1) = 2$ degrees of freedom for the chi-square distribution. Using $\alpha = 0.05$, the Excel function CHISQ.INV.RT(0.05,2) returns the critical value 5.99146. Because the test statistic exceeds the critical value, we reject the null hypothesis that the two categorical variables are independent.

Alternatively, we could simply use the CHISQ.TEST function to find the *p*-value for the test and base our conclusion on that without computing the chi-square statistic. For this example, the function CHISQ.TEST(F6:H7,F12:H13) returns the *p*-value of 0.0389, which is less than $\alpha = 0.05$; therefore, we reject the null hypothesis.

Cautions in Using the Chi-Square Test

First, when using PivotTables to construct a cross-tabulation and implement the chi-square test in Excel similar to Figure 7.17, be extremely cautious of blank cells in the PivotTable. Blank cells will not be counted in the chi-square calculations and will lead to errors. If you have blank cells in the PivotTable, simply replace them by zeros, or right-click in the PivotTable, choose *PivotTable Options*, and enter 0 in the field for the checkbox *For empty cells show*.

Second, the chi-square test assumes adequate expected cell frequencies. A rule of thumb is that there be no more than 20% of cells with expected frequencies smaller than 5, and no expected frequencies of zero. More advanced statistical procedures exist to handle this, but you might consider aggregating some of the rows or columns in a logical fashion to enforce this assumption. This, of course, results in fewer rows or columns.

Analytics in Practice: Using Hypothesis Tests and Business Analytics in a Help Desk Service Improvement Project[1]

Schlumberger is an international oilfield-services provider headquartered in Houston, Texas. Through an outsourcing contract, they supply help-desk services for a global telecom company that offers wireline communications and integrated telecom services to more than 2 million cellular subscribers. The help desk, located in Ecuador, faced increasing customer complaints and losses in dollars and cycle times. The company drew upon the analytics capability of one of the help-desk managers to investigate and solve the problem. The data showed that the average solution time for issues reported to the help desk was 9.75 hours. The company set a goal to reduce the average solution time by 50%. In addition, the number of issues reported to the help desk had reached an average of 30,000 per month. Reducing the total number of issues reported to the help desk would allow the company to address those issues that hadn't been resolved because of a lack of time, and to reduce the number of abandoned calls. They set a goal to identify preventable issues so that customers would not have to contact the help desk in the first place, and set a target of 15,000 issues.

As part of their analysis, they observed that the average solution time for help-desk technicians working at the call center seemed to be lower than the average for technicians working on site with clients. They conducted a hypothesis test structured around the question: Is there a difference between having help desk employees working at an off-site facility rather than on site within the client's main office? The null hypothesis was that there was no significant difference; the alternative hypothesis was that there was a significant difference. Using a two-sample *t*-test to assess whether the call center and the help desk are statistically different from each other, they found no statistically significant advantage in keeping help-desk employees working at the call center. As a result, they moved help-desk agents to the client's main office area. Using a variety of other analytical techniques, they were able to make changes to their process, resulting in the following:

StockLite/Shutterstock.com

- a decrease in the number of help-desk issues of 32%
- improved capability to meet the target of 15,000 total issues
- a reduction in the average desktop solution time from 9.75 hours to 1 hour, an improvement of 89.5%
- a reduction in the call-abandonment rate from 44% to 26%
- a reduction of 69% in help-desk operating costs

Key Terms

Alternative hypothesis
Analysis of variance (ANOVA)
Chi-square distribution
Chi-square statistic
Confidence coefficient
Factor
Hypothesis
Hypothesis testing
Level of significance
Null hypothesis
One-sample hypothesis test
One-tailed test of hypothesis
p-Value (observed significance level)
Power of the test
Statistical inference
Two-tailed test of hypothesis
Type I error
Type II error

[1]Based on Francisco, Endara M. "Help Desk Improves Service and Saves Money with Six Sigma," American Society for Quality, http://asq.org/economic-case/markets/pdf/help-desk-24490.pdf, accessed 8/19/11.

Problems and Exercises

For all hypothesis tests, assume that the level of significance is 0.05 unless otherwise stated.

1. Create an Excel workbook with worksheet templates (similar to the Excel workbook *Confidence Intervals*) for one-sample hypothesis tests for means and proportions. Apply your templates to the example problems in this chapter. (For subsequent problems, you should use the formulas in this chapter to perform the calculations, and use this template only to verify your results!)
2. A company is considering two different campaigns, A and B, for the promotion of their product. Two tests are conducted in two market areas with identical consumer characteristics, and in a random sample of 60 customers who saw campaign A, 18 tried the product. In a random sample of 100 customers who saw campaign B, 22 tried the product. What conclusion can management reach? (Assume that the population variance is not known.)
3. A management institute checked the past records of applicants and the mean score calculated was 350. The administration is interested to know whether the quality of new applicants has changed or not. From the recent scores of 100 applicants, the mean is 365 with a standard deviation of 38. Does this data provide statistical evidence that the quality of recent applicants has improved?
4. A retailer believes that its new advertising strategy will increase sales. Previously, the mean spending in 15 categories of consumer items in both the 18–34 and 35+ age groups was $70.00.
   a. Formulate a hypothesis test to determine if the mean spending in these categories has statistically increased.
   b. After the new advertising campaign was launched, a marketing study found that the mean spending for 300 respondents in the 18–34 age group was $75.86, with a standard deviation of $50.90. Is there sufficient evidence to conclude that the advertising strategy significantly increased sales in this age group?
   c. For 700 respondents in the 35+ age group, the mean and standard deviation were $68.53 and $45.29, respectively. Is there sufficient evidence to conclude that the advertising strategy significantly increased sales in this age group?
5. A financial advisor believes that the proportion of investors who are risk–averse (i.e., try to avoid risk in their investment decisions) is at least 0.7. A survey of 32 investors found that 20 of them were risk-averse. Formulate and test the appropriate hypotheses to determine whether his belief is valid.
6. Metropolitan Press hypothesizes that the average life of its largest Web press is 14,500 hours. They know that the standard deviation of press life is 2,100 hours. From a sample of 25 presses, the company find sample mean of 13,000 hours. At a 0.01 significance level, should the company conclude that the average life of the presses is less than the hypothesized 14,500 hours?
7. Ice Cream Manufacture is to produce a new ice cream flavor. The company's marketing research department surveyed 6,000 families and 335 of them showed interest in purchasing the new flavor. A similar study made two year ago showed that 5% of the families would purchase the flavor. What should the company conclude regarding the new flavor?
8. Call centers typically have high turnover. The director of human resources for a large bank has compiled data on about 70 former employees at one of the bank's call centers in the Excel file *Call Center Data*. In writing an article about call center working conditions, a reporter has claimed that the average tenure is no more than 2 years. Formulate and test a hypothesis using these data to determine if this claim can be disputed.
9. The manager of a store claims that 60% of the shoppers entering the store leave without making a purchase. Out of a sample of 50, it is found that 35 shoppers left without buying. Is the result consistent with the claim?
10. A sample of 400 athletes is found to have mean height of 171.38 cm. Can we call it a sample from a large population of mean height 171.17 and standard deviation of 3.30 cm?
11. The State of Ohio Department of Education has a mandated ninth-grade proficiency test that covers writing, reading, mathematics, citizenship (social studies), and science. The Excel file *Ohio Education Performance* provides data on success rates (defined as the percent of students passing) in school districts in the greater Cincinnati metropolitan area along with state averages. Test null hypotheses that the average scores in the Cincinnati area are equal to the state averages in each test and also for the composite score.
12. Formulate and test hypotheses to determine if statistical evidence suggests that the graduation rate for (1) top liberal arts colleges or (2) research universities in the sample *Colleges and Universities* exceeds 90%. Do the data support a conclusion that the graduation rates exceed 85%? Would your conclusions

change if the level of significance was 0.01 instead of 0.05?

13. The Excel file *Sales Data* provides data on a sample of customers. An industry trade publication stated that the average profit per customer for this industry was at least $4,500. Using a test of hypothesis, do the data support this claim or not?

14. The Excel file *Room Inspection* provides data for 100 room inspections at each of 25 hotels in a major chain. Management would like the proportion of nonconforming rooms to be less than 2%. Test an appropriate hypothesis to determine if management can make this claim.

15. An employer is considering negotiating its pricing structure for health insurance with its provider if there is sufficient evidence that customers will be willing to pay a lower premium for a higher deductible. Specifically, they want at least 30% of their employees to be willing to do this. Using the sample data in the Excel file *Insurance Survey*, determine what decision they should make.

16. Using the data in the Excel file *Consumer Transportation Survey*, test the following null hypotheses:

- **a.** Individuals spend at least 8 hours per week in their vehicles.
- **b.** Individuals drive an average of 600 miles per week.
- **c.** The average age of SUV drivers is no greater than 35.
- **d.** At least 80% of individuals are satisfied with their vehicles.

17. Using the Excel file *Facebook Survey*, determine if the mean number of hours spent online per week is the same for males as it is for females.

18. Determine if there is evidence to conclude that the mean number of vacations taken by married individuals is less than the number taken by single/divorced individuals using the data in the Excel file *Vacation Survey*. Use a level of significance of 0.05. Would your conclusion change if the level of significance is 0.01?

19. The Excel file *Accounting Professionals* provides the results of a survey of 27 employees in a tax division of a *Fortune* 100 company.

- **a.** Test the null hypothesis that the average number of years of service is the same for males and females.
- **b.** Test the null hypothesis that the average years of undergraduate study is the same for males and females.

20. In the Excel file *Cell Phone Survey*, test the hypothesis that the mean responses for Value for the Dollar and Customer Service do not differ by gender.

21. A sample size of 22 with a mean of 8 and a standard deviation of 12.5 test the hypothesis that the value of the population mean is 70 against the assumption that it is more than 70. Use the 0.025 significant levels.

22. Determine if there is evidence to conclude that the mean GPA of males who plan to attend graduate school is larger than that of females who plan to attend graduate school using the data in the Excel file *Graduate School Survey*.

23. The director of human resources for a large bank has compiled data on about 70 former employees at one of the bank's call centers (see the Excel file *Call Center Data*). For each of the following, assume equal variances of the two populations.

- **a.** Test the null hypothesis that the average length of service for males is the same as for females.
- **b.** Test the null hypothesis that the average length of service for individuals without prior call center experience is the same as those with experience.
- **c.** Test the null hypothesis that the average length of service for individuals with a college degree is the same as for individuals without a college degree.
- **d.** Now conduct tests of hypotheses for equality of variances. Were your assumptions of equal variances valid? If not, repeat the test(s) for means using the unequal variance test.

24. A producer of computer-aided design software for the aerospace industry receives numerous calls for technical support. Tracking software is used to monitor response and resolution times. In addition, the company surveys customers who request support using the following scale: 0—did not exceed expectations; 1—marginally met expectations; 2—met expectations; 3—exceeded expectations; 4—greatly exceeded expectations. The questions are as follows:

Q1: Did the support representative explain the process for resolving your problem?

Q2: Did the support representative keep you informed about the status of progress in resolving your problem?

Q3: Was the support representative courteous and professional?

Q4: Was your problem resolved?

Q5: Was your problem resolved in an acceptable amount of time?

Q6: Overall, how did you find the service provided by our technical support department?

A final question asks the customer to rate the overall quality of the product using a scale of 0—very poor; 1—poor; 2—good; 3—very good; 4—excellent. A sample of survey responses and associated resolution and response data are provided in the Excel file *Customer Support Survey.*

a. The company has set a service standard of 1 day for the mean resolution time. Does evidence exist that the response time is more than 1 day? How do the outliers in the data affect your result? What should you do about them?

b. Test the hypothesis that the average service index is equal to the average engineer index.

25. Using the data in the Excel file *Ohio Education Performance*, test the hypotheses that the mean difference in writing and reading scores is zero and that the mean difference in math and science scores is zero. Use the paired-sample procedure.

26. The Excel file *Unions and Labor Law Data* reports the percent of public- and private-sector employees in unions in 1982 for each state, along with indicators whether the states had a bargaining law that covered public employees or right-to-work laws.

a. Test the hypothesis that the mean percent of employees in unions for both the public sector and private sector is the same for states having bargaining laws as for those who do not.

b. Test the hypothesis that the mean percent of employees in unions for both the public sector and private sector is the same for states having right-to-work laws as for those who do not.

27. Using the data in the Excel file *Student Grades*, which represent exam scores in one section of a large statistics course, test the hypothesis that the variance in grades is the same for both tests.

28. In the Excel file *Restaurant Sales*, determine if the variance of weekday sales is the same as that of weekend sales for each of the three variables (lunch, dinner, and delivery).

29. A college is trying to determine if there is a significant difference in the mean GMAT score of students from different undergraduate backgrounds who apply to the MBA program. The Excel file *GMAT Scores* contain data from a sample of students. What conclusion can be reached using ANOVA?

30. Using the data in the Excel file *Cell Phone Survey*, apply ANOVA to determine if the mean response for Value for the Dollar is the same for different types of cell phones.

31. Using the data in the Excel file *Freshman College Data*, use ANOVA to determine whether significant differences exist in the mean retention rate for the different colleges over the 4-year period. Second, use ANOVA to determine if significant differences exist in the mean ACT and SAT scores among the different colleges.

32. A car manufacturing firm is bringing out a new model. To figure out its advertising campaign, they want to determine whether the model appeal will be dependent on a particular age group. A sample of a customer survey revealed the following:

	Under 20	20–40	40–50	50 and over	Total
Liked	140	70	70	25	305
Disliked	60	40	30	65	195
Total	200	110	100	90	500

What can the manufacturer conclude?

33. A survey of college students determined the preference for cell phone providers. The following data were obtained:

	Provider			
Gender	T-Mobile	AT&T	Verizon	Other
Male	12	39	27	16
Female	8	22	24	12

Can we conclude that gender and cell phone provider are independent? If not, what implications does this have for marketing?

34. For the data in the Excel file *Accounting Professionals*, perform a chi-square test of independence to determine if age group is independent of having a graduate degree.

35. For the data in the Excel file *Graduate School Survey*, perform a chi-square test for independence to determine if plans to attend graduate school are independent of gender.

36. For the data in the Excel file *New Account Processing*, perform chi-square tests for independence to determine if certification is independent of gender, and if certification is independent of having prior industry background.

Case: Drout Advertising Research Project

The background for this case was introduced in Chapter 1. This is a continuation of the case in Chapter 6. For this part of the case, propose and test some meaningful hypotheses that will help Ms. Drout understand and explain the results. Include two-sample tests, ANOVA, and/or Chi-Square tests for independence as appropriate. Write up your conclusions in a formal report, or add your findings to the report you completed for the case in Chapter 6 as per your instructor's requirements. If you have accumulated all sections of this case into one report, polish it up so that it is as professional as possible, drawing final conclusions about the perceptions of the role of advertising in the reinforcement of gender stereotypes and the impact of empowerment advertising.

Case: Performance Lawn Equipment

Elizabeth Burke has identified some additional questions she would like you to answer.

1. Are there significant differences in ratings of specific product/service attributes in the *2014 Customer Survey* worksheet?
2. In the worksheet *On-Time Delivery*, has the proportion of on-time deliveries in 2014 significantly improved since 2010?
3. Have the data in the worksheet *Defects After Delivery* changed significantly over the past 5 years?
4. Although engineering has collected data on alternative process costs for building transmissions in the worksheet *Transmission Costs*, why didn't they reach a conclusion as to whether one of the proposed processes is better than the current process?
5. Are there differences in employee retention due to gender, college graduation status, or whether the employee is from the local area in the data in the worksheet *Employee Retention*?

Conduct appropriate statistical analyses and hypothesis tests to answer these questions and summarize your results in a formal report to Ms. Burke.