Chapter 8

Hypothesis Testing

Chapter Outline

8.1 The Null and Alternative Hypotheses and Errors in Hypothesis Testing

8.2 Large Sample Tests about a Population Mean: One-Sided Alternatives

8.3 Large Sample Tests about a Population Mean: Two-Sided Alternatives

8.4 Small Sample Tests about a Population Mean

8.5 Tests about a Population Proportion

*8.6 Type II Error Probabilities and Sample Size Determination

*8.7 The Chi-Square Distribution

*8.8 Statistical Inference for a Population Variance

*Optional section

Hypothesis testing is a statistical procedure used to provide evidence in favor of some statement (called a *hypothesis*). For instance, hypothesis testing might be used to assess whether a population parameter, such as a population mean, differs from a specified standard or previous value. In this chapter we discuss testing hypotheses about population means, proportions, and variances.

In order to illustrate how hypothesis testing works, we revisit several cases introduced in previous chapters:

The Accounts Receivable Case: The consulting firm uses hypothesis testing to provide strong evidence that the new electronic billing system has reduced the mean payment time by more than 50 percent.

The Cheese Spread Case: The cheese spread producer uses hypothesis testing to supply extremely strong evidence that fewer than 10 percent of all current purchasers would stop buying the cheese spread if the new spout were used.

The Electronic Article Surveillance Case: A company that sells and installs EAS systems claims that no more than 5 percent of all consumers would say they would never shop in a store again if the store subjected them to a false EAS alarm. A store considering the purchase of such a system uses hypothesis testing to provide extremely strong evidence that this claim is not true.

In addition, we introduce two new cases that illustrate how to test a hypothesis:

The Trash Bag Case: A marketer of trash bags uses hypothesis testing to support its claim that the mean breaking strength of its new trash bag is greater than 50 pounds. As a result, a television network approves use of this claim in a commercial.

The Camshaft Case: An automobile manufacturer uses hypothesis testing to study an important quality characteristic affecting V6 engine camshafts. It finds that the mean "case hardness depth" differs from its desired target value and that this problem is one reason why some of the case hardness depths fail to meet specifications.

8.1 The Null and Alternative Hypotheses and Errors in Hypothesis Testing

One of the authors' former students is employed by a major television network in the standards and practices division. One of the division's responsibilities is to reduce the chances that advertisers will make false claims in commercials run on the network. Our former student reports that the network uses a statistical methodology called **hypothesis testing** to do this.

CHAPTER 9

To see how this might be done, suppose that a company wishes to advertise a claim, and suppose that the network has reason to doubt that this claim is true. The network assumes for the sake of argument that **the claim is not valid.** This assumption is called the **null hypothesis.** The statement that **the claim is valid** is called the **alternative,** or **research, hypothesis.** The network will run the commercial only if the company making the claim provides **sufficient sample evidence** to reject the null hypothesis that the claim is not valid in favor of the alternative hypothesis that the claim is valid. Explaining the exact meaning of *sufficient sample evidence* is quite involved and will be discussed in the next section.

The Null Hypothesis and the Alternative Hypothesis

In hypothesis testing:

1. The **null hypothesis,** denoted H_0, is the statement being tested. Usually this statement represents the *status quo* and is not rejected unless there is convincing sample evidence that it is false.
2. The **alternative,** or **research, hypothesis,** denoted H_a, is a statement that will be accepted only if there is convincing sample evidence that it is true.

Setting up the null and alternative hypotheses in a practical situation can be tricky. In some situations there is a condition for which we need to attempt to find supportive evidence. We then formulate (1) the alternative hypothesis to be the statement that this condition exists and (2) the null hypothesis to be the statement that this condition does not exist. To illustrate this, we consider the following case study.

EXAMPLE 8.1 The Trash Bag Case[1]

A leading manufacturer of trash bags produces the strongest trash bags on the market. The company has developed a new 30-gallon bag using a specially formulated plastic that is stronger and more biodegradable than other plastics. This plastic's increased strength allows the bag's thickness to be reduced, and the resulting cost savings will enable the company to lower its bag price by 25 percent. The company also believes the new bag is stronger than its current 30-gallon bag.

The manufacturer wants to advertise the new bag on a major television network. In addition to promoting its price reduction, the company also wants to claim the new bag is better for the environment and stronger than its current bag. The network is convinced of the bag's environmental advantages on scientific grounds. However, the network questions the company's claim of increased strength and requires statistical evidence to justify this claim. Although there are various measures of bag strength, the manufacturer and the network agree to employ "breaking strength." A bag's breaking strength is the amount of a representative trash mix (in pounds) that, when loaded into a bag suspended in the air, will cause the bag to rip or tear. Tests show that the current bag has a mean breaking strength that is very close to (but does not exceed) 50 pounds. The new bag's mean breaking strength μ is unknown and in question. The alternative hypothesis H_a is the statement for which we wish to find supportive evidence. Because we hope the new bags are stronger than the current bags, H_a says that μ is greater than 50. The null hypothesis states that H_a is false. Therefore, H_0 says that μ is less than or equal to 50. We summarize these hypotheses by stating that we are testing

$$H_0\colon \mu \leq 50 \quad \text{versus} \quad H_a\colon \mu > 50$$

The network will run the manufacturer's commercial if a random sample of n new bags provides sufficient evidence to reject $H_0\colon \mu \leq 50$ in favor of $H_a\colon \mu > 50$.

In the trash bag example, the decision to run the commercial is based solely on whether H_0 will be rejected in favor of H_a. In many situations, however, the result of a hypothesis test is used as input into a more complex decision-making process, and the hypothesis test is not the sole basis for the decision that will be made. This is true in the next two cases.

EXAMPLE 8.2 The Accounts Receivable Case

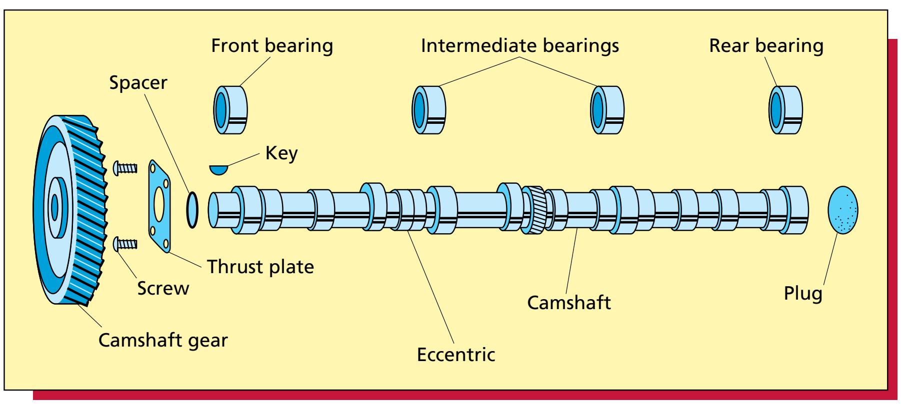
Recall that a management consulting firm has installed a new computer-based, electronic billing system in a Hamilton, Ohio, trucking company. Because of the system's advantages, and because the trucking company's clients are receptive to using this system, the management consulting firm believes that the new system will reduce the mean bill payment time by more than 50 percent. The mean payment time using the old billing system was approximately equal to, but no less than, 39 days. Therefore, if μ denotes the mean payment time using the new system, the consulting firm believes that μ will be less than 19.5 days. Because it is hoped that the new billing system *reduces* mean payment time, we formulate the alternative hypothesis as $H_a\colon \mu < 19.5$ and the null hypothesis as $H_0\colon \mu \geq 19.5$. The consulting firm will randomly select a sample of n invoices and determine how much evidence their payment times provide to reject H_0 in favor of H_a. The firm will use the results of the hypothesis test to demonstrate the benefits of the new billing system both to the Hamilton company and to other trucking companies that are considering using such a system. Note, however, that a potential user will decide whether to install the new system by considering factors beyond the results of the hypothesis test. For example, the cost of the new billing system and the receptiveness of the company's clients to using the new system are among other factors that must be considered. In complex business and industrial situations such as this, hypothesis testing is used to accumulate knowledge about and understand the problem at hand. The ultimate decision (such as whether to adopt the new billing system) is made on the basis of nonstatistical considerations, intuition, and the results of one or more hypothesis tests.

[1]This case is based on conversations by the authors with several employees working for a leading producer of trash bags. For purposes of confidentiality, we have agreed to withhold the company's name.

EXAMPLE 8.3 The Camshaft Case

An automobile manufacturer produces the parts for its vehicles in many different locations and transports them to assembly plants. In order to keep the assembly operations running efficiently, it is vital that all parts be within specification limits. One important part used in the assembly of V6 engines is the engine camshaft, and one important quality characteristic of this camshaft is the case hardness depth of its eccentrics. A camshaft *eccentric* is a metal disk positioned on the camshaft so that as the camshaft turns, the eccentric drives a lifter that opens and closes an engine valve. The V6 engine camshaft and its eccentrics are illustrated in Figure 8.1. These eccentrics are hardened by passing the camshaft through an electrical coil that "cooks" or "bakes" the metal. Studies indicate that the hardness depth of the eccentric pointed out in Figure 8.1 is representative of the hardness depth of all the eccentrics on the camshaft. Therefore, the hardness depth of this representative eccentric is measured at a specific location and is regarded as the **hardness depth of the camshaft.** The *optimal* or *target* hardness depth for a camshaft is 4.5 mm. In addition, specifications state that, in order for the camshaft to wear properly, the hardness depth of the camshaft must be between 3.0 mm and 6.0 mm. Unfortunately, however, the process has been producing far too many out-of-specification camshafts.

FIGURE 8.1 **A Camshaft and Related Parts**



Suppose that, in order to investigate why the process is not meeting specifications, a quality control analyst decides to randomly select a sample of n camshafts from the population of all camshafts produced on a particular day. The analyst measures the hardness depth of each camshaft in the sample and uses the resulting sample of n hardness depths to evaluate the population of the hardness depths of all camshafts produced on the particular day. Specifically, if μ denotes the mean of this population, the analyst will evaluate whether μ differs from the target value of 4.5 mm by testing the null hypothesis H_0: $\mu = 4.5$ versus the alternative hypothesis H_a: $\mu \neq 4.5$. Of course, μ differing from 4.5 is not the only reason why the hardness depths of the camshafts produced on the particular day might be out of specification. Another reason is that the variation of the hardness depths might be too large.

Below we summarize the sets of null and alternative hypotheses that we have thus far considered.

H_0: $\mu \leq 50$	H_0: $\mu \geq 19.5$	H_0: $\mu = 4.5$
versus	versus	versus
H_a: $\mu > 50$	H_a: $\mu < 19.5$	H_a: $\mu \neq 4.5$

The alternative hypothesis H_a: $\mu > 50$ is called a **one-sided, greater than** alternative hypothesis, whereas H_a: $\mu < 19.5$ is called a **one-sided, less than** alternative hypothesis, and H_a: $\mu \neq 4.5$ is called a **two-sided, not equal to** alternative hypothesis. Many of the alternative hypotheses we consider in this book are one of these three types. Also, note that each null hypothesis we have considered involves an **equality.** For example, the null hypothesis H_0: $\mu \leq 50$ says that

μ is either less than or **equal to** 50. We will see that, in general, the approach we use to test a null hypothesis versus an alternative hypothesis requires that the null hypothesis involve an equality.

The idea of a test statistic Suppose that in the trash bag case the manufacturer randomly selects a sample of $n = 40$ new trash bags. Each of these bags is tested for breaking strength, and the sample mean $\bar{x}$ and sample standard deviation s of the 40 breaking strengths are calculated. In order to test H_0: $\mu \leq 50$ versus H_a: $\mu > 50$, we utilize the **test statistic**

$$z = \frac{\bar{x} - 50}{\sigma_{\bar{x}}} = \frac{\bar{x} - 50}{\sigma/\sqrt{n}}$$

Here, since σ is unknown and the sample size is large, σ is estimated by s.

The test statistic z measures the distance between $\bar{x}$ and 50. The division by $\sigma_{\bar{x}}$ says that this distance is measured in units of the standard deviation of all possible sample means. For example, a value of z equal to, say, 2.4 would tell us that $\bar{x}$ is 2.4 such standard deviations above 50. In general, a value of the test statistic that is less than or equal to zero results when $\bar{x}$ is less than or equal to 50. This provides no evidence to support rejecting H_0 in favor of H_a because the point estimate $\bar{x}$ indicates that μ is probably less than or equal to 50. However, a value of the test statistic that is greater than zero results when $\bar{x}$ is greater than 50. This provides evidence to support rejecting H_0 in favor of H_a because the point estimate $\bar{x}$ indicates that μ might be greater than 50. Furthermore, the farther the value of the test statistic is above 0 (the farther $\bar{x}$ is above 50), the stronger is the evidence to support rejecting H_0 in favor of H_a.

Hypothesis testing and the legal system If the value of the test statistic z is far enough above 0, we reject H_0 in favor of H_a. To see how large z must be in order to reject H_0, we must understand that **a hypothesis test rejects a null hypothesis H_0 only if there is strong statistical evidence against H_0.** This is similar to our legal system, which rejects the innocence of the accused only if evidence of guilt is beyond a reasonable doubt. For instance, the network will reject H_0: $\mu \leq 50$ and run the trash bag commercial only if the test statistic z is far enough above 0 to show beyond a reasonable doubt that H_0: $\mu \leq 50$ is false and H_a: $\mu > 50$ is true. A test statistic that is only slightly greater than 0 might not be convincing enough. However, because such a test statistic would result from a sample mean $\bar{x}$ that is slightly greater than 50, it would provide some evidence to support rejecting H_0: $\mu \leq 50$, and it certainly would not provide strong evidence supporting H_0: $\mu \leq 50$. Therefore, if the value of the test statistic is not large enough to convince us to reject H_0, **we do not say that we accept H_0. Rather we say that we do not reject H_0** because the evidence against H_0 is not strong enough. Again, this is similar to our legal system, where the lack of evidence of guilt beyond a reasonable doubt results in a verdict of **not guilty,** but does not prove that the accused is innocent.

Type I and Type II errors and their probabilities To determine exactly how much statistical evidence is required to reject H_0, we consider the errors and the correct decisions that can be made in hypothesis testing. These errors and correct decisions, as well as their implications in the trash bag advertising example, are summarized in Tables 8.1 and 8.2. Across the top of each table are listed the two possible "states of nature." Either H_0: $\mu \leq 50$ is true, which says the manufacturer's claim that μ is greater than 50 is false, or H_0 is false, which says the claim is true. Down the left side of each table are listed the two possible decisions we can make in the hypothesis test. Using the sample data, we will either reject H_0: $\mu \leq 50$, which implies that the claim will be advertised, or we will not reject H_0, which implies that the claim will not be advertised.

In general, the two types of errors that can be made in hypothesis testing are defined here:

Type I and Type II Errors

If we reject H_0 when it is true, this is a **Type I error.**
If we do not reject H_0 when it is false, this is a **Type II error.**

TABLE 8.1 Type I and Type II Errors

	State of Nature	
Decision	**H_0 True**	**H_0 False**
Reject H_0	Type I error	Correct decision
Do not reject H_0	Correct decision	Type II error

TABLE 8.2 The Implications of Type I and Type II Errors in the Trash Bag Example

	State of Nature	
Decision	**Claim False**	**Claim True**
Advertise the claim	Advertise a false claim	Advertise a true claim
Do not advertise the claim	Do not advertise a false claim	Do not advertise a true claim

As can be seen by comparing Tables 8.1 and 8.2, if we commit a Type I error, we will advertise a false claim. If we commit a Type II error, we will fail to advertise a true claim.

We now let the symbol $\boldsymbol{\alpha}$ (pronounced **alpha**) **denote the probability of a Type I error,** and we let $\boldsymbol{\beta}$ (pronounced **beta**) **denote the probability of a Type II error.** Obviously, we would like both α and β to be small. A common (but not the only) procedure is to base a hypothesis test on taking a sample of a fixed size (for example, $n = 40$ trash bags) and on setting α equal to a small prespecified value. Setting α low means there is only a small chance of rejecting H_0 when it is true. This implies that we are requiring strong evidence against H_0 before we reject it.

We sometimes choose α as high as .10, but we usually choose α between .05 and .01. A frequent choice for α is .05. In fact, our former student tells us that the network often tests advertising claims by setting the probability of a Type I error equal to .05. That is, the network will run a commercial making a claim if the sample evidence allows it to reject a null hypothesis that says the claim is not valid in favor of an alternative hypothesis that says the claim is valid with α set equal to .05. Since a Type I error is deciding that the claim is valid when it is not, the policy of setting α equal to .05 says that, in the long run, the network will advertise only 5 percent of all invalid claims made by advertisers.

One might wonder why the network does not set α lower—say at .01. One reason is that **it can be shown that, for a fixed sample size, the lower we set $\boldsymbol{\alpha}$, the higher is $\boldsymbol{\beta}$, and the higher we set $\boldsymbol{\alpha}$, the lower is $\boldsymbol{\beta}$.** Setting α at .05 means that β, the probability of failing to advertise a true claim (a Type II error), will be smaller than it would be if α were set at .01. As long as (1) the claim to be advertised is plausible and (2) the consequences of advertising the claim even if it is false are not terribly serious, then it is reasonable to set α equal to .05. However, if either (1) or (2) is not true, then we might set α lower than .05. For example, suppose a pharmaceutical company wishes to advertise that it has developed an effective treatment for a disease that has formerly been very resistant to treatment. Such a claim is (perhaps) difficult to believe. Moreover, if the claim is false, patients suffering from the disease would be subjected to false hope and needless expense. In such a case, it might be reasonable for the network to set α at .01 because this would lower the chance of advertising the claim if it is false. We usually do not set α lower than .01 because doing so often leads to an unacceptably large value of β. We explain some methods for computing the probability of a Type II error in optional Section 8.6. However, β can be difficult or impossible to calculate in many situations, and we often must rely on our intuition when deciding how to set α.

Exercises for Section 8.1

CONCEPTS

8.1 Which hypothesis (the null hypothesis, H_0, or the alternative hypothesis, H_a) is the "status quo" hypothesis (that is, the hypothesis that states that things are remaining "as is")? Which hypothesis is the hypothesis that says that a "hoped for" or "suspected" condition exists?

8.2 Which hypothesis (H_0 or H_a) is not rejected unless there is convincing sample evidence that it is false? Which hypothesis (H_0 or H_a) will be accepted only if there is convincing sample evidence that it is true?

8.4, 8.8

8.3 Define each of the following:
a Type I error **b** Type II error **c** α **d** β

8.4 For each of the following situations, indicate whether an error has occurred and, if so, indicate what kind of error (Type I or Type II) has occurred.
a We do not reject H_0 and H_0 is true.
b We reject H_0 and H_0 is true.
c We do not reject H_0 and H_0 is false.
d We reject H_0 and H_0 is false.

8.5 Suppose that when we test a hypothesis
a We reject H_0. In this case, what is the only type of error that we could be making? Explain.
b We do not reject H_0. In this case, what is the only type of error that we could be making? Explain.

8.6 When testing a hypothesis, why don't we set the probability of a Type I error to be extremely small? Explain.

METHODS AND APPLICATIONS

8.7 **THE CUSTOMER SATISFACTION RATING CASE** CustSat

Recall that "very satisfied" customers typically give a rating of the VAC-5000 vacuum cleaner that exceeds 42. Suppose that the manufacturer of the VAC-5000 wishes to use the 65 satisfaction ratings to provide evidence supporting the claim that the mean composite satisfaction rating for the VAC-5000 exceeds 42.
a Letting μ represent the mean composite satisfaction rating for the VAC-5000, set up the null and alternative hypotheses needed if we wish to attempt to provide evidence supporting the claim that μ exceeds 42.
b In the context of this situation, interpret making a Type I error; interpret making a Type II error.

8.8 **THE BANK CUSTOMER WAITING TIME CASE** WaitTime

Recall that a bank manager has developed a new system to reduce the time customers spend waiting for teller service during peak hours. The manager hopes the new system will reduce waiting times from the current 9 to 10 minutes to less than 6 minutes.

Suppose the manager wishes to use the 100 waiting times to support the claim that the mean waiting time under the new system is shorter than six minutes.
a Letting μ represent the mean waiting time under the new system, set up the null and alternative hypotheses needed if we wish to attempt to provide evidence supporting the claim that μ is shorter than six minutes.
b In the context of this situation, interpret making a Type I error; interpret making a Type II error.

8.9 Recall that, in an article in *Accounting and Business Research,* Carslaw and Kaplan (1991) studied factors that influence audit delay for firms in New Zealand. One such factor studied by Carslaw and Kaplan is whether a firm is owner controlled or manager controlled. Suppose we wish to support the claim that the mean audit delay for all public owner-controlled companies in New Zealand is smaller than 90 days.
a Letting μ represent the mean audit delay for all public owner-controlled companies in New Zealand, set up the null and alternative hypotheses needed if we wish to provide evidence supporting the claim that μ is smaller than 90 days.
b In the context of this situation, interpret making a Type I error; interpret making a Type II error.

8.10 An automobile parts supplier owns a machine that produces a cylindrical engine part. This part is supposed to have an outside diameter of three inches. Parts with diameters that are too small or too large do not meet customer requirements and must be rejected. Lately, the company has experienced problems meeting customer requirements. The technical staff feels that the mean diameter produced by the machine is off target. In order to verify this, a special study will randomly sample 100 parts produced by the machine. The 100 sampled parts will be measured, and if the results obtained cast a substantial amount of doubt on the hypothesis that the mean diameter equals the target value of three inches, the company will assign a problem-solving team to intensively search for the causes of the problem.
a The parts supplier wishes to set up a hypothesis test so that the problem-solving team will be assigned when the null hypothesis is rejected. Set up the null and alternative hypotheses for this situation.
b In the context of this situation, interpret making a Type I error; interpret making a Type II error.
c Suppose it costs the company \$3,000 a day to assign the problem-solving team to a project. Is this \$3,000 figure the daily cost of a Type I error or a Type II error? Explain.

8.11 The Crown Bottling Company has just installed a new bottling process that will fill 16-ounce bottles of the popular Crown Classic Cola soft drink. Both overfilling and underfilling bottles are undesirable: Underfilling leads to customer complaints and overfilling costs the company considerable money. In order to verify that the filler is set up correctly, the company wishes to see whether the mean bottle fill, μ, is close to the target fill of 16 ounces. To this end, a random sample of 36 filled bottles is selected from the output of a test filler run. If the sample results cast a substantial amount of doubt on the hypothesis that the mean bottle fill is the desired 16 ounces, then the filler's initial setup will be readjusted.

a The bottling company wants to set up a hypothesis test so that the filler will be readjusted if the null hypothesis is rejected. Set up the null and alternative hypotheses for this hypothesis test.

b In the context of this situation, interpret making a Type I error; interpret making a Type II error.

8.12 Consolidated Power, a large electric power utility, has just built a modern nuclear power plant. This plant discharges waste water that is allowed to flow into the Atlantic Ocean. The Environmental Protection Agency (EPA) has ordered that the waste water may not be excessively warm so that thermal pollution of the marine environment near the plant can be avoided. Because of this order, the waste water is allowed to cool in specially constructed ponds and is then released into the ocean. This cooling system works properly if the mean temperature of waste water discharged is 60°F or cooler. Consolidated Power is required to monitor the temperature of the waste water. A sample of 100 temperature readings will be obtained each day, and if the sample results cast a substantial amount of doubt on the hypothesis that the cooling system is working properly (the mean temperature of waste water discharged is 60°F or cooler), then the plant must be shut down and appropriate actions must be taken to correct the problem.

a Consolidated Power wishes to set up a hypothesis test so that the power plant will be shut down when the null hypothesis is rejected. Set up the null and alternative hypotheses that should be used.

b In the context of this situation, interpret making a Type I error; interpret making a Type II error.

c The EPA periodically conducts spot checks to determine whether the waste water being discharged is too warm. Suppose the EPA has the power to impose very severe penalties (for example, very heavy fines) when the waste water is excessively warm. Other things being equal, should Consolidated Power set the probability of a Type I error equal to $\alpha = .01$ or $\alpha = .05$? Explain.

d Suppose Consolidated Power has been experiencing technical problems with the cooling system. Because the system has been unreliable, the company feels it must take precautions to avoid failing to shut down the plant when its waste water is too warm. Other things being equal, should Consolidated Power set the probability of a Type I error equal to $\alpha = .01$ or $\alpha = .05$? Explain.

8.13 THE DISK BRAKE CASE

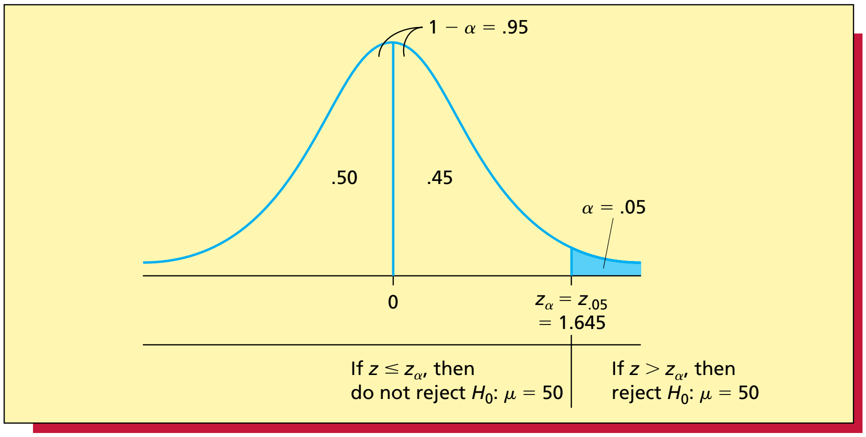
National Motors has equipped the ZX-900 with a new disk brake system. We define the stopping distance for a ZX-900 as the distance (in feet) required to bring the automobile to a complete stop from a speed of 35 mph under normal driving conditions using this new brake system. In addition, we define μ to be the mean stopping distance of all ZX-900s. One of the ZX-900's major competitors is advertised to achieve a mean stopping distance of 60 ft. National Motors would like to claim in a new television commercial that the ZX-900 achieves a shorter mean stopping distance. The standards and practices division of a major television network will permit National Motors to run the commercial if H_0: $\mu \geq 60$ can be rejected in favor of H_a: $\mu < 60$ by setting $\alpha = .05$. Interpret what it means to set α at .05.

8.2 Large Sample Tests about a Population Mean: One-Sided Alternatives

CHAPTER 9

Testing a "greater than" alternative hypothesis by using a rejection point Recall in the trash bag case that the manufacturer will randomly select a sample of $n = 40$ trash bags and calculate the mean $\bar{x}$ and the standard deviation s of the breaking strengths of these trash bags. In order to test H_0: $\mu \leq 50$ versus H_a: $\mu > 50$ by setting the probability of a Type I error equal to α, we will test the modified null hypothesis H_0: $\mu = 50$ versus H_a: $\mu > 50$. The idea here is that if there is sufficient evidence to reject the hypothesis that μ equals 50 in favor of $\mu > 50$, then there

FIGURE 8.2 The Rejection Point for Testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ by Setting $\alpha = .05$



is certainly also sufficient evidence to reject the hypothesis that μ is less than or equal to 50. In order to test H_0: $\mu = 50$ versus H_a: $\mu > 50$, we utilize the **test statistic**

$$z = \frac{\bar{x} - 50}{\sigma_{\bar{x}}} = \frac{\bar{x} - 50}{\sigma/\sqrt{n}}$$

Here, since σ is unknown and the sample size is large, σ will be estimated by s. A positive value of the test statistic results from an $\bar{x}$ that is greater than 50 and thus provides evidence against H_0: $\mu = 50$ and in favor of H_a: $\mu > 50$. To decide how large the test statistic must be to reject H_0 in favor of H_a by setting the probability of a Type I error equal to α, we use the following two-step procedure:

1 Place the probability of a Type I error, α, in the right-hand tail of the standard normal curve and use the normal table (see Table A.3, page 814) to find the normal point z_α. Here z_α, which we call a **rejection point** (or **critical point**), is the point on the horizontal axis under the standard normal curve that gives a right-hand tail area equal to α.

2 Reject H_0: $\mu = 50$ in favor if H_a: $\mu > 50$ if and only if the test statistic z is greater than the rejection point z_α.

Figure 8.2 illustrates that, if we set $\alpha = .05$, then we use the rejection point $z_\alpha = z_{.05} = 1.645$ (see Table A.3). This says that we should reject H_0 if $z > 1.645$. To understand this procedure, recall that because the sample size $n = 40$ is large, the Central Limit Theorem tells us that the sampling distribution of $(\bar{x} - \mu)/\sigma_{\bar{x}}$ is (approximately) a standard normal distribution. It follows that if the null hypothesis H_0: $\mu = 50$ is true, then the sampling distribution of the test statistic $z = (\bar{x} - 50)/\sigma_{\bar{x}}$ is (approximately) a standard normal distribution. Therefore, using the rejection point z_α says that (as illustrated in Figure 8.2):

1 If H_0: $\mu = 50$ is true, the probability that z will be less than or equal to z_α is $(1 - \alpha)$. For instance, if we set α equal to .05, then, when μ equals 50, $100(1 - \alpha)\% = 95\%$ of all possible values of z are less than or equal to $z_{.05} = 1.645$. This says that 95 percent of all possible values of z would tell us to not reject H_0: $\mu = 50$ when H_0 is true—a correct decision.

2 If H_0: $\mu = 50$ is true, the probability that z will be greater than z_α is α. For instance, if we set α equal to .05, then, when μ equals 50, $100\alpha\% = 5\%$ of all possible values of z are greater than $z_{.05} = 1.645$. This says that 5 percent of all possible values of z would tell us to reject H_0: $\mu = 50$ when H_0 is true—a Type I error.

Therefore, using the rejection point z_α implies that the probability of Type I error equals α.

Suppose that the trash bag manufacturer randomly samples 40 of the newly developed trash bags and tests them for breaking strength. Using the results, the sample mean and sample standard deviation are calculated to be $\bar{x} = 50.575$ and $s = 1.6438$. We have seen that if we set $\alpha = .05$, then we use the rejection point $z_{.05} = 1.645$. Because

$$z = \frac{\bar{x} - 50}{s/\sqrt{n}} = \frac{50.575 - 50}{1.6438/\sqrt{40}} = 2.2123$$

is greater than $z_{.05} = 1.645$, we can reject H_0: $\mu = 50$ in favor of H_a: $\mu > 50$. Therefore, we conclude (at an α of .05) that the mean breaking strength of the new trash bag exceeds 50 pounds. It follows that the television network will allow the manufacturer to run a commercial claiming that the new trash bag is stronger than the former bag. Furthermore, we can be intuitively confident that H_0: $\mu = 50$ is false and H_a: $\mu > 50$ is true. This is because, since we have rejected H_0 by setting α equal to .05, we have rejected H_0 by using a test that allows only a 5 percent chance of wrongly rejecting H_0. Note, however, that the sample mean $\bar{x} = 50.575$ suggests that μ is not much larger than 50. Therefore, the trash bag manufacturer can claim only that its new bag is slightly stronger than its former bag. Of course, this might be important to many consumers who feel that, because the new bag is 25 percent less expensive and is more environmentally sound, it is definitely worth purchasing if it has any strength advantage.

Considerations in setting α We have reasoned in Section 8.1 that the television network has set α equal to .05 rather than .01 because doing so means that β, the probability of failing to advertise a true claim (a Type II error), will be smaller than it would be if α were set at .01. It is informative, however, to see what would have happened if the network had set α equal to .01. Figure 8.3 illustrates that as we decrease α from .05 to .01, the rejection point z_α increases from $z_{.05} = 1.645$ to $z_{.01} = 2.33$. Because the test statistic value $z = 2.2123$ is less than $z_{.01} = 2.33$, we cannot reject H_0: $\mu = 50$ in favor of H_a: $\mu > 50$ by setting α equal to .01. This illustrates the point that, the smaller we set α, the larger is the rejection point, and thus the stronger is the statistical evidence that we are requiring to reject the null hypothesis H_0. Some statisticians have concluded (somewhat subjectively) that (1) **if we set α equal to .05, then we are requiring strong evidence to reject H_0;** and (2) **if we set α equal to .01, then we are requiring very strong evidence to reject H_0.** These considerations will be useful in future applications of hypothesis testing.

FIGURE 8.3 The Rejection Points for Testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ by Setting $\alpha = .05$ and .01

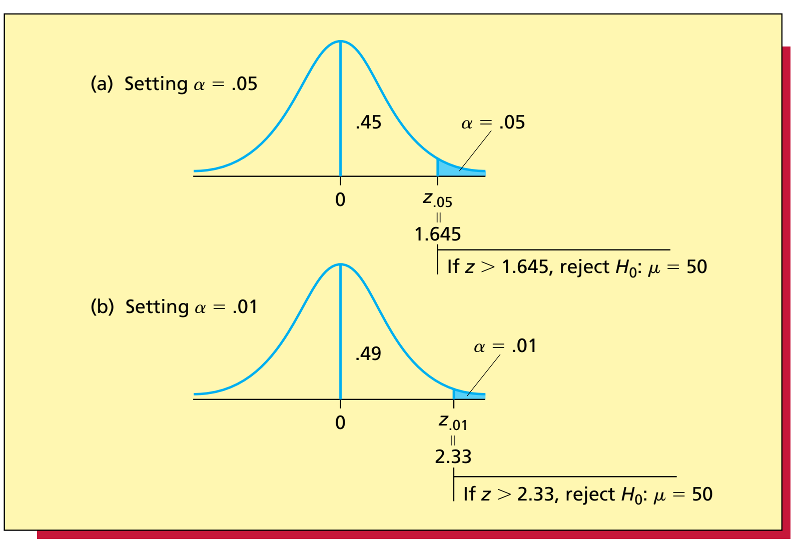
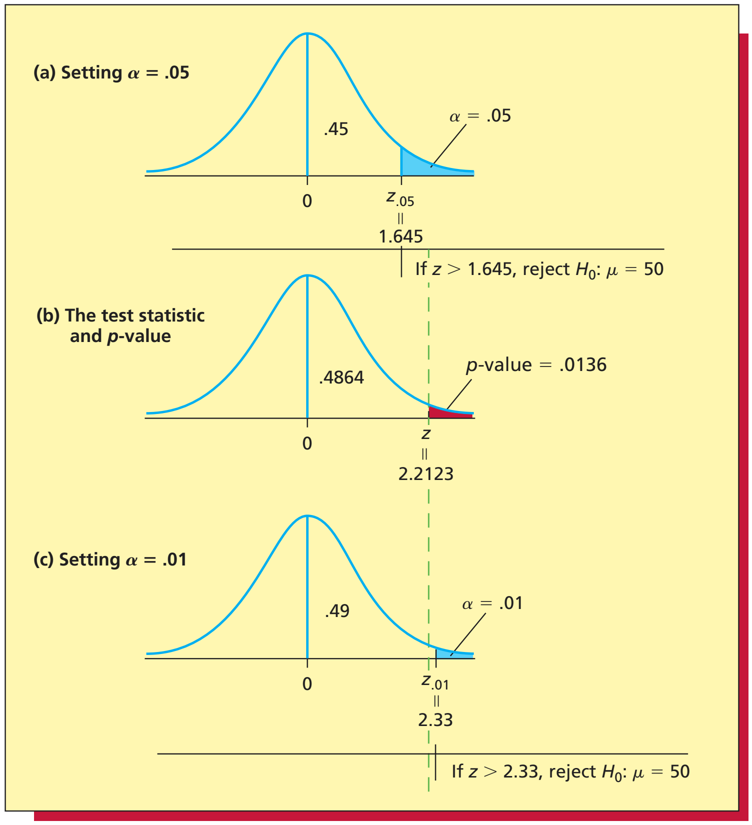


FIGURE 8.4 Testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ by Using Rejection Points and the p-Value



A *p*-value for testing a "greater than" alternative hypothesis Another approach to hypothesis testing is based on calculating a ***p*-value,** which is a measure of how likely the sample results are if the null hypothesis H_0 is true. Sample results that are not likely if H_0 is true are evidence that H_0 is not true. The p-value for testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ in the trash bag case is the area under the standard normal curve to the right of the test statistic value $z = 2.2123$. As illustrated in Figure 8.4(b), this area is $.5 - .4864 = .0136$. The p-value is the probability, computed assuming that H_0: $\mu = 50$ is true, of observing a value of the test statistic that is greater than or equal to the value $z = 2.2123$ that we have actually computed from the sample data. The p-value of .0136 says that, if H_0: $\mu = 50$ is true, then only 136 in 10,000 of all possible test statistic values are at least as large, or extreme, as the value $z = 2.2123$. That is, if we are to believe that H_0 is true, we must believe that we have observed a test statistic value that can be described as a 136 in 10,000 chance. Because it is difficult to believe that we have observed a 136 in 10,000 chance, we intuitively have strong evidence that H_0: $\mu = 50$ is false and H_a: $\mu > 50$ is true.

In addition to its interpretation as a probability, the p-value can be used to decide whether we can reject H_0: $\mu = 50$ in favor of H_a: $\mu > 50$ **by setting the probability of a Type I error equal to α** (or, as we will sometimes say, **at level of significance α**). To understand this, suppose we set α equal to .05. Then, comparing the two normal curves in Figures 8.4(a) and (b), we see that since the ***p*-value of .0136 is less than the α of .05,** the test statistic $z = 2.2123$ is greater than the rejection point $z_{.05} = 1.645$. Therefore, we can **reject H_0 by setting α equal to .05.** As another example, suppose that we set α equal to .01. Then, comparing the two normal curves in Figures 8.4(b) and (c), we see that since the ***p*-value of .0136 is greater than the α of .01,** the test statistic $z = 2.2123$ is less than the rejection point $z_{.01} = 2.33$. Therefore, we **cannot reject H_0 by setting α equal to .01.** Generalizing these examples, we conclude that the test statistic z will be greater than the rejection point z_α if and only if the p-value is less than α. That is, **we can reject H_0 in favor of H_a at level of significance α if and only if the p-value is less than α.**

Thus far we have considered two methods for testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ at the .05 and .01 values of α. Using the first method, we determine if the test statistic value $z = 2.2123$ is greater than the rejection points $z_{.05} = 1.645$ and $z_{.01} = 2.33$. Using the second method, we determine if the p-value of .0136 is less than .05 and .01. Whereas the rejection point method requires that we look up a different rejection point for each different α value, the p-value method requires only that we calculate a single p-value and compare it directly with the different α values. It follows that the p-value method is the most efficient way to test a hypothesis at different α values. This can be useful when there are different decision makers who might use different α values. For example, television networks do not always evaluate advertising claims by setting α equal to .05. The reason is that the consequences of a Type I error (advertising a false claim) are more serious for some claims than for others. For example, the consequences of a Type I error would be fairly serious for a claim about the effectiveness of a drug or for a claim that the product of one company is better than the product of another company. However, these consequences might not be as serious for a noncomparative claim about an inexpensive and safe product, such as a cosmetic.[2] Networks sometimes use α values between .01 and .04 for claims having more serious Type I error consequences, and they sometimes use α values between .06 and .10 for claims having less serious Type I error consequences. Furthermore, the policies for setting α of one network can differ somewhat from those of another. As a result, reporting an advertising claim's p-value to each network is the most efficient way to tell the network whether to allow the claim to be advertised. For example, most networks would evaluate the trash bag claim by choosing an α value between .025 and .10. Since the p-value of .0136 is less than all these α values, most networks would allow the trash bag claim to be advertised.

Because a single p-value can help different decision makers to make their own independent decisions, statistical software packages use p-values to report the results of hypothesis tests (as we will begin to see in Section 8.4). However, the rejection point approach also has advantages. One is that understanding rejection points helps us to better understand the probability of a Type I error (and the probability of a Type II error—see optional Section 8.6). Furthermore, in some situations (for example, in Section 8.4) statistical tables are not complete enough to calculate the p-value, and a computer software package or an electronic calculator with statistical capabilities is needed. If these tools are not immediately available, rejection points can be used to carry out hypothesis tests, because statistical tables are almost always complete enough to give the needed rejection points. Throughout this book we will continue to present both the rejection point and the p-value approaches to hypothesis testing.

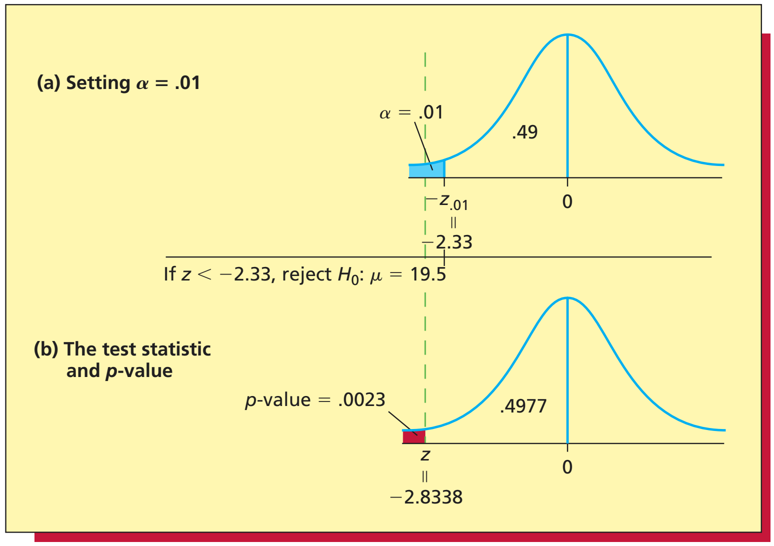
Testing a "less than" alternative hypothesis Consider the accounts receivable case. In order to study whether the new electronic billing system reduces the mean bill payment time by more than 50 percent, the management consulting firm will use the random sample of $n = 65$ payment times to test H_0: $\mu \geq 19.5$ versus H_a: $\mu < 19.5$. To perform this test, we test the modified null hypothesis H_0: $\mu = 19.5$ versus H_a: $\mu < 19.5$. The idea here is that if there is sufficient evidence to reject the hypothesis that μ equals 19.5 in favor of $\mu < 19.5$, then there is certainly also sufficient evidence to reject the hypothesis that μ is greater than or equal to 19.5. In order to test H_0: $\mu = 19.5$ versus H_a: $\mu < 19.5$, we compute the sample mean $\bar{x}$ and the sample standard deviation s of the payment times and utilize the test statistic

$$z = \frac{\bar{x} - 19.5}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 19.5}{s/\sqrt{n}}$$

A value of the test statistic z that is less than zero results when $\bar{x}$ is less than 19.5. This provides evidence to support rejecting H_0 in favor of H_a because the point estimate $\bar{x}$ indicates that μ might be less than 19.5. To decide how much less than zero the test statistic must be to reject H_0 in favor

[2]The authors wish to thank CBS and NBC for telephone conversations providing the examples and other information used in this paragraph. NBC also sent us a 14-page advertising standards booklet. The NBC logo on the booklet's cover is shown on the page margin.

FIGURE 8.5 Testing H_0: μ = 19.5 versus H_a: μ < 19.5 by Using Rejection Points and the *p*-Value



of H_a by setting the probability of a Type I error equal to α, we use the following two-step procedure:

1. Place the probability of a Type I error, α, in the left-hand tail of the standard normal curve and use the normal table to find the rejection point $-z_\alpha$. Here $-z_\alpha$ is the negative of the normal point z_α. That is, $-z_\alpha$ is the point on the horizontal axis under the standard normal curve that gives a left-hand tail area equal to α.
2. Reject H_0: $\mu = 19.5$ in favor of H_a: $\mu < 19.5$ if and only if the test statistic z is less than the rejection point $-z_\alpha$.

For example, suppose the management consulting firm wants to be very sure that it truthfully describes the benefits of the new system both to the company in which it has been installed and to other companies that are considering installing such a system. Therefore, the firm will require very strong evidence to conclude that μ is less than 19.5, which implies that it will test H_0: μ = 19.5 versus H_a: $\mu < 19.5$ by setting α equal to .01. Figure 8.5(a) illustrates that the rejection point $-z_\alpha$ for the hypothesis test is $-z_{.01} = -2.33$. The mean and the standard deviation of the random sample of 65 payment times are $\bar{x} = 18.1077$ and $s = 3.9612$. Because

$$z = \frac{\bar{x} - 19.5}{s/\sqrt{n}} = \frac{18.1077 - 19.5}{3.9612/\sqrt{65}} = -2.8338$$

is less than $-z_{.01} = -2.33$, we can reject H_0: $\mu = 19.5$ in favor of H_a: $\mu < 19.5$ by setting α equal to .01. Therefore, we conclude (at an α of .01) that the mean payment time for the new electronic billing system is less than 19.5 days. This, along with the fact that the sample mean $\bar{x} = 18.1077$ is slightly less than 19.5, implies that it is reasonable for the management consulting firm to conclude that the new electronic billing system has reduced the mean bill payment time by slightly more than 50 percent (a substantial improvement over the old system).

We next note from Figure 8.5(b) that the area under the standard normal curve to the left of the test statistic value $z = -2.8338$ is $.5 - .4977 = .0023$. This area is the *p*-value for the hypothesis test. The *p*-value is the probability, computed assuming that H_0: $\mu = 19.5$ is true, of observing a value of the test statistic that is less than or equal to the value $z = -2.8338$ that we have actually computed from the sample data. The *p*-value of .0023 says that, if H_0: $\mu = 19.5$ is true, then only 23 in 10,000 of all possible test statistic values are at least as negative, or extreme, as the value $z = -2.8338$. That is, if we are to believe that H_0 is true, we must believe that we have observed a test statistic value that can be described as a 23 in 10,000 chance.

In addition to its interpretation as a probability, the p-value can be used to decide whether we can reject H_0: $\mu = 19.5$ in favor of H_a: $\mu < 19.5$ at level of significance α. To understand this, suppose that we set α equal to .01. Then, comparing the two normal curves in Figures 8.5(a) and (b), we see that since the **p-value of .0023 is less than the α of .01,** the test statistic $z = -2.8338$ is less than the rejection point $-z_{.01} = -2.33$. Therefore, we can **reject H_0 by setting α equal to .01.** In general, the test statistic z will be less than the rejection point $-z_\alpha$ if and only if the p-value is less than α. That is, **we can reject H_0 in favor of H_a at level of significance α if and only if the p-value is less than α.**

A summary of testing a one-sided alternative hypothesis As illustrated in the previous examples, we can test two types of one-sided alternative hypotheses. First, we sometimes wish to test H_0: $\mu \le \mu_0$ versus H_a: $\mu > \mu_0$, where μ_0 is a specific number. By the same reasoning used to test H_0: $\mu \le 50$ versus H_a: $\mu > 50$ in the trash bag case, we test H_0: $\mu \le \mu_0$ versus H_a: $\mu > \mu_0$, by testing H_0: $\mu = \mu_0$ versus H_a: $\mu > \mu_0$. Second, we sometimes wish to test H_0: $\mu \ge \mu_0$ versus H_a: $\mu < \mu_0$. By the same reasoning used to test H_0: $\mu \ge 19.5$ versus H_a: $\mu <$ 19.5 in the accounts receivable case, we test H_0: $\mu \ge \mu_0$ versus H_a: $\mu < \mu_0$ by testing H_0: $\mu = \mu_0$ versus H_a: $\mu < \mu_0$. To summarize, we may think of testing a one-sided alternative hypothesis about a population mean as testing H_0: $\mu = \mu_0$ versus H_a: $\mu > \mu_0$ or as testing H_0: $\mu = \mu_0$ versus H_a: $\mu < \mu_0$. In addition, as illustrated in the previous examples, there is a rejection point condition that tells us whether we can reject H_0: $\mu = \mu_0$ in favor of a particular one-sided alternative hypothesis **at level of significance α** (that is, **by setting the probability of a Type I error equal to α**). We summarize the rejection point conditions and the p-values in the following box.

A Hypothesis Test about a Population Mean: Testing H_0: $\mu = \mu_0$ versus a One-Sided Alternative Hypothesis

Define the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

and assume that the population sampled is normally distributed, or that the sample size n is large. We can reject H_0: $\mu = \mu_0$ in favor of a particular alternative hypothesis at level of significance α if and only if the appropriate rejection point condition holds, or, equivalently, the corresponding p-value is less than α.

Alternative Hypothesis	Rejection Point Condition: Reject H_0 if	p-Value
H_a: $\mu > \mu_0$	$z > z_\alpha$	The area under the standard normal curve to the right of z
H_a: $\mu < \mu_0$	$z < -z_\alpha$	The area under the standard normal curve to the left of z

Note that if σ is unknown, and n is large, we estimate σ by s.

We next present a five-step procedure that can help us to use the summary box.

The Five Steps of Hypothesis Testing

1. Determine the null and alternative hypotheses that are appropriate for the hypothesis testing application being considered.
2. Specify the level of significance α (that is, the probability of a Type I error) for the test.
3. Select the test statistic that will be used to decide whether to reject the null hypothesis. Collect the sample data and compute the value of the test statistic.

(steps continued on next page)

4 **a** If you wish to use a rejection point condition, use the summary box to find the rejection point condition corresponding to the alternative hypothesis. Use the specified value of α to find the rejection point given in the rejection point condition. Reject the null hypothesis at level of significance α if a comparison of the test statistic value with the rejection point shows that the rejection point condition holds.

b If you wish to use a *p*-value, use the summary box to find the *p*-value corresponding to the alternative hypothesis. Calculate the *p*-value by using the test statistic value. Reject the null hypothesis at level of significance α if the *p*-value is less than α.

5 Interpret your statistical result in managerial (real-world) terms.

For example, consider the trash bag case. Step 1 Since we wish to show that the mean breaking strength μ is greater than 50, we will test the null hypothesis H_0: $\mu \leq 50$ versus the alternative hypothesis H_a: $\mu > 50$. Step 2 The television network requires that we specify a level of significance of $\alpha = .05$. Step 3 Since the sample size $n = 40$ is large, the summary box tells us that the test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 50}{\sigma/\sqrt{n}}$$

where we estimate σ by s. When we select the $n = 40$ trash bags, we find that $\bar{x} = 50.575$ and $s = 1.6438$. Therefore, the value of the test statistic is

$$z = \frac{50.575 - 50}{1.6438/\sqrt{40}} = 2.2123$$

Step 4a If we wish to use a rejection point condition, the summary box tells us, since the alternative hypothesis H_a: $\mu > 50$ is of the form H_a: $\mu > \mu_0$, that we should reject H_0 if $z > z_\alpha$. Because $\alpha = .05$, the rejection point z_α is $z_{.05} = 1.645$. Since the test statistic value $z = 2.2123$ is greater than the rejection point $z_{.05} = 1.645$, we reject H_0 at level of significance $\alpha = .05$. Step 4b If we wish to use a *p*-value, the summary box tells us, since the alternative hypothesis H_a: $\mu > 50$ is of the form H_a: $\mu > \mu_0$, that the *p*-value is the area under the standard normal curve to the right of the test statistic value $z = 2.2123$. This area equals $.5 - .4864 = .0136$. Because the *p*-value of .0136 is less than the α of .05, we reject H_0 at level of significance $\alpha = .05$. Step 5 We conclude (at an α of .05) that the mean breaking strength of the new trash bag exceeds 50 pounds and that the manufacturer's commercial will be run on the television network.

Measuring the weight of evidence against the null hypothesis In general, the decision to take an action (for example, run the trash bag commercial) is sometimes based solely on whether there is sufficient sample evidence to reject a null hypothesis (H_0: $\mu = 50$) by setting α equal to a single, prespecified value (.05). In such situations, it is often also useful to know all of the information—called the **weight of evidence**—that the hypothesis test provides against the null hypothesis and in favor of the alternative hypothesis. For example, the trash bag manufacturer would almost certainly wish to know *how much* evidence there is that its new bag is stronger than its former bag. Furthermore, although we tested a hypothesis in the accounts receivable case by setting α equal to a single, prespecified value, the hypothesis test did not immediately lead to a decision as to whether to take an action. In a situation such as this, when hypothesis testing is used more as a way to achieve evolving understanding of an industrial or scientific process, it is particularly important to know the weight of evidence against the null hypothesis and in favor of the alternative hypothesis.

The most informative way to measure the weight of evidence is to use the *p*-value. For every hypothesis test considered in this book we can interpret the *p*-value to be the **probability, computed assuming that the null hypothesis H_0 is true, of observing a value of the test statistic that is at least as extreme as the value actually computed from the sample data.** The **smaller**

the *p*-value is, the less likely are the sample results if the null hypothesis H_0 is true. Therefore, the stronger is the evidence that H_0 is false and that the alternative hypothesis H_a is true. We can use the *p*-value to test H_0 versus H_a at level of significance α as follows:

We reject H_0 in favor of H_a at level of significance α if and only if the *p*-value is less than α.

Experience with hypothesis testing has resulted in statisticians making the following (somewhat subjective) conclusions:

Interpreting the Weight of Evidence against the Null Hypothesis

If the *p*-value for testing H_0 is less than

- .10, we have **some evidence** that H_0 is false.
- .05, we have **strong evidence** that H_0 is false.
- .01, we have **very strong evidence** that H_0 is false.
- .001, we have **extremely strong evidence** that H_0 is false.

For example, recall that the *p*-value for testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ in the trash bag case is .0136. This *p*-value is less than .05 but not less than .01. Therefore, we have strong evidence, but not very strong evidence, that H_0: $\mu = 50$ is false and H_a: $\mu > 50$ is true. That is, we have strong evidence that the mean breaking strength of the new trash bag exceeds 50 pounds. As another example, the *p*-value for testing H_0: $\mu = 19.5$ versus H_a: $\mu < 19.5$ in the accounts receivable case is .0023. This *p*-value is less than .01 but not less than .001. Therefore, we have very strong evidence, but not extremely strong evidence, that H_0: $\mu = 19.5$ is false and H_a: $\mu < 19.5$ is true. That is, we have very strong evidence that the new billing system has reduced the mean payment time to less than 19.5 days.

To conclude this section we note that, while many statisticians believe in assessing the weight of evidence, other statisticians do not. These other statisticians, who might be called **decision theorists,** contend that measuring the weight of evidence lets the results of a single sample bias our view too much as to the relative validity of H_0 and H_a. Decision theorists feel that, even in establishing your own personal belief, you should make a choice between H_0 and H_a by setting α equal to a single value that is prespecified before the sample is taken. Even for decision theorists, however, the *p*-value has the advantage of being the most efficient way to report the results of a hypothesis test to different decision makers who might use different prespecified values of α. In this book we will continue to assess the weight of evidence, and the decision theorist can regard the *p*-value as simply an efficient way to report the results of a hypothesis test.

Exercises for Section 8.2

CONCEPTS

8.14 Explain what a rejection point is, and explain how it is used to test a hypothesis.

8.15 Explain what a *p*-value is, and explain how it is used to test a hypothesis.

www.mhhe.com/bowerman3e 8.31, 8.33

METHODS AND APPLICATIONS

In Exercises 8.16 through 8.22 we consider using a random sample of 100 measurements to test H_0: $\mu = 80$ versus H_a: $\mu > 80$.

8.16 If $\bar{x} = 85$ and $s = 20$, calculate the value of the test statistic z.

8.17 Use a rejection point to test H_0 versus H_a by setting α equal to .10.

8.18 Use a rejection point to test H_0 versus H_a by setting α equal to .05.

8.19 Use a rejection point to test H_0 versus H_a by setting α equal to .01.

8.20 Use a rejection point to test H_0 versus H_a by setting α equal to .001.

8.21 Calculate the *p*-value and use it to test H_0 versus H_a at each of α = .10, .05, .01, and .001.

8.22 How much evidence is there that H_0: $\mu = 80$ is false and H_a: $\mu > 80$ is true?

In Exercises 8.23 through 8.29 we consider using a random sample of 49 measurements to test H_0: $\mu = 20$ versus H_a: $\mu < 20$.

8.23 If $\bar{x} = 18$ and $s = 7$, calculate the value of the test statistic z.

8.24 Use a rejection point to test H_0 versus H_a by setting α equal to .10.

8.25 Use a rejection point to test H_0 versus H_a by setting α equal to .05.

8.26 Use a rejection point to test H_0 versus H_a by setting α equal to .01.

8.27 Use a rejection point to test H_0 versus H_a by setting α equal to .001.

8.28 Calculate the p-value and use it to test H_0 versus H_a at each of $\alpha = .10, .05, .01$, and .001.

8.29 How much evidence is there that H_0: $\mu = 20$ is false and H_a: $\mu < 20$ is true?

8.30 THE CUSTOMER SATISFACTION RATING CASE CustSat

Recall from Exercise 8.7 (page 306) that "very satisfied" customers typically give a rating of the VAC-5000 vacuum cleaner that exceeds 42. Letting μ be the mean composite satisfaction rating for the VAC-5000, we found in Exercise 8.7 that we should test H_0: $\mu \leq 42$ versus H_a: $\mu > 42$ in order to attempt to provide evidence supporting the claim that μ exceeds 42. The random sample of 65 satisfaction ratings yields a sample mean of $\bar{x} = 42.954$ and a sample standard deviation of $s = 2.6424$.

a Use rejection points to test H_0 versus H_a at each of $\alpha = .10, .05, .01$, and .001.
b Calculate the p-value and use it to test H_0 versus H_a at each of $\alpha = .10, .05, .01$, and .001.
c How much evidence is there that a typical customer is very satisfied?

8.31 THE BANK CUSTOMER WAITING TIME CASE WaitTime

Letting μ be the mean waiting time under the new system, we found in Exercise 8.8 (page 306) that we should test H_0: $\mu \geq 6$ versus H_a: $\mu < 6$ in order to attempt to provide evidence that μ is less than six minutes. The random sample of 100 waiting times yields a sample mean of $\bar{x} = 5.46$ minutes and a sample standard deviation of $s = 2.475$. Moreover, Figure 8.6 gives the MINITAB output obtained when we use the waiting time data to test H_0: $\mu = 6$ versus H_a: $\mu < 6$. On this output the label "SE Mean," which stands for "the standard error of the mean," denotes the quantity $s/\sqrt{n}$, and the label "Z" denotes the calculated test statistic.

a Use rejection points to test H_0 versus H_a at each of $\alpha = .10, .05, .01$, and .001.
b Calculate the p-value and verify that it equals .0146 (.015 rounded), as shown on the MINITAB output. Use the p-value to test H_0 versus H_a at each of $\alpha = .10, .05, .01$, and .001.
c How much evidence is there that the new system has reduced the mean waiting time to below six minutes?

8.32 Again consider the audit delay situation originally discussed in Exercise 7.11 (page 263). Letting μ be the mean audit delay for all public owner-controlled companies in New Zealand, we found in Exercise 8.9 that we should test H_0: $\mu \geq 90$ versus H_a: $\mu < 90$ in order to attempt to provide evidence supporting the claim that μ is less than 90 days. Suppose that a random sample of 100 public owner-controlled companies in New Zealand is found to give a mean audit delay of $\bar{x} = 86.6$ days with a standard deviation of $s = 32.83$ days. Calculate the p-value for testing H_0 versus H_a and determine how much evidence there is that the mean audit delay for all public-owner controlled companies in New Zealand is less than 90 days.

8.33 THE CAR MILEAGE CASE GasMiles

Suppose the federal government proposes to give a substantial tax break to automakers producing midsize cars that get a mean mileage exceeding 31 mpg. Letting μ be a midsize car's mean mileage, the government will award the tax credit if an automaker is able to reject H_0: $\mu \leq 31$ in favor of H_a: $\mu > 31$ at the .05 level of significance. Recall that the sample of 49 mileages has mean $\bar{x} = 31.5531$ and standard deviation $s = .7992$. Use these sample results and a rejection point to test H_0 versus H_a at the .05 level of significance. Will the automaker be awarded the tax

FIGURE 8.6 MINITAB Output of the Test of H_0: $\mu = 6$ versus H_a: $\mu < 6$ in the Bank Customer Waiting Time Case

```
Test of mu = 6.000 vs mu < 6.000. The assumed sigma = 2.47.
Variable       N      Mean     StDev    SE Mean         Z        P
WaitTime     100     5.460     2.475      0.247     -2.18    0.015
```

Note: Because the test statistic z has a theoretical denominator $\sigma/\sqrt{n}$ that uses the population standard deviation σ, MINITAB makes the user specify an assumed value for σ. We have used $s = 2.475$ as the assumed value for σ.

break? Calculate the p-value. Would the tax break have been awarded if the federal government had set α equal to .01? Justify your answer.

8.34 Consider the Consolidated Power waste water situation and the hypotheses you set up in Exercise 8.12 (page 307). Recall the power plant will be shut down and corrective action will be taken on the cooling system if the null hypothesis H_0: $\mu \leq 60$ is rejected in favor of H_a: $\mu > 60$. Suppose Consolidated Power decides to use a level of significance of $\alpha = .05$, and suppose a random sample of 100 temperature readings is obtained. For each of the following sample results, determine whether the power plant should be shut down and the cooling system repaired.

a $\bar{x} = 60.482$ and $s = 2$. **b** $\bar{x} = 60.262$ and $s = 2$. **c** $\bar{x} = 60.618$ and $s = 2$.

8.3 Large Sample Tests about a Population Mean: Two-Sided Alternatives

Testing a "not equal to" alternative hypothesis In the camshaft case, the quality control analyst will randomly select $n = 35$ camshafts from the population of all camshafts produced on a particular day. The analyst will measure the hardness depth of each camshaft and use the resulting sample of 35 hardness depths to test H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$. To perform this test, we compute the sample mean $\bar{x}$ and the sample standard deviation s of the hardness depths and utilize the test statistic

$$z = \frac{\bar{x} - 4.5}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 4.5}{s/\sqrt{n}}$$

A value of the test statistic that is greater than 0 results when $\bar{x}$ is greater than 4.5. This provides evidence to support rejecting H_0 in favor of H_a because the point estimate $\bar{x}$ indicates that μ might be greater than 4.5. Similarly, a value of the test statistic that is less than 0 results when $\bar{x}$ is less than 4.5. This also provides evidence to support rejecting H_0 in favor of H_a because the point estimate $\bar{x}$ indicates that μ might be less than 4.5. To decide how much greater than 0 or less than 0 the test statistic value must be to reject H_0 in favor of H_a by setting the probability of a Type I error equal to α, we use the following two-step procedure:

1. Divide the probability of a Type I error, α, into two equal parts, and place the area $\alpha/2$ in the right-hand tail of the standard normal curve and the area $\alpha/2$ in the left-hand tail of the standard normal curve. Then use the normal table to find the rejection points $z_{\alpha/2}$ and $-z_{\alpha/2}$. Here $z_{\alpha/2}$ is the point on the horizontal axis under the standard normal curve that gives a right-hand tail area equal to $\alpha/2$, and $-z_{\alpha/2}$ is the negative of $z_{\alpha/2}$.
2. Reject H_0: $\mu = 4.5$ in favor of H_a: $\mu \neq 4.5$ if and only if the test statistic z is greater than the rejection point $z_{\alpha/2}$ or less than the rejection point $-z_{\alpha/2}$. Note that this rejection point condition is equivalent to saying that the absolute value of the test statistic, $|z|$, is greater than the rejection point $z_{\alpha/2}$.

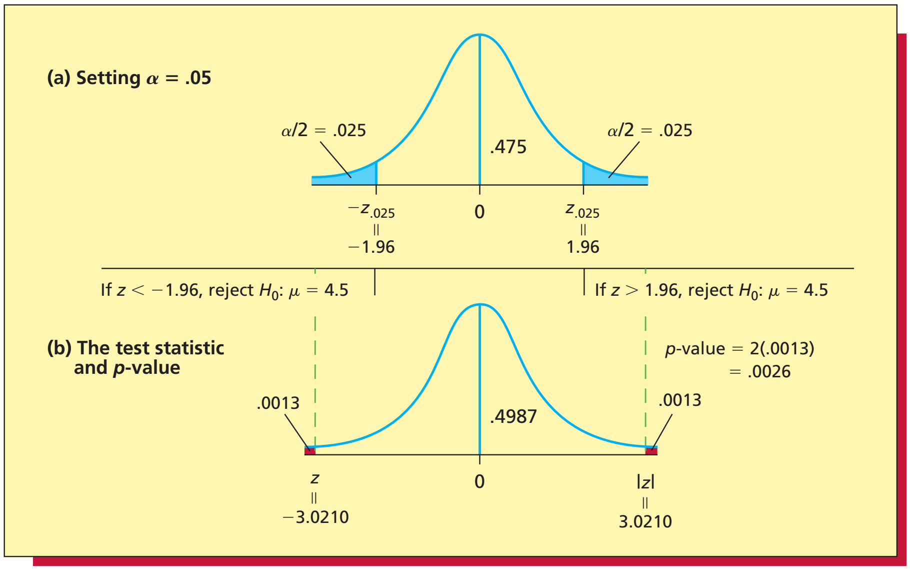
For example, assume that the quality control analyst will test H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$ by setting α equal to .05. To understand this choice of α, recall that the camshaft hardening process has not been meeting specifications. For this reason, the quality control analyst has decided that it is very important to avoid committing a Type II error. That is, it is very important to avoid failing to reject H_0: $\mu = 4.5$ if μ for the day's production does differ from 4.5 mm. Setting α equal to .05 rather than .01 makes the probability of this Type II error smaller than it would be if α were set at .01. Figure 8.7(a) illustrates that the rejection points for the hypothesis test are

$$z_{\alpha/2} = z_{.05/2} = z_{.025} = 1.96 \qquad \text{and} \qquad -z_{\alpha/2} = -z_{.025} = -1.96$$

Suppose that the quality control analyst randomly selects $n = 35$ camshafts from the day's production of camshafts and calculates the mean and the standard deviation of the hardness depths of these camshafts to be $\bar{x} = 4.26$ and $s = .47$. Because

$$z = \frac{\bar{x} - 4.5}{s/\sqrt{n}} = \frac{4.26 - 4.5}{.47/\sqrt{35}} = -3.0210$$

FIGURE 8.7 Testing H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$ by Using Rejection Points and the p-Value



is less than $-z_{.025} = -1.96$ (or, equivalently, because $|z| = 3.0210$ is greater than $z_{.025} = 1.96$), we can reject H_0: $\mu = 4.5$ in favor of H_a: $\mu \neq 4.5$ by setting α equal to .05. Therefore, we conclude (at an α of .05) that the mean camshaft hardness depth differs from 4.5 mm. Furthermore, because $\bar{x}$ equals 4.26 mm., it appears that μ is less than 4.5. This being said, it is important to ask if μ is far enough below 4.5 to matter from a practical standpoint. One way to approach this question is to assess whether the difference between μ and 4.5 is large enough to contribute to the production of individual camshafts that fail to meet hardness depth specifications. We will investigate this issue later in this section.

We next note from Figure 8.7(b) that the area under the standard normal curve to the right of $|z| = 3.0210$ is .0013. Twice this area—that is, 2(.0013) = .0026—is the p-value for the hypothesis test. To interpret the p-value as a probability, note that the symmetry of the standard normal curve implies that twice the area under the curve to the right of $|z| = 3.0210$ equals the area under this curve to the right of 3.0210 plus the area under the curve to the left of -3.0210 [see Figure 8.7(b)]. Also, note that since both positive and negative test statistic values count against H_0: $\mu = 4.5$, a test statistic value that is either greater than or equal to 3.0210 or less than or equal to -3.0210 is at least as extreme as the observed test statistic value $z = -3.0210$. It follows that the p-value of .0026 says that, if H_0: $\mu = 4.5$ is true, then only 26 in 10,000 of all possible test statistic values are at least as extreme as $z = -3.0210$. That is, if we are to believe that H_0 is true, we must believe that we have observed a test statistic value that can be described as a 26 in 10,000 chance.

In addition to its interpretation as a probability, the p-value can be used to decide whether we can reject H_0: $\mu = 4.5$ in favor of H_a: $\mu \neq 4.5$ at level of significance α. To understand this, suppose that we set α equal to .05. Then, since the **p-value of .0026 is less than the α of .05,** it follows that the area under the standard normal curve to the right of $|z| = 3.0210$, which is one-half the p-value and equals .0013, is less than $\alpha/2 = .025$. Comparing the two normal curves in Figures 8.7(a) and (b), we see that this implies that $|z| = 3.0210$ is greater than the rejection point $z_{.025} = 1.96$. Therefore, we can **reject H_0 by setting α equal to .05.** In general, the absolute value of the test statistic z will be greater than the rejection point $z_{\alpha/2}$ if and only if the p-value is less than α. That is, **we can reject H_0 in favor of H_a at level of significance α if and only if the p-value is less than α.** For example, the p-value of .0026 is less than an α of .01 but not less than an α of .001. It follows that we have very strong evidence, but not extremely strong evidence, that H_0: $\mu = 4.5$ is false and H_a: $\mu \neq 4.5$ is true. That is, we have very strong evidence that the mean camshaft hardness depth differs from 4.5 mm.

A summary of testing H_0: $\mu = \mu_0$ In the following box we summarize how to use rejection points and p-values to test H_0: $\mu = \mu_0$ versus either H_a: $\mu > \mu_0$, H_a: $\mu < \mu_0$, or H_a: $\mu \neq \mu_0$:

Testing a Hypothesis about a Population Mean: Testing H_0: $\mu = \mu_0$

Define the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

and assume that the population sampled is normally distributed, or that the sample size n is large. We can reject H_0: $\mu = \mu_0$ in favor of a particular alternative hypothesis at level of significance α if and only if the appropriate rejection point condition holds, or, equivalently, the corresponding p-value is less than α.

Alternative Hypothesis	Rejection Point Condition: Reject H_0 if	p-Value
H_a: $\mu > \mu_0$	$z > z_\alpha$	The area under the standard normal curve to the right of z
H_a: $\mu < \mu_0$	$z < -z_\alpha$	The area under the standard normal curve to the left of z
H_a: $\mu \neq \mu_0$	$\lvert z \rvert > z_{\alpha/2}$—that is, $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	Twice the area under the standard normal curve to the right of $\lvert z \rvert$

Note that if σ is unknown, and n is large, we estimate σ by s.

We now illustrate using this summary box and the five hypothesis testing steps presented in the last section. Suppose that airline flights arriving at a metropolitan airport from the New York/Boston corridor have routinely experienced delays for the last several years. Last year, these delayed flights were an average of 35 minutes late. In order to see whether this average has changed, an airport administrator will randomly sample 36 New York/Boston flights that were delayed during the last two months. Step 1 We wish to decide whether μ, the mean number of minutes that delayed New York/Boston flights were late during the last two months, differs from last year's average of 35 minutes. Therefore, we will test H_0: $\mu = 35$ versus H_a: $\mu \neq 35$. Step 2 Since the administrator does not expect to take immediate action based on the test results, the consequences of a Type I error are not particularly serious. Thus, the administrator will specify a level of significance of $\alpha = .10$. Step 3 Since the sample size $n = 36$ is large, the summary box tells us that the test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - 35}{\sigma / \sqrt{n}}$$

where we estimate σ by s. When the random sample of 36 delayed flights is selected and the length of each flight's delay is determined, we find that $\bar{x} = 33$ minutes and $s = 12$ minutes. Therefore, the value of the test statistic is

$$z = \frac{33 - 35}{12 / \sqrt{36}} = -1$$

Step 4a If we wish to use a rejection point condition, the summary box tells us, since the alternative hypothesis H_a: $\mu \neq 35$ is of the form H_a: $\mu \neq \mu_0$, that we should reject H_0 if $\lvert z \rvert > z_{\alpha/2}$. Because $\alpha = .10$, the rejection point is $z_{\alpha/2} = z_{.10/2} = z_{.05} = 1.645$. Since $\lvert z \rvert = \lvert -1 \rvert = 1$ is less than $z_{.05} = 1.645$, we cannot reject H_0 at level of significance $\alpha = .10$. Step 4b If we wish to use a p-value, the summary box tells us, since the alternative hypothesis H_a: $\mu \neq 35$ is of the form H_a: $\mu \neq \mu_0$, that the p-value is twice the area under the standard normal curve to the right of $\lvert z \rvert = 1$. Using Table A.3 (page 814), we find that the p-value equals $2(.5 - .3413) = 2(.1587) = .3174$. Because the p-value is not less than the α of .10, we cannot reject H_0: $\mu = 35$ at level of significance $\alpha = .10$. Step 5 The airport administrator concludes (at an α of .10) that there is insufficient evidence to conclude that the mean delay μ differs from last year's 35-minute figure.

Using confidence intervals to test hypotheses Confidence intervals can be used to test hypotheses. Specifically, it can be proven that we can reject H_0: $\mu = \mu_0$ in favor of H_a: $\mu \neq \mu_0$ by setting the probability of a Type I error equal to α if and only if the $100(1 - \alpha)$ percent confidence interval for μ does not contain μ_0. For example, consider the camshaft case and testing H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$ by setting α equal to .05. To do this, we use the mean $\bar{x} = 4.26$ and the standard deviation $s = .47$ of the sample of $n = 35$ camshafts to calculate the 95 percent confidence interval for μ to be

$$\left[\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\right] = \left[4.26 \pm 1.96\frac{.47}{\sqrt{35}}\right] = [4.10, 4.42]$$

Because this interval does not contain 4.5, we can reject H_0: $\mu = 4.5$ in favor of H_a: $\mu \neq 4.5$ by setting α equal to .05.

A confidence interval that has both a **lower limit** and an **upper limit** is called a **two-sided confidence interval.** All of the confidence intervals that we study in this book are two-sided. Furthermore, we have just seen that we can test H_0: $\mu = \mu_0$ versus the two-sided alternative hypothesis H_a: $\mu \neq \mu_0$ by using the two-sided confidence interval for μ. If we wish to use a confidence interval to test H_0: $\mu = \mu_0$ versus either of the one-sided alternative hypotheses H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$, we need to use a **one-sided confidence interval.** We will not study one-sided confidence intervals in this book. However, it should be emphasized that we do not need to use confidence intervals (one-sided or two-sided) to test hypotheses. We can test hypotheses by using test statistics and rejection points or p-values. Furthermore, as illustrated in the next subsection, the meaningful application of confidence intervals is in helping to evaluate **practical importance** by estimating the population mean after we have established **statistical significance** by performing a hypothesis test.

Statistical significance versus practical importance If we can reject a null hypothesis by setting the probability of a Type I error equal to α, we say that we have **statistical significance at the α level.** Whether we have statistical significance at a given level often depends greatly on the size of the sample we have selected. To see this, recall that the trash bag manufacturer wishes to test H_0: $\mu \leq 50$ versus H_a: $\mu > 50$ and has obtained the sample mean $\bar{x} = 50.575$ and the sample standard deviation $s = 1.6438$ based on a sample of $n = 40$ trash bags. Since the p-value associated with

$$z = \frac{\bar{x} - 50}{s/\sqrt{n}} = \frac{50.575 - 50}{1.6438/\sqrt{40}} = \frac{.575}{.2599} = 2.2123$$

is .0136, we have statistical significance at the .05 level but not at the .01 level. However, suppose that the manufacturer had obtained the same $\bar{x}$ and s based on a larger sample of $n = 100$ new bags. The test statistic value is then

$$z = \frac{\bar{x} - 50}{s/\sqrt{n}} = \frac{50.575 - 50}{1.6438/\sqrt{100}} = \frac{.575}{.16438} = 3.4980$$

and the p-value is the area under the standard normal curve to the right of $z = 3.4980$. Looking at the normal table (see Table A.3, page 814), we see that the area to the right of 3.09 is $.5 - .4990 = .001$. Therefore, because 3.4980 is greater than 3.09, the p-value is less than .001, and we have statistical significance at the .001 level. Intuitively, this is saying that, if H_0: $\mu = 50$ is true, then chance alone would be extremely unlikely to produce such a large test statistic value. What has happened here is that the numerator of the test statistic has remained the same, while the larger sample size makes the denominator of the test statistic smaller. This results in a larger and more statistically significant value of the test statistic. Intuitively, the larger sample size makes us surer that H_a: $\mu > 50$ is true.

On the basis of statistical significance at the .001 level, some practitioners would report that "there is a highly statistically significant difference between the mean breaking strength of the new bag and the mean breaking strength of the former bag." This statement can be easily misinterpreted. **Statistical significance** at the .001 level simply means that $\bar{x}$ is far enough above 50 so that, when the test statistic z is calculated, it allows us to reject H_0 by setting α equal to .001. In other words, **statistical significance** means only that $\bar{x}$ (when used in the test statistic) **signifies** that μ is greater than 50. It does not necessarily mean, however, that the difference between μ and 50 is large. In fact, the sample mean $\bar{x} = 50.575$ indicates that μ is not much larger than 50. *As far as potential customers are concerned,* it is doubtful that a difference of about .575 pounds would represent a large increase in bag strength.

Suppose, however, that when the manufacturer undertakes a research and development program to further improve bag breaking strengths, a sample of 40 bags yields $\bar{x} = 58.213$ and $s = 1.4286$. Because the p-value associated with

$$z = \frac{\bar{x} - 50}{s/\sqrt{n}} = \frac{58.213 - 50}{1.4286/\sqrt{40}} = 36.3598$$

is less than .001, we have statistical significance at the .001 level. Furthermore, $\bar{x} = 58.213$ says that we estimate that μ exceeds 50 by 8.213 pounds. The 95 percent confidence interval for μ

$$\left[\bar{x} \pm z_{.025}\frac{s}{\sqrt{n}}\right] = \left[58.213 \pm 1.96\frac{1.4286}{\sqrt{40}}\right]$$
$$= [57.7703,\ 58.6557]$$

says that we are 95 percent confident that the newest trash bag increases mean breaking strength by between 7.7703 and 8.6557 pounds. The increases in mean breaking strength implied by the point estimate and confidence interval would probably represent fairly large increases in strength to many customers.

In general, whether a statistically significant result has **practical importance** is a personal decision. Different individuals will make different assessments of practical importance. For example, concluding that the new trash bag has any strength advantage (that is, concluding that H_a: $\mu > 50$ is true) certainly has practical importance to the trash bag manufacturer because this means that the television network will approve running commercials claiming that the new bag is stronger. This conclusion might also have practical importance to consumers who feel that, because the new bag is less expensive and more environmentally sound, it is definitely worth purchasing if it has any strength advantage. Therefore, to the manufacturer and to these consumers, both the sample with mean $\bar{x} = 50.575$ and the sample with mean $\bar{x} = 58.213$ give statistically significant and practically important results. On the other hand, some consumers may feel that only a reasonably large increase in mean breaking strength has practical importance. To these consumers, it is quite likely that only the sample with mean $\bar{x} = 58.213$ would represent a practically important result.

When conducting a hypothesis test, the alternative hypothesis usually makes a statement that is of some practical importance to the person who has formulated the hypothesis test. It follows that, as far as this person is concerned, a finding of statistical significance (that is, concluding that the alternative hypothesis is true) usually has some practical importance. Other people, however, may disagree. Therefore, when we are confronted with a statistically significant result, we should also carefully investigate practical importance by using the point estimate of and a confidence interval for the parameter (mean or the like) involved in the hypothesis test. The point estimate and confidence interval tell us by how much the parameter differs from the value specified in the null hypothesis. Furthermore, as just illustrated in the trash bag example, whether the alternative hypothesis is one-sided or two-sided, we can use a two-sided confidence interval to help estimate the population parameter. Once we have evaluated the size of the difference between the parameter and the value specified in the null hypothesis, the decision as to whether this difference is practically important becomes subjective; different people might reach different conclusions. For instance, a particular estimated increase in breaking strength might be important to one person, but not to another.

To conclude this subsection, we consider two other examples of evaluating practical importance. First, consider the accounts receivable case and testing H_0: $\mu = 19.5$ versus H_a: $\mu < 19.5$. Since the management consulting firm hopes that the new electronic billing system reduces the mean bill payment time by more than 50 percent, the alternative hypothesis H_a: $\mu < 19.5$ expresses what is, to the consulting firm, a practically important result. Therefore, because the p-value associated with testing H_0 versus H_a is .0023, strong evidence supporting this practically important result exists. To fully evaluate practical importance, we need to estimate μ. The point estimate $\bar{x} = 18.1077$ says we estimate that the mean bill payment time is 18.1077 days. The

95 percent confidence interval for μ, [17.1, 19.1], says that we are 95 percent confident that the mean bill payment time is between 17.1 days and 19.1 days. Second, consider the camshaft case and testing H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$. The p-value of .0026 says that we have very strong evidence that μ does not equal 4.5. To fully evaluate practical importance, note that the sample mean $\bar{x} = 4.26$ says that we estimate that μ is .24 mm less than the target value of 4.5 mm. The 95 percent confidence interval for μ, [4.10, 4.42], says that we are 95 percent confident that μ is somewhere between .40 mm and .08 mm less than the target value of 4.5 mm. Unfortunately, however, the differences implied by the point estimate and confidence interval do not alone give us a good sense of whether the difference between μ and 4.5 is practically important. To investigate further, recall that specifications state that each individual hardness depth should be between 3 mm and 6 mm. Using $\bar{x} = 4.26$ and $s = .47$, we estimate that 99.73 percent of all individual camshaft hardness depths are in the interval $[\bar{x} \pm 3s] = [4.26 \pm 3(.47)] = [2.85, 5.67]$. This estimated tolerance interval says that, because of a somewhat small $\bar{x}$ of 4.26 and a somewhat large s of .47, we estimate that we are producing some hardness depths that are below the lower specification limit of 3 mm. This is a practically important result. In Chapter 14 we will learn how to use **statistical process control** to improve the camshaft hardening process.

Exercises for Section 8.3

www.mhhe.com/bowerman3e

8.45, 8.48

CONCEPTS

8.35 Explain how rejection points and the p-value are used to test a two-sided alternative hypothesis.

8.36 Discuss the difference between statistical significance and practical importance.

METHODS AND APPLICATIONS

In Exercises 8.37 through 8.43 we consider using a random sample of $n = 81$ measurements to test H_0: $\mu = 40$ versus H_a: $\mu \neq 40$.

8.37 If $\bar{x} = 34$ and $s = 18$, calculate the value of the test statistic z.

8.38 Use rejection points to test H_0 versus H_a by setting α equal to .10.

8.39 Use rejection points to test H_0 versus H_a by setting α equal to .05.

8.40 Use rejection points to test H_0 versus H_a by setting α equal to .01.

8.41 Use rejection points to test H_0 versus H_a by setting α equal to .001.
Hint: $z_{.0005}$ can be shown to equal 3.29.

8.42 Calculate the p-value and use it to test H_0 versus H_a at each of α = .10, .05, .01, and .001.

8.43 How much evidence is there that H_0: $\mu = 40$ is false and H_a: $\mu \neq 40$ is true?

8.44 Consider the automobile parts supplier and the hypotheses you set up in Exercise 8.10 (page 306). Suppose that a problem-solving team will be assigned to rectify the process producing cylindrical engine parts if the null hypothesis H_0: $\mu = 3$ can be rejected in favor of H_a: $\mu \neq 3$ by setting α equal to .05.

a If a sample of 40 parts yields a sample mean diameter of $\bar{x} = 3.006$ inches and a sample standard deviation equal to $s = .016$ inches, use rejection points and a p-value to test H_0 versus H_a by setting α equal to .05. Should the problem-solving team be assigned?

b Suppose that product specifications state that each and every part must have a diameter between 2.95 and 3.05 inches—that is, the specifications are 3" $\pm$ .05". Use the sample information given in part a to estimate an interval that contains almost all (99.73 percent) of the diameters. Compare this estimated interval with the specification limits. Are the specification limits being met, or are some diameters outside the specification limits? Explain.

8.45 Consider the Crown Bottling Company fill process and the hypotheses you set up in Exercise 8.11 (page 307). Recall that the initial setup of the filler will be readjusted if the null hypothesis H_0: $\mu = 16$ is rejected in favor of H_a: $\mu \neq 16$. Suppose that Crown Bottling Company decides to use a level of significance of $\alpha = .01$, and suppose a random sample of 36 bottle fills is obtained from a test run of the filler. For each of the following sample results, determine whether the filler's initial setup should be readjusted. In each case perform the hypothesis test by using rejection points and a p-value.

a $\bar{x} = 16.05$ and $s = .1$. **b** $\bar{x} = 15.96$ and $s = .1$. **c** $\bar{x} = 16.02$ and $s = .1$.

8.46 Again consider the Crown Bottling Company fill process. Suppose a random sample of 36 bottle fills yields a sample mean of $\bar{x} = 16.05$ and a sample standard deviation of $s = .1$. Use this sample information to compute a 95 percent confidence interval for the mean fill. Then, use this interval to test H_0: $\mu = 16$ versus H_a: $\mu \neq 16$ by setting $\alpha = .05$. What do you conclude? What considerations might influence your decision as to whether the result has practical importance?

8.47 Recall that, as discussed in Exercise 7.12 (page 263), in an article in the *Journal of Marketing,* Bayus (1991) studied differences between early and late replacement buyers. One variable investigated by Bayus is the number of auto dealers visited during the replacement process.

a Letting μ be the mean number of dealers visited by early replacement buyers, suppose that we wish to test H_0: $\mu = 4$ versus H_a: $\mu \neq 4$. If a random sample of 800 early replacement buyers yields a mean number of dealers visited of $\bar{x} = 3.3$ with a standard deviation of $s = .71$, calculate the p-value for the hypothesis test. Do we estimate that μ is less than 4 or greater than 4?

b Letting μ be the mean number of dealers visited by late replacement buyers, suppose that we wish to test H_0: $\mu = 4$ versus H_a: $\mu \neq 4$. If a random sample of 500 late replacement buyers yields a mean number of dealers visited of $\bar{x} = 4.3$ with a standard deviation of $s = .66$, calculate the p-value for the hypothesis test. Do we estimate that μ is less than 4 or greater than 4?

8.48 THE DISK BRAKE CASE

Recall from Exercise 8.13 (page 307) that the television network will permit National Motors to claim that the ZX-900 achieves a shorter mean stopping distance if H_0: $\mu \geq 60$ can be rejected in favor of H_a: $\mu < 60$ by setting α equal to .05. If the stopping distances of a random sample of $n = 81$ ZX-900s have a mean of $\bar{x} = 57.8$ ft and a standard deviation of $s = 6.02$ ft, will National Motors be allowed to run the commercial? Calculate a 95 percent confidence interval for μ. Do the point estimate of μ and confidence interval for μ indicate that μ might be far enough below 60 feet to suggest that we have a practically important result?

8.4 Small Sample Tests about a Population Mean

In order to employ the hypothesis tests in Sections 8.2 and 8.3, the sampling distribution of

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

must be (at least approximately) a standard normal distribution. This will be true if the population sampled is normally distributed or if the sample size is large. Since the value of σ is rarely known, and because the sample size must be large (at least 30) to replace σ by s in the above formula, the methods of Sections 8.2 and 8.3 are most often used when we take a large sample. If we take a small sample and do not know σ, we can base a hypothesis test about μ on the sampling distribution of

CHAPTER 9

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

If the sampled population is normally distributed (or at least mound-shaped), then this sampling distribution is a **t distribution having $n - 1$ degrees of freedom.** This leads to the following results:

A Small Sample Test about a Population Mean: Testing H_0: $\mu = \mu_0$

Define the test statistic

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

and assume that the population sampled is normally distributed. We can reject H_0: $\mu = \mu_0$ in favor of a particular alternative hypothesis at level of significance α if and only if the appropriate rejection point condition holds, or, equivalently, the corresponding p-value is less than α.

Alternative Hypothesis	Rejection Point Condition: Reject H_0 if	p-Value
H_a: $\mu > \mu_0$	$t > t_\alpha$	The area under the t distribution curve to the right of t
H_a: $\mu < \mu_0$	$t < -t_\alpha$	The area under the t distribution curve to the left of t
H_a: $\mu \neq \mu_0$	$\lvert t \rvert > t_{\alpha/2}$—that is, $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$	Twice the area under the t distribution curve to the right of $\lvert t \rvert$

Here t_α, $t_{\alpha/2}$, and the p-values are based on $n - 1$ degrees of freedom.

EXAMPLE 8.4

In 1991 the average interest rate charged by U.S. credit card issuers was 18.8 percent. Since that time, there has been a proliferation of new credit cards affiliated with retail stores, oil companies, alumni associations, professional sports teams, and so on. A financial officer wishes to study whether the increased competition in the credit card business has reduced interest rates. In order to carry out this study, the officer randomly selects a sample of 15 credit cards and determines their interest rates. The interest rates for the 15 sampled cards are given in Table 8.3. A stem-and-leaf display and MINITAB box plot are given in Figure 8.8. The stem-and-leaf display looks reasonably mound-shaped, and both the stem-and-leaf display and the box plot look reasonably symmetrical. Letting μ denote the current mean interest rate, the financial officer will study whether the current mean rate is lower than the mean interest rate charged in 1991 by using the sample of 15 interest rates to test H_0: $\mu = 18.8\%$ versus H_a: $\mu < 18.8\%$. If we set α equal to .05, then, since there are $n - 1 = 15 - 1 = 14$ degrees of freedom, we use the rejection point $-t_{.05} = -1.761$ [see Figure 8.9(a)]. The mean and standard deviation of the interest rates in Table 8.3 are $\bar{x} = 16.827$ and $s = 1.538$. Because

$$t = \frac{\bar{x} - 18.8}{s/\sqrt{n}} = \frac{16.827 - 18.8}{1.538/\sqrt{15}} = -4.97$$

is less than $-t_{.05} = -1.761$, we can reject H_0: $\mu = 18.8\%$ in favor of H_a: $\mu < 18.8\%$ by setting α equal to .05. The p-value for testing H_0 versus H_a is the area under the curve of the t distribution having 14 degrees of freedom to the left of -4.97 [see Figure 8.9(b)]. Tables of t points (such as Table A.4, page 815) are not complete enough to give such areas for the majority of t statistic values, so we use computer software packages to calculate p-values that are based on the t distribution. For example, the MINITAB output in Figure 8.10(a) and the MegaStat output in Figure 8.11 tell us that the p-value for testing H_0: $\mu = 18.8\%$ versus H_a: $\mu < 18.8\%$ is .0001. As another example, since Excel can be used to calculate areas under t distribution curves, Excel can be used to calculate p-values. To illustrate, the p-value for testing H_0: $\mu = 18.8\%$ versus H_a: $\mu <$ 18.8% is the area under the t distribution curve having 14 degrees of freedom to the left of -4.97. This area equals the area under this t distribution curve to the right of 4.97, which Excel tells us is .000103 [see Figure 8.10(b)]. Because the p-value of .0001 (rounded) is less than .05, .01, and .001, we can reject H_0 at the .05, .01, and .001 levels of significance. Notice that the p-value of .0001 on the MegaStat output is shaded dark yellow. This indicates that we can reject H_0 at the .01 level of significance (light yellow shading would indicate significance at the .05 level, but not at the .01 level). As a probability, the p-value of .0001 says that if we are to believe that H_0: $\mu =$ 18.8% is true, we must believe that we have observed a t statistic value ($t = -4.97$) that can be described as a 1 in 10,000 chance. In summary, we have extremely strong evidence that H_0: $\mu =$ 18.8% is false and H_a: $\mu < 18.8\%$ is true. That is, we have extremely strong evidence that the current mean credit card interest rate is less than 18.8%. Therefore, the financial officer concludes

TABLE 8.3 Interest Rates Charged by 15 Randomly Selected Credit Cards CreditCD

15.6%	15.3%	19.2%
17.8	16.4	15.8
14.6	18.4	18.1
17.3	17.6	16.6
18.7	14.0	17.0

FIGURE 8.8 Stem-and-Leaf Display and Box Plot of the Interest Rates

14	0 6
15	3 6 8
16	4 6
17	0 3 6 8
18	1 4 7
19	2

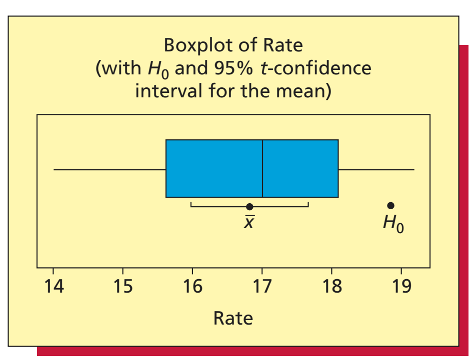


FIGURE 8.9 Testing H_0: μ = 18.8% versus H_a: μ < 18.8% by Using a Rejection Point and a p-Value

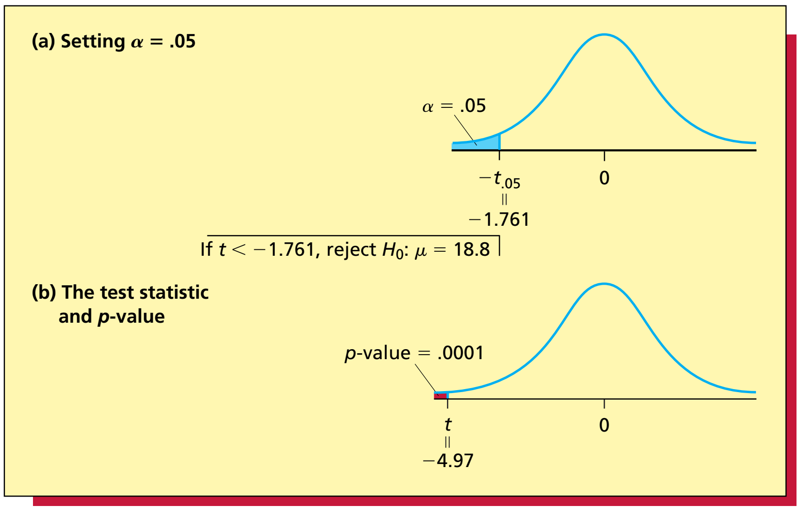


FIGURE 8.10 The MINITAB and Excel Outputs for Testing H_0: μ = 18.8% versus H_a: μ < 18.8%

(a) The MINITAB output

```
Test of mu = 18.800 vs mu < 18.800

Variable    N     Mean   StDev   SE Mean       T        P
Rate       15   16.827   1.538     0.397   -4.97   0.0001
```

(b) The Excel output

t-statistic
4.97
p-value
0.000103

FIGURE 8.11 The MegaStat Output for Testing H_0: μ = 18.8% versus H_a: μ < 18.8%

Hypothesis Test: Mean vs. Hypothesized Value

18.8000 hypothesized value	1.5378 std. dev.	15 n	**−4.97** t
16.8267 mean Rate	0.3971 std. error	14 df	**.0001** p-value (one-tailed, lower)

= Significant at .05 level   = Significant at .01 level

that the current mean credit card interest rate is lower than the mean interest rate in 1991. Furthermore, the sample mean $\bar{x}$ = 16.827 says that we estimate the mean interest rate is 18.8% − 16.827% = 1.973% lower than it was in 1991.

We have seen (in Section 8.3) that a large sample can give a statistically significant result that is not practically important. On the other hand, if a sample is too small, we might not be able to reject a null hypothesis involving a population parameter even though its point estimate suggests that the alternative hypothesis might be true. This is one reason why we say that "we do not reject H_0" rather than saying that "we accept H_0." For example, suppose in the camshaft case that a random sample of n = 5 camshafts yields $\bar{x}$ = 4.82 and s = .43. In order to test H_0: μ = 4.5 versus H_a: $\mu \neq 4.5$ by setting α = .05, we compare the absolute value of the test statistic t with the rejection point $t_{.05/2} = t_{.025} = 2.776$ (based on $n - 1 = 4$ degrees of freedom). Since the absolute value of

$$t = \frac{\bar{x} - 4.5}{s/\sqrt{n}} = \frac{4.82 - 4.5}{.43/\sqrt{5}} = 1.6641$$

is not greater than $t_{.025} = 2.776$, we cannot reject H_0 by setting $\alpha = .05$ (the reader can also verify that we cannot reject H_0 by setting α equal to .10, .01, or .001). Here, although the sample mean $\bar{x} = 4.82$ suggests that H_a: $\mu \neq 4.5$ might be true, the sample size $n = 5$ is not sufficiently large to make the denominator of t small enough to yield a value of t that is large enough to tell us to reject H_0. Intuitively, the value of the test statistic is telling us that the sample mean and sample standard deviation are not based on enough observations to provide strong evidence that H_a: $\mu \neq 4.5$ is true. However, this certainly does not provide strong evidence supporting H_0: $\mu = 4.5$ either. A larger sample might provide strong evidence against H_0.

Because the t test (the test about μ based on the t distribution) is a statistically correct test that does not require that we know σ, it can be argued that it is best, if we do not know σ, to use this test for any size sample—even a large sample. For example, in the trash bag case we have tested H_0: $\mu = 50$ versus H_a: $\mu > 50$ by calculating the test statistic

$$z = \frac{\bar{x} - 50}{s/\sqrt{n}} = \frac{50.575 - 50}{1.6438/\sqrt{40}} = 2.2123$$

and using the large sample z test (the test about μ based on the normal distribution). Since we do not know σ, it might be best to regard this test statistic as a t statistic and use the t test. If we do this, the p-value for testing H_0 versus H_a is the area under the t distribution curve having $n - 1 = 39$ degrees of freedom to the right of 2.2123. Using Excel, we find that this p-value is .016435, which is slightly larger than the p-value of .0136 calculated in Section 8.2 using the normal distribution. In general, when the sample size n is large (say, at least 30), the t test and the large sample z test give similar results, and thus it is reasonable to use the large sample z test as an approximation to the t test. Furthermore, the Central Limit Theorem tells us that, if n is at least 30, then both tests are approximately valid no matter how the sampled population is distributed. If the sample size n is small (less than 30) and we do not know σ, then we must use the t test, which is valid for small samples if the sampled population is normally distributed (or at least mound-shaped). If the sample size n is small and the sampled population is not mound-shaped, or if the sampled population is highly skewed, then it might be appropriate to use a **nonparametric test about the population median.** Such a test is discussed in Chapter 15.

Exercises for Section 8.4

8.52, 8.55, 8.56, 8.59, 8.60

CONCEPTS

8.49 What assumptions must be met in order to carry out the small sample test about a population mean based on the t distribution?

8.50 Suppose that the t statistic for testing H_0: $\mu = 20$ versus H_a: $\mu > 20$ based on a sample of size $n = 7$ equals 1.1264. What might we do to obtain a more statistically significant result?

8.51 Is it appropriate to employ a t test when a large (≥ 30) sample has been taken? How do the results of the t test and z test compare when the sample size is large?

METHODS AND APPLICATIONS

8.52 For each of the following hypothesis tests find the rejection point condition needed. In each case assume that the population is normally distributed.

a Testing H_0: $\mu \leq 1$ versus H_a: $\mu > 1$ with $n = 15$ and $\alpha = .05$.
b Testing H_0: $\mu \geq 1$ versus H_a: $\mu < 1$ with $n = 7$ and $\alpha = .10$.
c Testing H_0: $\mu = 1$ versus H_a: $\mu \neq 1$ with $n = 24$ and $\alpha = .01$.

8.53 Suppose that a random sample of 16 measurements from a normally distributed population gives a sample mean of $\bar{x} = 13.5$ and a sample standard deviation of $s = 6$. Use rejection points to test H_0: $\mu \leq 10$ versus H_a: $\mu > 10$ using levels of significance $\alpha = .10$, $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$. What do you conclude at each value of α?

8.54 Suppose that a random sample of nine measurements from a normally distributed population gives a sample mean of $\bar{x} = 2.57$ and a sample standard deviation of $s = .3$. Use rejection points to test H_0: $\mu = 3$ versus H_a: $\mu \neq 3$ using levels of significance $\alpha = .10$, $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$. What do you conclude at each value of α?

8.55 The *bad debt ratio* for a financial institution is defined to be the dollar value of loans defaulted divided by the total dollar value of all loans made. Suppose that a random sample of seven Ohio banks is selected and that the bad debt ratios (written as percentages) for these banks are 7%, 4%, 6%, 7%, 5%, 4%, and 9%. BadDebt

a Banking officials claim that the mean bad debt ratio for all Midwestern banks is 3.5 percent and that the mean bad debt ratio for Ohio banks is higher. Set up the null and alternative hypotheses needed to attempt to provide evidence supporting the claim that the mean bad debt ratio for Ohio banks exceeds 3.5 percent.

b Assuming that bad debt ratios for Ohio banks are approximately normally distributed, use rejection points and the given sample information to test the hypotheses you set up in part *a* by setting α equal to .10, .05, .01, and .001. How much evidence is there that the mean bad debt ratio for Ohio banks exceeds 3.5 percent? What does this say about the banking official's claim?

c Are you qualified to decide whether we have a practically important result? Who would be? How might practical importance be defined in this situation?

8.56 Consider Exercise 8.55. Below we give the MINITAB output of the test statistic and p-value for testing H_0: $\mu = 3.5$ versus H_a: $\mu > 3.5$. BadDebt

```
Test of mu = 3.500 vs mu > 3.500

Variable    N    Mean   StDev   SE Mean      T        P
d-ratio     7   6.000   1.826     0.690   3.62   0.0055
```

a Use the p-value to test H_0 versus H_a by setting α equal to .10, .05, .01, and .001. What do you conclude at each value of α?

b How much evidence is there that the mean bad debt ratio for Ohio banks exceeds 3.5 percent?

8.57 In the book *Business Research Methods,* Donald R. Cooper and C. William Emory (1995) discuss using hypothesis testing to study receivables outstanding. To quote Cooper and Emory:

> . . . the controller of a large retail chain may be concerned about a possible slowdown in payments by the company's customers. She measures the rate of payment in terms of the average number of days receivables outstanding. Generally, the company has maintained an average of about 50 days with a standard deviation of 10 days. Since it would be too expensive to analyze all of a company's receivables frequently, we normally resort to sampling.

a Set up the null and alternative hypotheses needed to attempt to show that there has been a slowdown in payments by the company's customers (there has been a slowdown if the average days outstanding exceeds 50).

b Assume approximate normality and suppose that a random sample of 25 accounts gives an average days outstanding of $\bar{x} = 54$ with a standard deviation of $s = 8$. Use rejection points to test the hypotheses you set up in part *a* at levels of significance $\alpha = .10$, $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$. How much evidence is there of a slowdown in payments?

c Are you qualified to decide whether this result has practical importance? Who would be?

8.58 Again consider the length of time it takes Dutch companies to complete the five stages in the adoption of total quality control (TQC) as originally discussed in Exercise 7.19 (page 270).

a Letting μ be the mean duration of the implementation stage for Dutch companies, set up the null and alternative hypotheses needed to attempt to provide evidence that μ exceeds one year.

b Suppose that a random sample of five Dutch firms that have adopted TQC is selected. Each firm is asked to report how long it took to complete the implementation stage. The firms report the following durations (in years) for this stage: 2.5, 1.5, 1.25, 3.5, and 1.25. Assuming that durations are approximately normally distributed, the MINITAB output of the test statistic and p-value for testing $H_0: \mu = 1$ versus H_a: $\mu > 1$ is as follows: TQC

```
Test of mu = 1.000 vs mu > 1.000

Variable  N    Mean   StDev  SE Mean     T   P-Value
YEARS     5   2.000   0.984    0.440  2.27     0.043
```

Use the sample data to verify that the values of $\bar{x}$, s, and t given on the output are correct.

c Using the test statistic and rejection points, test H_0 versus H_a by setting α equal to .10, .05, .01, and .001. How much evidence is there that the mean length of time it takes Dutch companies to complete the implementation stage exceeds one year?

d Using the p-value on the MINITAB output, test H_0 versus H_a by setting α equal to .10, .05, .01, and .001.

e Do you think that this sample result has practical importance? Explain.

f Suppose that a sample of 25 firms gave the same sample mean and standard deviation as those obtained from the sample in part *b*. How much evidence would there be that the mean length of time exceeds one year?

8.59 Consider the chemical yield situation discussed in Exercise 7.22 (page 271). Recall that the chemical company wishes to determine whether a new catalyst, catalyst XA-100, changes the mean hourly yield of the chemical process. When five trial runs are made using the new catalyst, the following yields (in pounds per hour) are recorded: 801, 814, 784, 836, and 820. ChemYield

a Let μ be the mean of all possible yields using the new catalyst. Assuming that chemical yields are approximately normally distributed, the MegaStat output of the test statistic and p-value, and the Excel output of the p-value, for testing H_0: $\mu = 750$ versus H_a: $\mu \neq 750$ are as follows:

Hypothesis Test: Mean vs. Hypothesized Value

750.000 hypothesized value	19.647 std. dev.	5 n	6.94 t
811.000 mean Hourly Yield	8.786 std. error	4 df	.0023 p-value (two-tailed)

t-statistic
6.942585
p-value
0.002261

(Here we had Excel calculate twice the area under the t distribution curve having 4 degrees of freedom to the right of 6.942585.) Use the sample data to verify that the values of $\bar{x}$, s, and t given on the output are correct.

b Use the test statistic and rejection points to test H_0 versus H_a by setting α equal to .10, .05, .01, and .001.

c Use the p-value to test H_0 versus H_a by setting α equal to .10, .05, .01, and .001.

d How much evidence is there that the new catalyst changes the mean hourly yield?

e Find the p-value for testing H_0: $\mu = 750$ versus H_a: $\mu > 750$.

8.60 As discussed in Exercise 7.21 (page 271), the October 7, 1991, issue of *Fortune* magazine reported on the rapid rise of fees and expenses charged by mutual funds.

a Suppose that 10 years ago the average annual expense for stock funds was 1.19 percent. Let μ be the current mean annual expense for all stock funds, and assume that stock fund annual expenses are approximately normally distributed. If a random sample of 12 stock funds gives a sample mean annual expense of $\bar{x} = 1.63\%$ with a standard deviation of $s = .31\%$, use rejection points and this sample information to test H_0: $\mu \leq 1.19\%$ versus H_a: $\mu > 1.19\%$ by setting α equal to .10, .05, .01, and .001. How much evidence is there that the current mean annual expense for stock funds exceeds the average of 10 years ago?

b Now let μ be the current mean annual expense for all municipal bond funds, and assume that municipal bond fund annual expenses are approximately normally distributed. If a random sample of 12 municipal bond funds gives a sample mean annual expense of $\bar{x} = 0.89\%$ with a standard deviation of $s = .23\%$, use rejection points and this sample information to test H_0: $\mu \geq 1\%$ versus H_a: $\mu < 1\%$ by setting α equal to .10, .05, .01, and .001. How much evidence is there that the current mean annual expense for all municipal bond funds is less than 1 percent?

c Do you think that the result in part *a* has practical importance? Explain your opinion. How might we define practical importance in this situation?

8.61 In Exercise 8.30 (page 316) we used the normal distribution to calculate the p-value for testing H_0: $\mu = 42$ versus H_a: $\mu < 42$ in the customer satisfaction situation. If we use Excel to calculate this p-value by using the t distribution, we find that the p-value equals .002478. Describe how Excel has calculated this p-value, and compare this p-value with the p-value calculated in Exercise 8.30. CustSat

8.62 In Exercise 8.61, why might the p-value calculated using the t distribution be considered more appropriate? CustSat

8.5 Tests about a Population Proportion

In this section we study a large sample hypothesis test about a population proportion (that is, about the fraction of population units that possess some qualitative characteristic). We begin with an example.

EXAMPLE 8.5 The Cheese Spread Case

Recall that the soft cheese spread manufacturer has decided that replacing the current spout with the new spout is profitable only if p, the true proportion of all current purchasers who would stop buying the cheese spread if the new spout were used, is less than .10. The manufacturer feels that it is unwise to change the spout unless it has very strong evidence that p is less than .10. Therefore, the spout will be changed if and only if the null hypothesis H_0: $p = .10$ can be rejected in favor of the alternative hypothesis H_a: $p < .10$ at the .01 level of significance.

In order to see how to test this kind of hypothesis, remember that when n is large, the sampling distribution of

$$\frac{\hat{p} - p}{\sqrt{\dfrac{p(1 - p)}{n}}}$$

is approximately a standard normal distribution. Let p_0 denote a specified value between 0 and 1 (its exact value will depend on the problem), and consider testing the null hypothesis H_0: $p = p_0$. We then have the following result:

A Large Sample Test about a Population Proportion: Testing H_0: $p = p_0$

Define the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

If the sample size n is large, we can reject H_0: $p = p_0$ in favor of a particular alternative hypothesis at level of significance α if and only if the appropriate rejection point condition holds, or, equivalently, the corresponding p-value is less than α.

Alternative Hypothesis	Rejection Point Condition: Reject H_0 if	p-Value
H_a: $p > p_0$	$z > z_\alpha$	The area under the standard normal curve to the right of z
H_a: $p < p_0$	$z < -z_\alpha$	The area under the standard normal curve to the left of z
H_a: $p \neq p_0$	$\lvert z \rvert > z_{\alpha/2}$— that is, $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	Twice the area under the standard normal curve to the right of $\lvert z \rvert$

Here n should be considered large if both np_0 and $n(1 - p_0)$ are at least 5.[3]

EXAMPLE 8.6 The Cheese Spread Case

Recall that p is the proportion of all current purchasers who would stop buying the cheese spread if the new spout were used. In order to provide evidence that the new spout is profitable, we wish to demonstrate that p is less than .10. We can attempt to do this by testing H_0: $p = .10$ versus H_a: $p < .10$. Suppose that 1,000 soft cheese spread purchasers are randomly selected to test the new spout. We find that 63 of these purchasers say they would stop buying the spread if the new spout were used. Therefore, the sample proportion of purchasers who would stop buying is $\hat{p} = 63/1{,}000 = .063$. Since $np_0 = 1{,}000(.10) = 100$ and $n(1 - p_0) = 1{,}000(1 - .10) = 900$ are both at least 5, we can test H_0 versus H_a by comparing the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \frac{.063 - .10}{\sqrt{\dfrac{.10(1 - .10)}{1{,}000}}} = -3.90$$

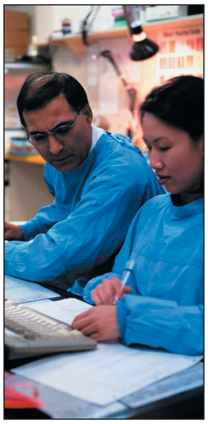
with the rejection point $-z_{.01} = -2.33$. Because $z = -3.90$ is less than $-z_{.01} = -2.33$, we can reject H_0 in favor of H_a by setting $\alpha = .01$. Therefore, we conclude (at level of significance .01) that the proportion of current purchasers who would stop buying the cheese spread if the new spout were used is less than .10. It follows that the company will use the new spout.

Although the manufacturer has made its decision by setting α equal to a single, prechosen value (.01), it would probably also wish to know the weight of evidence against H_0 and in favor of H_a. The p-value is the area under the standard normal curve to the left of $z = -3.90$. Since Table A.3 (page 814) tells us that the area to the left of -3.09 is .001, and since -3.90 is less than -3.09, the p-value is less than .001. Therefore, we have extremely strong evidence that H_a: $p < .10$ is true. That is, we have extremely strong evidence that fewer than 10 percent of current purchasers would stop buying the cheese spread if the new spout were used. Furthermore,

[3]Some statisticians suggest using the more conservative rule that both np_0 and $n(1 - p_0)$ must be at least 10.

the point estimate $\hat{p} = .063$ says we estimate that 6.3 percent of all current customers would stop buying the cheese spread if the new spout were used.

EXAMPLE 8.7



Recent medical research has sought to develop drugs that lessen the severity and duration of viral infections. Virol, a relatively new drug, has been shown to provide relief for 70 percent of all patients suffering from viral upper respiratory infections. A major drug company is developing a competing drug called Phantol. The drug company wishes to investigate whether Phantol is more effective than Virol. To do this, the company will randomly sample 300 patients having viral upper respiratory infections to attempt to reject the null hypothesis H_0: $p = .70$ in favor of the alternative hypothesis H_a: $p > .70$. Here p is the true proportion of all patients whose symptoms are relieved by Phantol. Suppose that Phantol provides relief for 231 of the 300 randomly selected patients. Then, the sample proportion of patients who obtain relief is $\hat{p} = 231/300 = .77$. Because $np_0 = 300(.70) = 210$ and $n(1 - p_0) = 300(1 - .70) = 90$ are both at least 5, we can test H_0 versus H_a by computing the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \frac{.77 - .70}{\sqrt{\dfrac{(.70)(1 - .70)}{300}}} = 2.65$$

Because $z = 2.65$ is greater than $z_{.05} = 1.645$, we can reject H_0 at the .05 level of significance. The p-value for the test is the area under the standard normal curve to the right of $z = 2.65$. This p-value is $(.5 - .4960) = .004$ (see Table A.3, page xxx), and it provides very strong evidence against H_0: $p = .70$ and in favor of H_a: $p > .70$. That is, we have very strong evidence that Phantol will provide relief for more than 70 percent of all patients suffering from viral upper respiratory infections. More specifically, the point estimate $\hat{p} = .77$ of p says that we estimate that Phantol will provide relief for 77 percent of all such patients. Comparing this estimate to the 70 percent of patients whose symptoms are relieved by Virol, we conclude that Phantol is somewhat more effective.

EXAMPLE 8.8 The Electronic Article Surveillance Case

Suppose that a company selling electronic article surveillance devices claims that the proportion, p, of all consumers who would say they would never shop in a store again if the store subjected them to a false alarm is no more than .05. A store considering installing such a device is concerned that p is greater than .05 and wishes to test H_0: $p = .05$ versus H_a: $p > .05$. Recall that 40 out of 250 consumers in a systematic sample said they would never shop in a store again if the store subjected them to a false alarm. Therefore, the sample proportion of consumers is $\hat{p} = 40/250 = .16$. Since $np_0 = 250(.05) = 12.5$ and $n(1 - p_0) = 250(1 - .05) = 237.5$ are both at least 5, the p-value for testing H_0 versus H_a is the area under the standard normal curve to the right of

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \frac{.16 - .05}{\sqrt{\dfrac{(.05)(.95)}{250}}} = 7.9802$$

The normal table tells us that the area under the standard normal curve to the right of 3.09 is .001. Therefore, the p-value is less than .001 and provides extremely strong evidence against H_0: $p = .05$ and in favor of H_a: $p > .05$. That is, we have extremely strong evidence that the proportion of all consumers who say they would never shop in a store again if the store subjected them to a false alarm is greater than .05. Furthermore, the point estimate $\hat{p} = .16$ says we estimate that the percentage of such consumers is 11 percent more than the 5 percent maximum claimed by the company selling the electronic article surveillance devices. A 95 percent confidence interval for p is

$$\left[\hat{p} \pm z_{.025}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n - 1}}\right] = \left[.16 \pm 1.96\sqrt{\frac{(.16)(.84)}{250 - 1}}\right]$$
$$= [.1145, .2055]$$

FIGURE 8.12 **The MegaStat Output for Testing H_0: $p = .16$ versus H_a: $p > .16$**

Hypothesis test for proportion vs hypothesized value

Observed	*Hypothesized*				
0.16	0.05 p (as decimal)	0.0138 std. error		0.1146 confidence interval 95.% lower	
40/250	13/250 p (as fraction)	7.98 z		0.2054 confidence interval 95.% upper	
40.	12.5 X	7.77E-16 p-value		0.0454	half-width
250	250 n	(one-tailed upper)			

This interval says we are 95 percent confident that the percentage of consumers who would say they would never shop in a store again if the store subjected them to a false alarm is between 6.45 percent and 15.55 percent more than the 5 percent maximum claimed by the company selling the electronic article surveillance devices. The rather large increases over the claimed 5 percent maximum implied by the point estimate and the confidence interval would mean substantially more lost customers and thus are practically important. Figure 8.12 gives the MegaStat output for testing H_0: $p = .16$ versus H_a: $p > .16$. Note that this output includes a 95 percent confidence interval for p.

Exercises for Section 8.5

CONCEPTS

8.67, 8.69, 8.76

8.63 If we test a hypothesis to provide evidence supporting the claim that a majority of voters prefer a political candidate, discuss the difference between p and $\hat{p}$.

8.64 If we test a hypothesis to provide evidence supporting the claim that more than 30 percent of all consumers prefer a particular brand of beer, discuss the difference between p and $\hat{p}$.

8.65 If we test a hypothesis to provide evidence supporting the claim that fewer than 5 percent of the units produced by a process are defective, discuss the difference between p and $\hat{p}$.

8.66 What condition must be satisfied in order to appropriately use the methods of this section?

METHODS AND APPLICATIONS

8.67 For each of the following sample sizes and hypothesized values of the population proportion p, determine whether the sample size is large enough to use the large sample test about p given in this section:

a $n = 400$ and $p_0 = .5$.
b $n = 100$ and $p_0 = .01$.
c $n = 10{,}000$ and $p_0 = .01$.
d $n = 100$ and $p_0 = .2$.
e $n = 256$ and $p_0 = .7$.
f $n = 200$ and $p_0 = .98$.
g $n = 1{,}000$ and $p_0 = .98$.
h $n = 25$ and $p_0 = .4$.

8.68 Suppose we wish to test H_0: $p \leq .8$ versus H_a: $p > .8$ and that a random sample of $n = 400$ gives a sample proportion $\hat{p} = .86$.

a Test H_0 versus H_a at the .05 level of significance by using a rejection point. What do you conclude?
b Find the p-value for this test.
c Use the p-value to test H_0 versus H_a by setting α equal to .10, .05, .01, and .001. What do you conclude at each value of α?

8.69 Suppose we test H_0: $p = .3$ versus H_a: $p \neq .3$ and that a random sample of $n = 100$ gives a sample proportion $\hat{p} = .20$.

a Test H_0 versus H_a at the .01 level of significance by using a rejection point. What do you conclude?
b Find the p-value for this test.
c Use the p-value to test H_0 versus H_a by setting α equal to .10, .05, .01, and .001. What do you conclude at each value of α?

8.70 Suppose we are testing H_0: $p \leq .5$ versus H_a: $p > .5$, where p is the proportion of all beer drinkers who have tried at least one brand of "ice beer." If a random sample of 500 beer drinkers has been taken and if $\hat{p}$ equals .57, how many beer drinkers in the sample have tried at least one brand of "ice beer"?

8.71 **THE MARKETING ETHICS CASE: CONFLICT OF INTEREST**

Recall that a conflict of interest scenario was presented to a sample of 205 marketing researchers and that 111 of these researchers disapproved of the actions taken.

a Let p be the proportion of all marketing researchers who disapprove of the actions taken in the conflict of interest scenario. Set up the null and alternative hypotheses needed to attempt to provide evidence supporting the claim that a majority (more than 50 percent) of all marketing researchers disapprove of the actions taken.

b Assuming that the sample of 205 marketing researchers has been randomly selected, use rejection points and the previously given sample information to test the hypotheses you set up in part *a* at the .10, .05, .01, and .001 levels of significance. How much evidence is there that a majority of all marketing researchers disapprove of the actions taken?

c Suppose a random sample of 1,000 marketing researchers reveals that 540 of the researchers disapprove of the actions taken in the conflict of interest scenario. Use rejection points to determine how much evidence there is that a majority of all marketing researchers disapprove of the actions taken.

d Note that in parts *b* and *c* the sample proportion $\hat{p}$ is (essentially) the same. Explain why the results of the hypothesis tests in parts *b* and *c* differ.

8.72 Last year, television station WXYZ's share of the 11 P.M. news audience was approximately equal to, but no greater than, 25 percent. The station's management believes that the current audience share is higher than last year's 25 percent share. In an attempt to substantiate this belief, the station surveyed a random sample of 400 11 P.M. news viewers and found that 146 watched WXYZ.

a Let p be the current proportion of all 11 P.M. news viewers who watch WXYZ. Set up the null and alternative hypotheses needed to attempt to provide evidence supporting the claim that the current audience share for WXYZ is higher than last year's 25 percent share.

b Use rejection points and the following MINITAB output to test the hypotheses you set up in part *a* at the .10, .05, .01, and .001 levels of significance. How much evidence is there that the current audience share is higher than last year's 25 percent share?

```
Test of p=0.25 vs p>0.25
Sample      X     N     Sample p      Z-Value     P-Value
1         146   400     0.365000       5.31        0.000
```

c Calculate the p-value for the hypothesis test in part *b*. Use the p-value to carry out the test by setting α equal to .10, .05, .01, and .001. Interpret your results.

d Do you think that the result of the station's survey has practical importance? Why or why not?

8.73 Suppose in January 1999 it was reported that 22 percent of Americans rated their own economic situation to be "very good." In a Gannett News Service poll[4] of 1,003 Americans taken October 15–20, 1999, and reported in the October 31, 1999, issue of the *Cincinnati Enquirer,* 18 percent of those polled rated their own economic situation to be "very good."

a Set up and carry out a hypothesis test that attempts to provide evidence supporting the claim that the proportion of Americans rating their own economic situation as very good in October 1999 is lower than the proportion of 22 percent in January 1999. Use $\alpha = .05$. What do you conclude?

b Use the sample results to determine how much evidence there is that the proportion of Americans rating their own economic situation as very good in October 1999 is lower than in January 1999.

8.74 In the book *Essentials of Marketing Research,* William R. Dillon, Thomas J. Madden, and Neil H. Firtle discuss a marketing research proposal to study day-after recall for a brand of mouthwash. To quote the authors:

> The ad agency has developed a TV ad for the introduction of the mouthwash. The objective of the ad is to create awareness of the brand. The objective of this research is to evaluate the awareness generated by the ad measured by aided- and unaided-recall scores.
>
> A minimum of 200 respondents who claim to have watched the TV show in which the ad was aired the night before will be contacted by telephone in 20 cities.
>
> The study will provide information on the incidence of unaided and aided recall.

Suppose a random sample of 200 respondents shows that 46 of the people interviewed were able to recall the commercial without any prompting (unaided recall).

a In order for the ad to be considered successful, the percentage of unaided recall must be above the category norm for a TV commercial for the product class. If this norm is 18 percent, set up the null and alternative hypotheses needed to attempt to provide evidence that the ad is successful.

[4]Source: C. Raasch, "At 2000, Americans Guardedly Optimistic," Gannett News Service, *Cincinnati Enquirer,* October 31, 1999, p. A1.

b Use the previously given sample information to compute the *p*-value for the hypothesis test you set up in part *a.* Use the *p*-value to carry out the test by setting α equal to .10, .05, .01, and .001. How much evidence is there that the TV commercial is successful?

c Do you think the result of the ad agency's survey has practical importance? Explain your opinion.

8.75 In Exercise 2.11 (page 54) we briefly described a series of international quality standards called ISO 9000. In the results of a Quality Systems Update/Deloitte & Touche survey of ISO 9000–registered companies published by CEEM Information Systems, 496 of 620 companies surveyed reported that ISO 9000 registration status influences their selection of suppliers.[5]

a Letting *p* be the proportion of all ISO 9000–registered companies that say ISO 9000 registration status influences their selection of suppliers, set up the null and alternative hypotheses needed to attempt to provide evidence supporting the claim that *p* exceeds .75.

b Use rejection points and the previously given sample information to test the hypotheses you set up in part *a* by setting α equal to .10, .05, .01, and .001. How much evidence is there that more than 75 percent of all ISO 9000–registered companies say that ISO 9000 registration status influences their selection of suppliers?

8.76 The manufacturer of the ColorSmart-5000 television set claims that 95 percent of its sets last at least five years without needing a single repair. In order to test this claim, a consumer group randomly selects 400 consumers who have owned a ColorSmart-5000 television set for five years. Of these 400 consumers, 316 say that their ColorSmart-5000 television sets did not need repair, while 84 say that their ColorSmart-5000 television sets did need at least one repair.

a Letting *p* be the proportion of ColorSmart-5000 television sets that last five years without a single repair, set up the null and alternative hypotheses that the consumer group should use to attempt to show that the manufacturer's claim is false.

b Use rejection points and the previously given sample information to test the hypotheses you set up in part *a* by setting α equal to .10, .05, .01, and .001. How much evidence is there that the manufacturer's claim is false?

c Do you think the results of the consumer group's survey have practical importance? Explain your opinion.

8.77 In its April 22, 1991, issue, *Fortune* magazine reported on the results of a survey of 203 CEOs of *Fortune* 500 and Service 500 companies. The survey investigated the CEO's opinions regarding corporate debt and competitiveness. In response to one question, 82 percent of the CEOs said that their companies' debt load did not inhibit their competitiveness in any way.[6]

a Suppose that a new survey of 200 randomly selected CEOs is taken and that 150 of these CEOs say that their companies' debt load does not inhibit their competitiveness in any way. Letting *p* be the proportion of all CEOs who feel that their companies' debt load does not inhibit competitiveness, test H_0: $p = .82$, which says that currently held opinion is the same as was found in the 1991 survey, versus H_a: $p \neq .82$, which says that current opinion differs from the 1991 survey result. Carry out this test for levels of significance .10, .05, .01, and .001 by using rejection points. Hint: $z_{.0005} = 3.29$.

b Find the *p*-value for testing H_0: $p = .82$ versus H_a: $p \neq .82$. Use the *p*-value to carry out this test by setting α equal to .10, .05, .025, .01, and .001.

c How much evidence is there that current opinion differs from the 1991 survey result (that is, that *p* differs from .82)? If opinion has changed, describe how it has changed. Do you think the change has any practical importance? Explain your opinion.

*8.6 Type II Error Probabilities and Sample Size Determination

As we have seen, we usually take action (for example, advertise a claim) on the basis of having rejected the null hypothesis. In this case, we know the chances that the action has been taken erroneously because we have prespecified α, the probability of rejecting a true null hypothesis. However, sometimes we must act (for example, use a day's production of camshafts to make V6 engines) on the basis of *not* rejecting the null hypothesis. If we must do this, it is best to know the probability of not rejecting a false null hypothesis (a Type II error). If this probability is not small enough, we may change the hypothesis testing procedure. In order to discuss this further, we must first see how to compute the probability of a Type II error.

CHAPTERS 9 AND 11

[5] Source: *Is ISO 9000 for You?* (Fairfax, VA: CEEM Information Services).

[6] Source: "What, Me? Too Leveraged?" *Fortune* (April 22, 1991), pp. 135–36.

As an example, the Federal Trade Commission (FTC) often tests claims that companies make about their products. Suppose coffee is being sold in cans that are labeled as containing three pounds, and also suppose that the FTC wishes to determine if the mean amount of coffee μ in all such cans is at least three pounds. To do this, the FTC tests H_0: $\mu \geq 3$ (or $\mu = 3$) versus H_a: $\mu < 3$ by setting $\alpha = .05$. Suppose that a sample of 35 coffee cans yields $\bar{x} = 2.9973$ and $s = .0147$. Because

$$z = \frac{\bar{x} - 3}{s/\sqrt{n}} = \frac{2.9973 - 3}{.0147/\sqrt{35}} = -1.0866$$

is not less than $-z_{.05} = -1.645$, we cannot reject H_0: $\mu \geq 3$ by setting $\alpha = .05$. Since we cannot reject H_0, we cannot have committed a Type I error, which is the error of rejecting a true H_0. However, we might have committed a Type II error, which is the error of not rejecting a false H_0. Therefore, before we make a final conclusion about μ, we should calculate the probability of a Type II error.

A Type II error is not rejecting H_0: $\mu \geq 3$ when H_0 is false. Because any value of μ that is less than 3 makes H_0 false, there is a different Type II error (and, therefore, a different Type II error probability) associated with each value of μ that is less than 3. In order to demonstrate how to calculate these probabilities, we will calculate the probability of not rejecting H_0: $\mu \geq 3$ when in fact μ equals 2.995. This is the probability of failing to detect an average underfill of .005 pounds. For a fixed sample size (for example, $n = 35$ coffee can fills), the value of β, the probability of a Type II error, depends upon how we set α, the probability of a Type I error. Since we have set $\alpha = .05$, we reject H_0 if

$$\frac{\bar{x} - 3}{\sigma/\sqrt{n}} < -z_{.05}$$

or, equivalently, if

$$\bar{x} < 3 - z_{.05}\frac{\sigma}{\sqrt{n}}$$

Estimating σ by s (because n is large), we reject H_0 if

$$\bar{x} < 3 - z_{.05}\frac{s}{\sqrt{n}} = 3 - 1.645\frac{.0147}{\sqrt{35}} = 2.9959126$$

Therefore, we do not reject H_0 if $\bar{x} \geq 2.9959126$. It follows that β, the probability of not rejecting H_0: $\mu \geq 3$ when μ equals 2.995, is

$$\begin{aligned}\beta &= P(\bar{x} \geq 2.9959126 \text{ when } \mu = 2.995)\\ &= P\left(\frac{\bar{x} - \mu}{s/\sqrt{n}} \geq \frac{2.9959126 - 2.995}{.0147/\sqrt{35}}\right)\\ &= P(z \geq .37) = .5 - .1443 = .3557\end{aligned}$$

This calculation is illustrated in Figure 8.13. Similarly, it follows that β, the probability of not rejecting H_0: $\mu \geq 3$ when μ equals 2.99, is

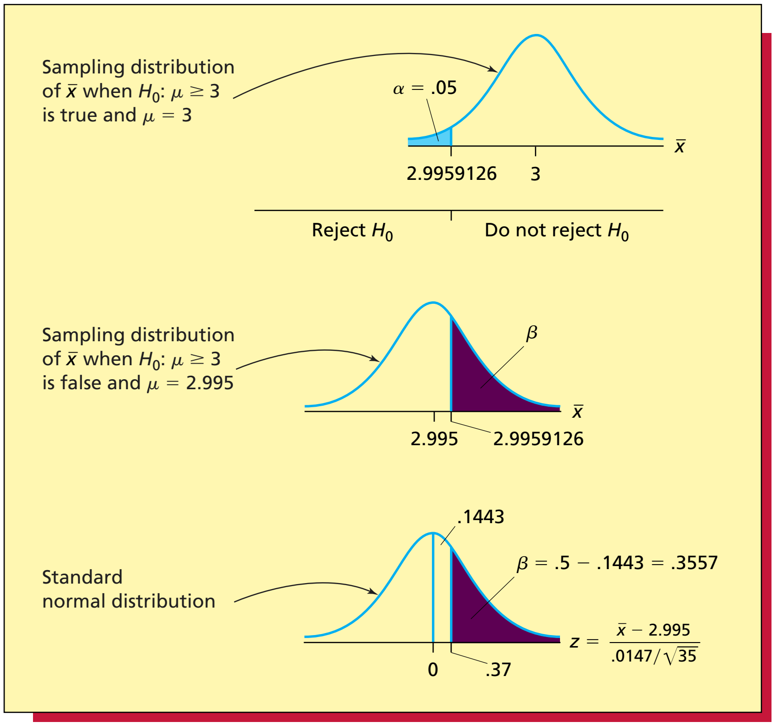
$$\begin{aligned}\beta &= P(\bar{x} \geq 2.9959126 \text{ when } \mu = 2.99)\\ &= P\left(\frac{\bar{x} - \mu}{s/\sqrt{n}} \geq \frac{2.9959126 - 2.99}{.0147/\sqrt{35}}\right)\\ &= P(z \geq 2.38) = .5 - .4913 = .0087\end{aligned}$$

It also follows that β, the probability of not rejecting H_0: $\mu \geq 3$ when μ equals 2.985, is

$$\begin{aligned}\beta &= P(\bar{x} \geq 2.9959126 \text{ when } \mu = 2.985)\\ &= P\left(\frac{\bar{x} - \mu}{s/\sqrt{n}} \geq \frac{2.9959126 - 2.985}{.0147/\sqrt{35}}\right)\\ &= P(z \geq 4.39)\end{aligned}$$

This probability is less than .001 (because z is greater than 3.09).

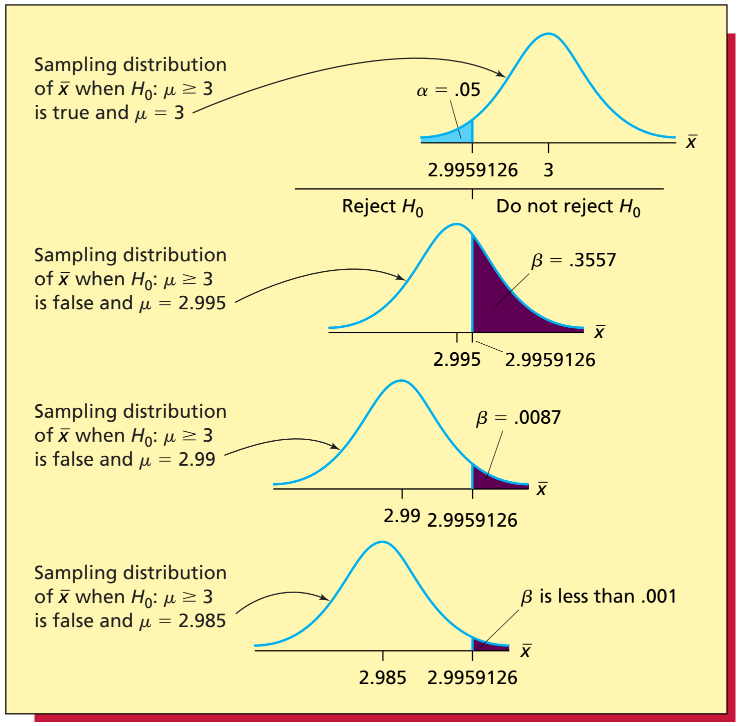
FIGURE 8.13 Calculating β When μ Equals 2.995



In Figure 8.14 we illustrate the values of β that we have calculated. Notice that the closer an alternative value of μ is to 3 (the value specified by H_0: $\mu = 3$), the larger is the associated value of β. Although alternative values of μ that are closer to 3 have larger associated probabilities of Type II errors, these values of μ have associated Type II errors with less serious consequences. For example, we are more likely to not reject H_0: $\mu = 3$ when $\mu = 2.995$ ($\beta = .3557$) than we are to not reject H_0: $\mu = 3$ when $\mu = 2.99$ ($\beta = .0087$). However, not rejecting H_0: $\mu = 3$ when $\mu = 2.995$, which means that we are failing to detect an average underfill of .005 pounds, is less serious than not rejecting H_0: $\mu = 3$ when $\mu = 2.99$, which means that we are failing to detect a larger average underfill of .01 pounds. In order to decide whether a particular hypothesis test adequately controls the probability of a Type II error, we must determine which Type II errors are serious, and then we must decide whether the probabilities of these errors are small enough. For example, suppose that the FTC and the coffee producer agree that failing to reject H_0: $\mu = 3$ when μ equals 2.99 is a serious error, but that failing to reject H_0: $\mu = 3$ when μ equals 2.995 is not a particularly serious error. Then, since the probability of not rejecting H_0: $\mu = 3$ when μ equals 2.99, which is .0087, is quite small, we might decide that the hypothesis test adequately controls the probability of a Type II error. To understand the implication of this, recall that the sample of 35 coffee cans, which has $\bar{x} = 2.9973$ and $s = .0147$, does not provide enough evidence to reject H_0: $\mu \geq 3$ by setting $\alpha = .05$. We have just shown that the probability that we have failed to detect a serious underfill is quite small (.0087), so the FTC might decide that no action should be taken against the coffee producer. Of course, this decision should also be based on the variability of the fills of the individual cans. Because $\bar{x} = 2.9973$ and $s = .0147$, we estimate that 99.73 percent of all individual coffee can fills are contained in the interval $[\bar{x} \pm 3s] = [2.9973 \pm 3(.0147)] = [2.9532, 3.0414]$. If the FTC believes it is reasonable to accept fills as low as (but no lower than) 2.9532 pounds, this evidence also suggests that no action against the coffee producer is needed.

Suppose, instead, that the FTC and the coffee producer had agreed that failing to reject H_0: $\mu \geq 3$ when μ equals 2.995 is a serious mistake. The probability of this Type II error, which

FIGURE 8.14 How β Changes as the Alternative Value of μ Changes



is .3557, is large. Therefore, we might conclude that the hypothesis test is not adequately controlling the probability of a serious Type II error. In this case, we have two possible courses of action. First, we have previously said that, for a fixed sample size, the lower we set α, the higher is β, and the higher we set α, the lower is β. Therefore, if we keep the sample size fixed at $n = 35$ coffee cans, we can reduce β by increasing α. To demonstrate this, suppose we increase α to .10. In this case we reject H_0 if

$$\frac{\bar{x} - 3}{\sigma/\sqrt{n}} < -z_{.10}$$

or, equivalently, if

$$\bar{x} < 3 - z_{.10}\frac{\sigma}{\sqrt{n}}$$

Estimating σ by $s = .0147$, we reject H_0 if

$$\bar{x} < 3 - z_{.10}\frac{s}{\sqrt{n}} = 3 - 1.282\frac{.0147}{\sqrt{35}} = 2.9968145$$

Therefore, we do not reject H_0 if $\bar{x} \geq 2.9968145$. It follows that β, the probability of not rejecting H_0: $\mu \geq 3$ when μ equals 2.995, is

$$\begin{aligned}\beta &= P(\bar{x} \geq 2.9968145 \text{ when } \mu = 2.995) \\ &= P\left(\frac{\bar{x} - \mu}{s/\sqrt{n}} \geq \frac{2.9968145 - 2.995}{.0147/\sqrt{35}}\right) \\ &= P(z \geq .73) = .5 - .2673 = .2327\end{aligned}$$

We thus see that increasing α from .05 to .10 reduces β from .3557 to .2327. However, β is still too large, and, besides, we might not be comfortable making α larger than .05. Therefore, if we wish to decrease β and maintain α at .05, we must increase the sample size. We will soon present a formula we can use to find the sample size needed to make both α and β as small as we wish.

Once we have computed β, we can calculate what we call the *power* of the test.

The **power** of a statistical test is the probability of rejecting the null hypothesis when it is false.

Just as β depends upon the alternative value of μ, so does the power of a test. In general, **the power associated with a particular alternative value of μ equals $1 - \beta$,** where β is the probability of a Type II error associated with the same alternative value of μ. For example, we have seen that, when we set $\alpha = .05$, the probability of not rejecting H_0: $\mu \geq 3$ when μ equals 2.99 is .0087. Therefore, the power of the test associated with the alternative value 2.99 (that is, the probability of rejecting H_0: $\mu \geq 3$ when μ equals 2.99) is $1 - .0087 = .9913$.

Thus far we have demonstrated how to calculate β when testing a *less than* alternative hypothesis. In the following box we present (without proof) a method for calculating the probability of a Type II error when testing a *less than,* a *greater than,* or a *not equal to* alternative hypothesis:

Calculating the Probability of a Type II Error

Assume that the sampled population is normally distributed, or that a large sample will be taken. Consider testing H_0: $\mu = \mu_0$ versus one of H_a: $\mu > \mu_0$, H_a: $\mu < \mu_0$, or H_a: $\mu \neq \mu_0$. Then, if we set the probability of a Type I error equal to α and randomly select a sample of size *n,* the probability, β, of a Type II error corresponding to the alternative value μ_a of μ is (exactly or approximately) equal to the area under the standard normal curve to the left of

$$z^* - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}$$

Here z^* equals z_α if the alternative hypothesis is one-sided ($\mu > \mu_0$ or $\mu < \mu_0$), in which case the method for calculating β is exact. Furthermore, z^* equals $z_{\alpha/2}$ if the alternative hypothesis is two-sided ($\mu \neq \mu_0$), in which case the method for calculating β is approximate.

EXAMPLE 8.9 The Camshaft Case

Reconsider the camshaft situation and testing H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$ by setting $\alpha = .05$. Assume that a random sample of 35 camshafts that has been selected from a day's production yields $\bar{x} = 4.5716$ and $s = .4128$. Because

$$z = \frac{\bar{x} - 4.5}{s/\sqrt{n}} = \frac{4.5716 - 4.5}{.4128/\sqrt{35}} = 1.0261$$

is between $-z_{.025} = -1.96$ and $z_{.025} = 1.96$, we cannot reject H_0: $\mu = 4.5$ by setting $\alpha = .05$. Since we cannot reject H_0, we might have committed a Type II error. Suppose that an automobile assembly plant that is considering using the camshafts decides that failing to reject H_0: $\mu = 4.5$ when μ differs from 4.5 by as much as .3 mm (that is, when μ is 4.2 or 4.8) is a serious Type II error. Because we have set α equal to .05, β for the alternative value $\mu_a = 4.8$ (that is, the probability of not rejecting H_0: $\mu = 4.5$ when μ equals 4.8) is the area under the standard normal curve to the left of

$$\begin{aligned} z^* - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}} &= z_{.025} - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}} \\ &= 1.96 - \frac{|4.5 - 4.8|}{.4128/\sqrt{35}} \\ &= -2.34 \end{aligned}$$

Here we have estimated σ by the sample standard deviation $s = .4128$, and $z^* = z_{\alpha/2} = z_{.05/2} = z_{.025}$ since the alternative hypothesis ($\mu \neq 4.5$) is two-sided. The area under the standard normal curve to the left of -2.34 is $.5 - .4904 = .0096$. Therefore, β for the alternative value $\mu_a = 4.8$ is .0096. Similarly, it can be verified that β for the alternative value $\mu_a = 4.2$ is .0096. It follows, since we cannot reject H_0: $\mu = 4.5$ by setting $\alpha = .05$, and since we have just shown that there is only a small (.0096) probability that we have failed to detect a serious (that is, a .3 mm) deviation of μ from 4.5, that the assembly plant might decide to use the day's production of camshafts. Of course, this decision should also be based on knowing the variability of the hardness depths of the individual camshafts. Because $\bar{x} = 4.5716$ and $s = .4128$, we estimate that 99.73 percent of all individual hardness depths are contained in the interval $[\bar{x} \pm 3s] = [4.5716 \pm 3(.4128)] = [3.3332, 5.81]$. The limits of this interval are within the specification limits of 3.00 mm to 6.00 mm, so the assembly plant has further evidence that it should use the camshafts.

In the following box we present (without proof) a formula that tells us the sample size needed to make both the probability of a Type I error and the probability of a Type II error as small as we wish:

Calculating the Sample Size Needed to Achieve Specified Values of α and β

Assume that the sampled population is normally distributed, or that a large sample will be taken. Consider testing H_0: $\mu = \mu_0$ versus one of H_a: $\mu > \mu_0$, H_a: $\mu < \mu_0$, or H_a: $\mu \neq \mu_0$. Then, in order to make the probability of a Type I error equal to α and the probability of a Type II error corresponding to the alternative value μ_a of μ equal to β, we should take a sample of size

$$n = \frac{(z^* + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2}$$

Here z^* equals z_α if the alternative hypothesis is one-sided ($\mu > \mu_0$ or $\mu < \mu_0$), and z^* equals $z_{\alpha/2}$ if the alternative hypothesis is two-sided ($\mu \neq \mu_0$). Also, z_β is the point on the scale of the standard normal curve that gives a right-hand tail area equal to β.

EXAMPLE 8.10

Again consider the coffee fill example and suppose we wish to test H_0: $\mu \geq 3$ (or $\mu = 3$) versus H_a: $\mu < 3$. If we wish α to be .05 and β for the alternative value $\mu_a = 2.995$ of μ to be .05, we should take a sample of size

$$\begin{aligned} n &= \frac{(z^* + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} \\ &= \frac{(z_{.05} + z_{.05})^2\sigma^2}{(\mu_0 - \mu_a)^2} \\ &= \frac{(1.645 + 1.645)^2(.0147)^2}{(3 - 2.995)^2} \\ &= 93.5592 = 94 \text{ (rounding up)} \end{aligned}$$

Here we have estimated σ by the sample standard deviation $s = .0147$ of the previous sample of $n = 35$ coffee can fills. Moreover, $z^* = z_\alpha = z_{.05} = 1.645$ because the alternative hypothesis ($\mu < 3$) is one-sided, and $z_\beta = z_{.05} = 1.645$.

Although we have set both α and β equal to the same value in the coffee fill situation, it is not necessary for α and β to be equal. As an example, again consider the camshaft situation in which we are testing H_0: $\mu = 4.5$ versus H_a: $\mu \neq 4.5$. Suppose that the assembly plant decides that failing to reject H_0: $\mu = 4.5$ when μ differs from 4.5 by as much as .2 mm (that is, when μ is 4.3 or 4.7) is a serious Type II error. Furthermore, suppose that it is also decided that this Type II error is more serious than a Type I error. Therefore, α will be set equal to .05 and β for the alternative value $\mu_a = 4.7$ (or $\mu_a = 4.3$) of μ will be set equal to .01. It follows that the assembly plant should take a sample of size

$$\begin{aligned} n &= \frac{(z^* + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} \\ &= \frac{(z_{.025} + z_{.01})^2\sigma^2}{(\mu_0 - \mu_a)^2} \\ &= \frac{(1.96 + 2.33)^2(.4128)^2}{(4.5 - 4.7)^2} \\ &= 78.4032 = 79 \text{ (rounding up)} \end{aligned}$$

Here we have estimated σ by the sample standard deviation $s = .4128$ of the previous sample of $n = 35$ camshaft hardness depths. Moreover, $z^* = z_{\alpha/2} = z_{.05/2} = z_{.025} = 1.96$ since the alternative hypothesis ($\mu \neq 4.5$) is two-sided, and $z_\beta = z_{.01} = 2.33$.

To conclude this section, we point out that the methods we have presented for calculating the probability of a Type II error and determining sample size can be extended to other hypothesis tests that utilize the normal distribution. We will not, however, present the extensions in this book.

Exercises for Section 8.6

CONCEPTS

8.81

8.78 We usually take action on the basis of having rejected the null hypothesis. When we do this, we know the chances that the action has been taken erroneously because we have prespecified α, the probability of rejecting a true null hypothesis. Here, it is obviously important to know (prespecify) α, the probability of a Type I error. When is it important to know the probability of a Type II error? Explain why.

8.79 Explain why we are able to compute many different values of β, the probability of a Type II error, for a single hypothesis test.

8.80 Explain what is meant by

a A serious Type II error.

b The power of a statistical test.

In general, do we want the power corresponding to a serious Type II error to be near 0 or near 1? Explain.

METHODS AND APPLICATIONS

8.81 Again consider the Consolidated Power waste water situation discussed in Exercises 8.12 and 8.34 (pages 307 and 317). Remember that the power plant will be shut down and corrective action will be taken on the cooling system if the null hypothesis H_0: $\mu \leq 60$ is rejected in favor of H_a: $\mu > 60$. In this exercise we calculate probabilities of various Type II errors in the context of this situation.

a Recall that Consolidated Power's hypothesis test is based on a sample of $n = 100$ temperature readings and that σ has been estimated by $s = 2$. If the power company sets $\alpha = .025$, calculate the probability of a Type II error for each of the following alternative values of μ: 60.1, 60.2, 60.3, 60.4, 60.5, 60.6, 60.7, 60.8, 60.9, 61.

b If we want the probability of making a Type II error when μ equals 60.5 to be very small, is Consolidated Power's hypothesis test adequate? Explain why or why not. If not, and if we wish to maintain the value of α at .025, what must be done?

c The **power curve** for a statistical test is a plot of the power $= 1 - \beta$ on the vertical axis versus values of μ that make the null hypothesis false on the horizontal axis. Plot the power curve for Consolidated Power's test of H_0: $\mu \leq 60$ versus H_a: $\mu > 60$ by plotting power $= 1 - \beta$ for each of the alternative values of μ in part *a*. What happens to the power of the test as the alternative value of μ moves away from 60?

8.82 Again consider the automobile parts supplier situation discussed in Exercises 8.10 and 8.44 (pages 306 and 322). Remember that a problem-solving team will be assigned to rectify the process producing the cylindrical engine parts if the null hypothesis H_0: $\mu = 3$ is rejected in favor of H_a: $\mu \neq 3$. In this exercise we calculate probabilities of various Type II errors in the context of this situation.

- **a** Suppose that the parts supplier's hypothesis test is based on a sample of $n = 100$ diameters and that σ has been estimated by $s = .023$. If the parts supplier sets $\alpha = .05$, calculate the probability of a Type II error for each of the following alternative values of μ: 2.990, 2.995, 3.005, 3.010.
- **b** If we want the probabilities of making a Type II error when μ equals 2.995 and when μ equals 3.005 to both be very small, is the parts supplier's hypothesis test adequate? Explain why or why not. If not, and if we wish to maintain the value of α at .05, what must be done?
- **c** Plot the power of the test versus the alternative values of μ in part *a*. What happens to the power of the test as the alternative value of μ moves away from 3?

8.83 In the Consolidated Power hypothesis test of H_0: $\mu \leq 60$ versus H_a: $\mu > 60$ (as discussed in Exercise 8.81) find the sample size needed to make the probability of a Type I error equal to .025 and the probability of a Type II error corresponding to the alternative value $\mu_a = 60.5$ equal to .025. Recall here that σ has been estimated by $s = 2$.

8.84 In the automobile parts supplier's hypothesis test of H_0: $\mu = 3$ versus H_a: $\mu \neq 3$ (as discussed in Exercise 8.82) find the sample size needed to make the probability of a Type I error equal to .05 and the probability of a Type II error corresponding to the alternative value $\mu_a = 3.005$ equal to .05. Recall here that σ has been estimated by $s = .023$.

*8.7 The Chi-Square Distribution

CHAPTER 5

Sometimes we can make statistical inferences by using the **chi-square distribution.** The probability curve of the χ^2 (pronounced *chi-square*) distribution is skewed to the right. Moreover, the exact shape of this probability curve depends on a parameter that is called the **number of degrees of freedom** (denoted *df*). Figure 8.15 illustrates chi-square distributions having 1, 5, and 10 degrees of freedom.

In order to use the chi-square distribution, we employ a **chi-square point,** which is denoted χ^2_α. As illustrated in the upper portion of Figure 8.16, χ^2_α is the point on the horizontal axis under the curve of the chi-square distribution that gives a right-hand tail area equal to α. The value of χ^2_α in a particular situation depends on the right-hand tail area α and the number of degrees of freedom (*df*) of the chi-square distribution. Values of χ^2_α are tabulated in a **chi-square table.** Such a table is given in Table A.17 of Appendix A (page 833); a portion of this table is reproduced as Table 8.4. Looking at the chi-square table, the rows correspond to the appropriate number of degrees of freedom (values of which are listed down the right side of the table), while the columns designate the right-hand tail area α. For example, suppose we wish to find the chi-square

FIGURE 8.15 **Chi-Square Distributions with 1, 5, and 10 Degrees of Freedom**

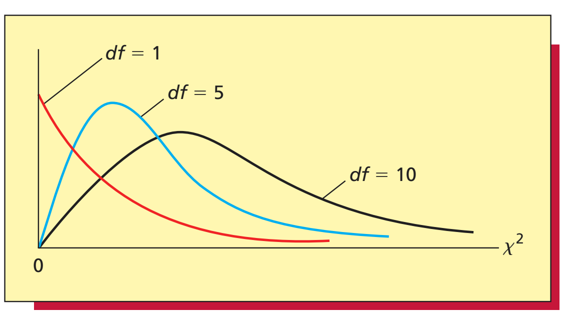


FIGURE 8.16 **Chi-Square Points**

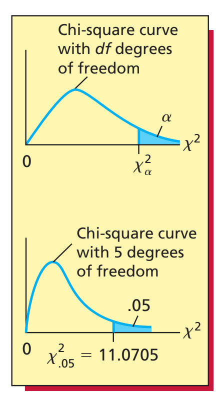


TABLE 8.4 A Portion of the Chi-Square Table

$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$	Degrees of Freedom (df)
2.70554	3.84146	5.02389	6.63490	7.87944	1
4.60517	5.99147	7.37776	9.21034	10.5966	2
6.25139	7.81473	9.34840	11.3449	12.8381	3
7.77944	9.48773	11.1433	13.2767	14.8602	4
9.23635	11.0705	12.8325	15.0863	16.7496	5
10.6446	12.5916	14.4494	16.8119	18.5476	6
12.0170	14.0671	16.0123	18.4753	20.2777	7
13.3616	15.5073	17.5346	20.0902	21.9550	8
14.6837	16.9190	19.0228	21.6660	23.5893	9
15.9871	18.3070	20.4831	23.2093	25.1882	10

point that gives a right-hand tail area of .05 under a chi-square curve having 5 degrees of freedom. To do this, we look in Table 8.4 at the row labeled 5 and the column labeled $\chi^2_{.05}$. We find that this $\chi^2_{.05}$ point is 11.0705 (see the lower portion of Figure 8.16).

*8.8 Statistical Inference for a Population Variance

Consider the camshaft case, and suppose the automobile manufacturer has determined that the variance of the population of all camshaft hardness depths produced by the current hardening process is approximately equal to, but no less than, .47. To reduce this variance, a new electrical coil is designed, and a random sample of $n = 30$ camshaft hardness depths produced by using this new coil has a mean of $\bar{x} = 4.50$ and a variance of $s^2 = .1885$. In order to attempt to show that the variance, σ^2, of the population of all camshaft hardness depths that would be produced by using the new coil is less than .47, we can use the following result:

CHAPTER 9

Statistical Inference for a Population Variance

Suppose that s^2 is the variance of a sample of n measurements randomly selected from a normally distributed population having variance σ^2. The sampling distribution of the statistic $(n - 1)s^2/\sigma^2$ is a chi-square distribution having $n - 1$ degrees of freedom. This implies that

1 **A 100(1 − α) percent confidence interval for σ^2** is

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}\right]$$

Here $\chi^2_{\alpha/2}$ and $\chi^2_{1-(\alpha/2)}$ are the points under the curve of the chi-square distribution having $n - 1$ degrees of freedom that give right-hand tail areas of, respectively, $\alpha/2$ and $1 - (\alpha/2)$.

2 We can test $H_0: \sigma^2 = \sigma_0^2$ by using the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Specifically, if we set the probability of a Type I error equal to α, then we can reject H_0 in favor of

a $H_a: \sigma^2 > \sigma_0^2$ if $\chi^2 > \chi^2_\alpha$

b $H_a: \sigma^2 < \sigma_0^2$ if $\chi^2 < \chi^2_{1-\alpha}$

c $H_a: \sigma^2 \neq \sigma_0^2$ if $\chi^2 > \chi^2_{\alpha/2}$ or $\chi^2 < \chi^2_{1-(\alpha/2)}$

Here χ^2_α, $\chi^2_{1-\alpha}$, $\chi^2_{\alpha/2}$, and $\chi^2_{1-(\alpha/2)}$ are based on $n - 1$ degrees of freedom.

The assumption that the sampled population is normally distributed must hold fairly closely for the statistical inferences just given about σ^2 to be valid. When we check this assumption in the camshaft case, we find that a histogram (not given here) of the sample of $n = 30$ hardness depths is bell-shaped and symmetrical. In order to compute a 95 percent confidence interval for σ^2, we note that $\chi^2_{\alpha/2}$ is $\chi^2_{.025}$ and $\chi^2_{1-(\alpha/2)}$ is $\chi^2_{.975}$. Table A.17 (page 833) tells us that these

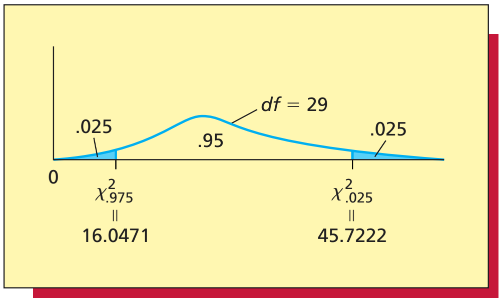


FIGURE 8.17 The Chi-Square Points $\chi^2_{.025}$ and $\chi^2_{.975}$

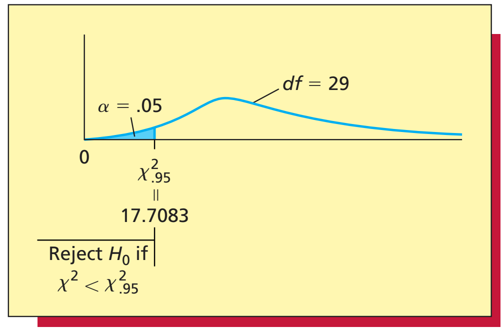


FIGURE 8.18 Testing H_0: $\sigma^2 = .47$ versus H_a: $\sigma^2 < .47$ by Setting $\alpha = .05$

points—based on $n - 1 = 29$ degrees of freedom—are $\chi^2_{.025} = 45.7222$ and $\chi^2_{.975} = 16.0471$ (see Figure 8.17). It follows that a 95 percent confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}\right] = \left[\frac{(29)(.1885)}{45.7222}, \frac{(29)(.1885)}{16.0471}\right]$$
$$= [.1196, .3407]$$

This interval provides strong evidence that σ^2 is less than .47. To formally test H_0: $\sigma^2 = .47$ versus H_a: $\sigma^2 < .47$, we calculate the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(29)(.1885)}{.47} = 11.6309$$

If we set $\alpha = .05$, we use the rejection point $\chi^2_{1-\alpha} = \chi^2_{.95} = 17.7083$ (see Figure 8.18). Because the test statistic is less than this rejection point, we reject H_0 by setting $\alpha = .05$. To see if there has been a practically important decrease in σ^2, note that $s^2 = .1885$ implies that $s = .4342$. It follows, since $\bar{x} = 4.5$, that we estimate that 99.73 percent of all individual camshaft hardness depths are in the interval $[\bar{x} \pm 3s] = [4.5 \pm 3(.4342)] = [3.20, 5.80]$ and are, therefore, within specifications (recall that hardness depth specifications are [3, 6]). This was not the case with the original σ^2 of .47 (or σ of .6856), because $[4.5 \pm 3(.6856)] = [2.44, 6.56]$ is not within the specifications of [3, 6]. Therefore, reducing σ^2 from .47 to an estimated value of .1885 seems to have practical importance. In Appendix 8.3 we show a MegaStat output of the above confidence interval for σ^2 and of the test of H_0: $\sigma^2 = .47$ versus H_a: $\sigma^2 < .47$.

Exercises for Sections 8.7 and 8.8

8.87, 8.88, 8.89, 8.90

CONCEPTS

8.85 What assumption must hold to use the chi-square distribution to make statistical inferences about a population variance?

8.86 Define the meaning of the chi-square points $\chi^2_{\alpha/2}$ and $\chi^2_{1-(\alpha/2)}$.

8.87 Give an example of a situation in which we might wish to compute a confidence interval for σ^2.

METHODS AND APPLICATIONS

Exercises 8.88 through 8.92 relate to the following situation: Consider the engine parts supplier discussed in Exercises 8.10 and 8.44 (pages 306 and 322), and suppose the supplier has determined that the variance of the population of all cylindrical engine part outside diameters produced by the current machine is approximately equal to, but no less than, .0005. To reduce this variance, a new machine is designed, and a random sample of $n = 25$ outside diameters produced by this new machine has a mean of $\bar{x} = 3$ and a variance of $s^2 = .00014$. Assume the population of all cylindrical engine part outside diameters that would be produced by the new machine is normally distributed, and let σ^2 denote the variance of this population.

8.88 Find a 95 percent confidence interval for σ^2.

8.89 Test H_0: $\sigma^2 = .0005$ versus H_a: $\sigma^2 < .0005$ by setting $\alpha = .05$.

8.90 Specifications state that the outside diameter of each cylindrical engine part must be between 2.95 inches and 3.05 inches. Do we estimate that we have reduced σ enough to meet these specifications? Explain your answer.

8.91 Find a 99 percent confidence interval for σ^2.

8.92 Test H_0: $\sigma^2 = .0005$ versus H_a: $\sigma^2 \neq .0005$ by setting $\alpha = .01$.

Chapter Summary

We began this chapter by learning about the two hypotheses that make up the structure of a hypothesis test. The **null hypothesis** is the statement being tested. Usually it represents the *status quo* and it is not rejected unless there is convincing sample evidence that it is false. The **alternative,** or, **research, hypothesis** is a statement that is accepted only if there is convincing sample evidence that it is true and that the null hypothesis is false. In some situations, the alternative hypothesis is a condition for which we need to attempt to find supportive evidence. We also learned that two types of errors can be made in a hypothesis test. A **Type I error** occurs when we reject a true null hypothesis, and a **Type II error** occurs when we do not reject a false null hypothesis.

We studied two commonly used ways to conduct a hypothesis test. The first involves comparing the value of a test statistic with what is called a **rejection point,** and the second employs what is called a ***p*-value.** The p-value measures the weight of evidence against the null hypothesis. The smaller the p-value, the more we doubt the null hypothesis. We learned that, if we can reject the null hypothesis with the probability of a Type I error equal to α, then we say that the test result has **statistical significance at the α level.** However, we also learned that, even if the result of a hypothesis test tells us that statistical significance exists, we must carefully assess whether the result is practically important. One good way to do this is to use a confidence interval.

The specific hypothesis tests we covered in this chapter all dealt with a hypothesis about one population parameter. First, we studied a **large sample test** about a **population mean,** which is based on the **normal distribution,** and second we studied a **small sample test** about a **population mean,** which is based on the ***t* distribution.** Figure 8.19 presents a flowchart summarizing how to select an appropriate test statistic to test a hypothesis about a population mean. Then we presented a test about a **population proportion** that is based on the **normal distribution.** Next we studied Type II error probabilities, and we showed how we can find the sample size needed to make both the probability of a Type I error and the probability of a serious Type II error as small as we wish. We concluded this chapter by discussing the **chi-square distribution** and its use in making statistical inferences about a population variance.

Glossary of Terms

alternative (research) hypothesis: A statement that will be accepted only if there is convincing sample evidence that it is true. Sometimes it is a condition for which we need to attempt to find supportive evidence. (page 301)

chi-square distribution: A useful continuous probability distribution. Its probability curve is skewed to the right, and the exact shape of the probability curve depends on the number of degrees of freedom associated with the curve. (page 340)

greater than alternative: An alternative hypothesis that is stated as a *greater than* (>) inequality. (page 303)

less than alternative: An alternative hypothesis that is stated as a *less than* (<) inequality. (page 303)

not equal to alternative: An alternative hypothesis that is stated as a *not equal to* (≠) inequality. (page 303)

null hypothesis: The statement being tested in a hypothesis test. It usually represents the status quo and it is not rejected unless there is convincing sample evidence that it is false. (page 301)

one-sided alternative hypothesis: An alternative hypothesis that is stated as either a *greater than* (>) or a *less than* (<) inequality. (page 303)

power (of a statistical test): The probability of rejecting the null hypothesis when it is false. (page 337)

***p*-value (probability value):** The probability, computed assuming that the null hypothesis is true, of observing a value of the test statistic that is at least as extreme as the value actually computed from the sample data. The p-value measures how much doubt is cast on the null hypothesis by the sample data. The smaller the p-value, the more we doubt the null hypothesis. (page 310)

rejection point (or **critical point**)**:** The value of the test statistic is compared with a rejection point in order to decide whether the null hypothesis can be rejected. (page 308)

statistical significance at the α level: When we can reject the null hypothesis by setting the probability of a Type I error equal to α. (page 320)

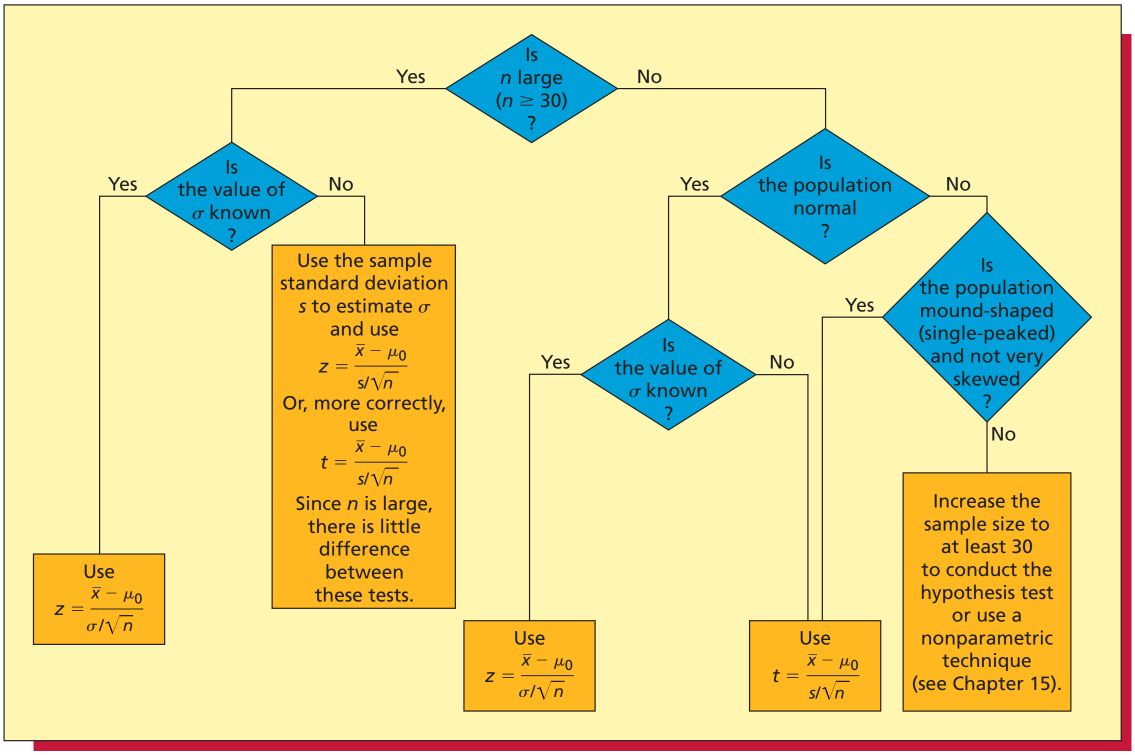
test statistic: A statistic computed from sample data in a hypothesis test. It is either compared with a rejection point or used to compute a p-value. (pages 304 and 308)

two-sided alternative hypothesis: An alternative hypothesis that is stated as a *not equal to* (≠) inequality. (page 303)

Type I error: Rejecting a true null hypothesis. (page 304)

Type II error: Failing to reject a false null hypothesis. (page 304)

FIGURE 8.19 Selecting an Appropriate Test Statistic to Test a Hypothesis about a Population Mean



Important Formulas and Tests

A large sample hypothesis test about a population mean: page 319
A small sample hypothesis test about a population mean: page 323
A large sample hypothesis test about a population proportion: page 329
Calculating the probability of a Type II error: page 337
Sample size determination to achieve specified values of α and β: page 338
Statistical inference about a population variance: page 341

Supplementary Exercises

8.93 The auditor for a large corporation routinely monitors cash disbursements. As part of this process, the auditor examines check request forms to determine whether they have been properly approved. Improper approval can occur in several ways. For instance, the check may have no approval, the check request might be missing, the approval might be written by an unauthorized person, or the dollar limit of the authorizing person might be exceeded.

a Last year the corporation experienced a 5 percent improper check request approval rate. Since this was considered unacceptable, efforts were made to reduce the rate of improper approvals. Letting p be the proportion of all checks that are now improperly approved, set up the null and alternative hypotheses needed to attempt to demonstrate that the current rate of improper approvals is lower than last year's rate of 5 percent.

b Suppose that the auditor selects a random sample of 625 checks that have been approved in the last month. The auditor finds that 18 of these 625 checks have been improperly approved. Use rejection points and this sample information to test the hypotheses you set up in part *a* at the .10, .05, .01, and .001 levels of significance. How much evidence is there that the rate of improper approvals has been reduced below last year's 5 percent rate?

c Find the p-value for the test of part *b*. Use the p-value to carry out the test by setting α equal to .10, .05, .01, and .001. Interpret your results.

d Suppose the corporation incurs a \$10 cost to detect and correct an improperly approved check. If the corporation disburses at least 2 million checks per year, does the observed reduction of the rate of improper approvals seem to have practical importance? Explain your opinion.

8.94 THE CIGARETTE ADVERTISEMENT CASE ModelAge

Recall that the cigarette industry requires that models in cigarette ads must appear to be at least 25 years old. Also recall that a sample of 50 people is randomly selected at a shopping mall. Each person in the sample is shown a "typical cigarette ad" and is asked to estimate the age of the model in the ad.

a Let μ be the mean perceived age estimate for all viewers of the ad, and suppose we consider the industry requirement to be met if μ is at least 25. Set up the null and alternative hypotheses needed to attempt to show that the industry requirement is not being met.

b Suppose that a random sample of 50 perceived age estimates gives a mean of $\bar{x} = 23.663$ years and a standard deviation of $s = 3.596$ years. Use these sample data to test the hypotheses of part *a* at the .10, .05, .01, and .001 levels of significance.

c How much evidence do we have that the industry requirement is not being met?

d Do you think that this result has practical importance? Explain your opinion.

8.95 THE CIGARETTE ADVERTISEMENT CASE ModelAge

Consider the cigarette ad situation discussed in Exercise 8.94. Using the sample information given in that exercise,

a Calculate the *p*-value for testing H_0 versus H_a.

b Determine whether H_0 would be rejected at each of $\alpha = .10$, $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.

c Describe how much evidence we have that the industry requirement is not being met.

8.96 In an article in the *Journal of Retailing,* Kumar, Kerwin, and Pereira study factors affecting merger and acquisition activity in retailing. As part of the study, the authors compare the characteristics of "target firms" (firms targeted for acquisition) and "bidder firms" (firms attempting to make acquisitions). Among the variables studied in the comparison were earnings per share, debt-to-equity ratio, growth rate of sales, market share, and extent of diversification.

a Let μ be the mean growth rate of sales for all target firms (firms that have been targeted for acquisition in the last five years and that have not bid on other firms), and assume growth rates are approximately normally distributed. Furthermore, suppose a random sample of 25 target firms yields a sample mean sales growth rate of $\bar{x} = 0.16$ with a standard deviation of $s = 0.12$. Use rejection points and this sample information to test H_0: $\mu \leq .10$ versus H_a: $\mu > .10$ by setting α equal to .10, .05, .01, and .001. How much evidence is there that the mean growth rate of sales for target firms exceeds .10 (that is, exceeds 10 percent)?

b Now let μ be the mean growth rate of sales for all firms that are bidders (firms that have bid to acquire at least one other firm in the last five years), and again assume growth rates are approximately normally distributed. Furthermore, suppose a random sample of 25 bidders yields a sample mean sales growth rate of $\bar{x} = 0.12$ with a standard deviation of $s = 0.09$. Use rejection points and this sample information to test H_0: $\mu \leq .10$ versus H_a: $\mu > .10$ by setting α equal to .10, .05, .01, and .001. How much evidence is there that the mean growth rate of sales for bidders exceeds .10 (that is, exceeds 10 percent)?

8.97 A consumer electronics firm has developed a new type of remote control button that is designed to operate longer before becoming intermittent. A random sample of 35 of the new buttons is selected and each is tested in continuous operation until becoming intermittent. The resulting lifetimes are found to have a sample mean of $\bar{x} = 1{,}241.2$ hours and a sample standard deviation of $s = 110.8$.

a Independent tests reveal that the mean lifetime (in continuous operation) of the best remote control button on the market is 1,200 hours. Letting μ be the mean lifetime of the population of all remote control buttons that will or could potentially be produced, set up the null and alternative hypotheses needed to attempt to provide evidence that the new button's mean lifetime exceeds the mean lifetime of the best remote button currently on the market.

b Using the previously given sample results, test the hypotheses you set up in part *a* by setting α equal to .10, .05, .01, and .001. What do you conclude for each value of α?

c Suppose that $\bar{x} = 1{,}241.2$ and $s = 110.8$ had been obtained by testing a sample of 100 buttons. Test the hypotheses you set up in part *a* by setting α equal to .10, .05, .01, and .001. Which sample (the sample of 35 or the sample of 100) gives a more statistically significant result? That is, which sample provides stronger evidence that H_a is true?

d If we define practical importance to mean that μ exceeds 1,200 by an amount that would be clearly noticeable to most consumers, do you think that the result has practical importance? Explain why the samples of 35 and 100 both indicate the same degree of practical importance.

e Suppose that further research and development effort improves the new remote control button and that a random sample of 35 buttons gives $\bar{x} = 1{,}524.6$ hours and $s = 102.8$ hours. Test your hypotheses of part *a* by setting α equal to .10, .05, .01, and .001.
   (1) Do we have a highly statistically significant result? Explain.
   (2) Do you think we have a practically important result? Explain.

8.98 Again consider the remote control button lifetime situation discussed in Exercise 8.97.

a Use the sample information given in part *a* of Exercise 8.97 to do the following:
   (1) Calculate the *p*-value for testing H_0 versus H_a.
   (2) Determine whether H_0 would be rejected at each of $\alpha = .10$, $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.
   (3) Describe how much evidence we have that the new button's mean lifetime exceeds the mean lifetime of the best remote button currently on the market.

b Repeat (1), (2), and (3) using the sample information given in part *c* of Exercise 8.97.

c Repeat (1), (2), and (3) using the sample information given in part *e* of Exercise 8.97.

8.99 Calculate and use an appropriate 95 percent confidence interval to help evaluate practical importance as it relates to the hypothesis test in each of the following situations discussed in previous review exercises. Explain what you think each confidence interval says about practical importance.

a The check approval situation of Exercise 8.93.

b The cigarette ad situation of Exercise 8.94.

c The remote control button situation of Exercise 8.97*a*, *c*, and *e*.

8.100 Several industries located along the Ohio River discharge a toxic substance called carbon tetrachloride into the river. The state Environmental Protection Agency monitors the amount of carbon tetrachloride pollution in the river. Specifically, the agency requires that the carbon tetrachloride contamination must average no more than 10 parts per million. In order to monitor the carbon tetrachloride contamination in the river, the agency takes a daily sample of 100 pollution readings at a specified location. If the mean carbon tetrachloride reading for this sample casts substantial doubt on the hypothesis that the average amount of carbon tetrachloride contamination in the river is at most 10 parts per million, the agency must issue a shutdown order. In the event of such a shutdown order, industrial plants along the river must be closed until the carbon tetrachloride contamination is reduced to a more acceptable level. Assume that the state Environmental Protection Agency decides to issue a shutdown order if a sample of 100 pollution readings implies that H_0: $\mu \leq 10$ can be rejected in favor of H_a: $\mu > 10$ by setting $\alpha = .01$. If we estimate that σ equals 2, calculate the probability of a Type II error for each of the following alternative values of μ: 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, 10.7, 10.8, 10.9, and 11.0.

8.101 THE INVESTMENT CASE InvestRet

Suppose that random samples of 50 returns for each of the following investment classes give the indicated sample mean and sample standard deviation:

Fixed annuities: $\bar{x} = 7.83\%$, $s = .51\%$
Domestic large-cap stocks: $\bar{x} = 13.42\%$, $s = 15.17\%$
Domestic midcap stocks: $\bar{x} = 15.03\%$, $s = 18.44\%$
Domestic small-cap stocks: $\bar{x} = 22.51\%$, $s = 21.75\%$

a For each investment class, set up the null and alternative hypotheses needed to test whether the current mean return differs from the historical (1970 to 1994) mean return given in Table 2.16 (page 107).

b Test each hypothesis you set up in part *a* at the .05 level of significance. What do you conclude? For which investment classes does the current mean return differ from the historical mean?

8.102 THE INTERNATIONAL BUSINESS TRAVEL EXPENSE CASE

Recall that the mean and the standard deviation of a random sample of 35 one-day travel expenses in Moscow are $\bar{x} = \$538$ and $s = \$41$.

a If μ denotes the mean of all one-day travel expenses in Moscow, set up the null and alternative hypotheses needed to attempt to provide evidence supporting the claim that μ is more than \$500.

b Test the hypotheses you set up in part *a* by setting $\alpha = .10, .05, .01$, and .001. How much evidence is there that the mean of all one-day travel expenses in Moscow is more than \$500?

8.103 THE UNITED KINGDOM INSURANCE CASE

Assume that the U.K. insurance survey is based on 1,000 randomly selected United Kingdom households and that 640 of these households spent on life insurance in 1993.

a If p denotes the proportion of all U.K. households that spent on life insurance in 1993, set up the null and alternative hypotheses needed to attempt to justify the claim that more than 60 percent of U.K. households spent on life insurance in 1993.

b Test the hypotheses you set up in part *a* by setting $\alpha = .10, .05, .01$, and .001. How much evidence is there that more than 60 percent of U.K. households spent on life insurance in 1993?

8.104 A bank's return on equity measures how well the bank uses its equity capital. Suppose that a random sample of six Ohio/Kentucky-area banks yields the following return on equity figures.[7]

Star Bank, Cincinnati	19.27%	First National Bank, Lebanon, Ohio	13.16%
Fifth Third Bank, Northern Kentucky	21.38%	Heritage Bank, Burlington, Ohio	9.65%
Liberty National Bank, Northern Kentucky	11.64%	Peoples Bank of Northern Kentucky	6.30%

RetEq

a The overall industry average return on equity is 14.82 percent. Set up the null and alternative hypotheses needed to test whether the mean return on equity for all Ohio/Kentucky banks differs from the overall industry average.

b Use the sample data to test the hypotheses you set up in part *a* at the .05 level of significance. Do you reject the null hypothesis?

c Based on the result of your test in part *b*, what do you conclude about whether the mean return on equity for all Ohio/Kentucky banks differs from the industry average?

8.105 Suppose that in January 1999 it was reported that 58 percent of Americans rated the condition of the economy to be "fairly good." In a Gannett News Service poll[8] of 1,003 Americans taken October 15–20, 1999, and reported in the October 31, 1999, issue of the *Cincinnati Enquirer,* 62 percent of those polled rated the condition of the economy to be "fairly good."

a Set up and carry out a hypothesis test that attempts to provide evidence supporting the claim that the proportion of Americans rating the condition of the economy to be "fairly good" in October 1999 is higher than the proportion of 58 percent in January 1999.

b Use the sample results to determine how much evidence there is that the proportion of Americans rating the condition of the economy to be "fairly good" in October 1999 is higher than the proportion of 58 percent in January 1999.

8.106 Internet Exercise

Are American consumers comfortable using their credit cards to make purchases over the Internet? Suppose that a noted authority suggests that credit cards will be firmly established on the Internet once the 80 percent barrier is broken; that is, as soon as more than 80 percent of those who make purchases over the Internet are willing to use a credit card to pay for their transactions. A recent Gallup Poll (story, survey results, and analysis can be found at http://www.gallup.com/poll/releases/pr000223.asp) found that, out of $n = 302$ Internet purchasers surveyed, 267 have paid for Internet purchases using a credit card. Based on the results of the Gallup survey, is there sufficient evidence to conclude that the proportion of Internet purchasers willing to use a credit card now exceeds 0.80? Set up the appropriate null and alternative hypotheses, test at the 0.05 and 0.01 levels of significance, and calculate a p-value for your test.

Go to the Gallup Organization website (http://www.gallup.com) and find the index of recent poll results (http://www.gallup.com/poll/index.asp). Select an interesting current poll and prepare a brief written summary of the poll or some aspect thereof. Include a statistical test for the significance of a proportion (you may have to make up your own value for the hypothesized proportion p_0) as part of your report. For example, you might select a political poll and test whether a particular candidate is preferred by a majority of voters ($p > 0.50$).

[7]Source: "A Report of Bank Profitability for Greater Cincinnati, Ohio Banks," *Cincinnati Enquirer,* June 4, 1995.

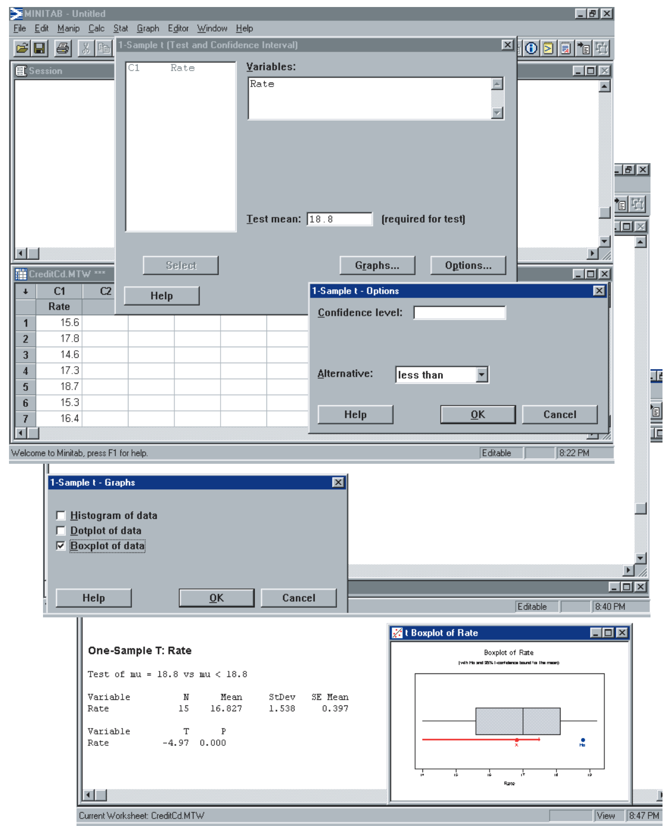
[8]Source: C. Raasch, "At 2000, Americans Guardedly Optimistic," Gannett News Service, *Cincinnati Enquirer,* October 31, 1999, p. A1.

Appendix 8.1 ■ One-Sample Hypothesis Testing Using MINITAB

The instruction block in this section begins by describing the entry of data into the MINITAB Data window. Alternatively, the data may be loaded directly from the data disk included with the text. The appropriate data file name is given at the top of the instruction block. Please refer to Appendix 1.1 for further information about entering data, saving, and printing results.

Hypothesis test for a population mean in Figure 8.10(a) on page 325 (data file: CreditCd.mtw):

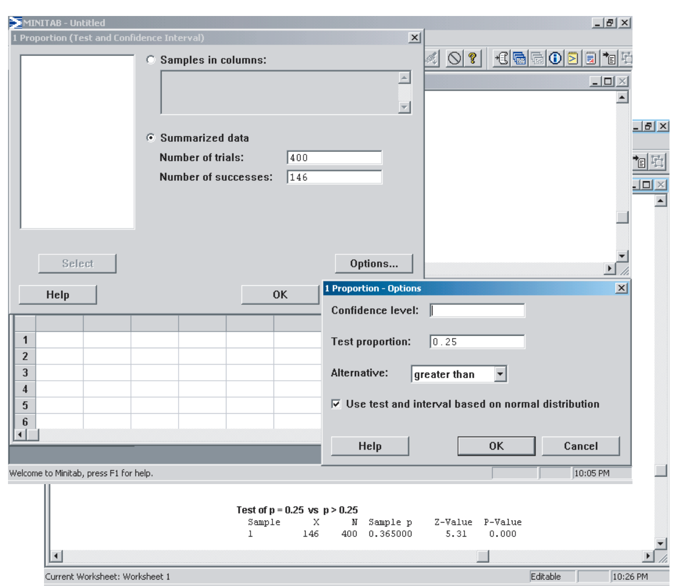
- In the Data window, enter the interest rate data from Table 8.3 (page 324) into a single column named Rate.
- Select **Stat : Basic Statistics : 1-Sample t.**
- In the "1-Sample t (Test and Confidence Interval)" dialog box, select the Rate variable into the Variables box.
- Enter 18.8 into the "Test mean" box.
- Click the Options . . . button, select "less than" in the Alternative box, and click OK in the "1-Sample t Options" dialog box.
- To produce a box plot of the data with a graphical representation of the hypothesis test, click the Graphs . . . button, check the "Boxplot of data" check box, and click OK in the "1-Sample t Graphs" dialog box.
- Click OK in the "1-Sample t (Test and Confidence Interval)" dialog box.
- The confidence interval is given in the session window, and the box plot appears in a graphics window.



A "1-Sample Z" test is also available in MINITAB under Basic Statistics. It requires a user-specified value of the population standard deviation, which is rarely known.

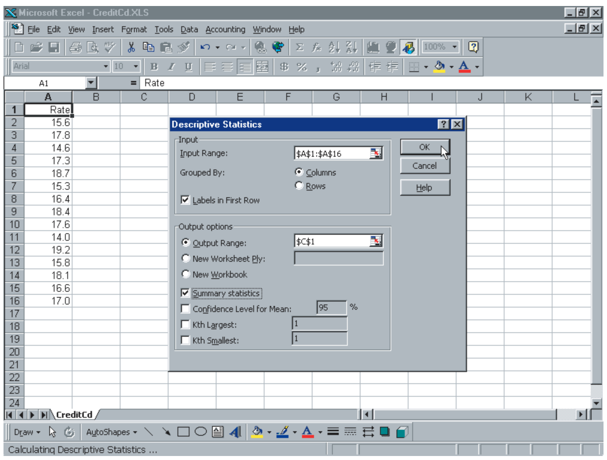
Hypothesis test for a population proportion in Exercise 8.72 on page 332:

- Select **Stat : Basic Statistics : 1 Proportion.**
- In the "1 Proportion (Test and Confidence Interval)" dialog box, click on "Summarized data".
- Enter the sample size (here equal to 400) into the "Number of trials" box.
- Enter the sample number of successes (here equal to 146) into the "Number of successes" box.
- Click on the Options button.
- In the "1 Proportion—Options" dialog box, enter the hypothesized proportion (here equal to 0.25) into the "Test proportion" box.
- Select the desired alternative (in this case "greater than") from the Alternative drop-down menu.
- Check the "Use test and interval based on normal distribution" check box.
- Click OK in the "1 Proportion—Options" dialog box.
- Click OK in the "1 Proportion (Test and Confidence Interval)" dialog box.



Appendix 8.2 ■ One-Sample Hypothesis Testing Using Excel

The instruction blocks in this section each begin by describing the entry of data into an Excel spreadsheet. Alternatively, the data may be loaded directly from the data disk included with the text. The appropriate data file name is given at the top of each instruction block. Please refer to Appendix 1.2 for further information about entering data, saving, and printing results.

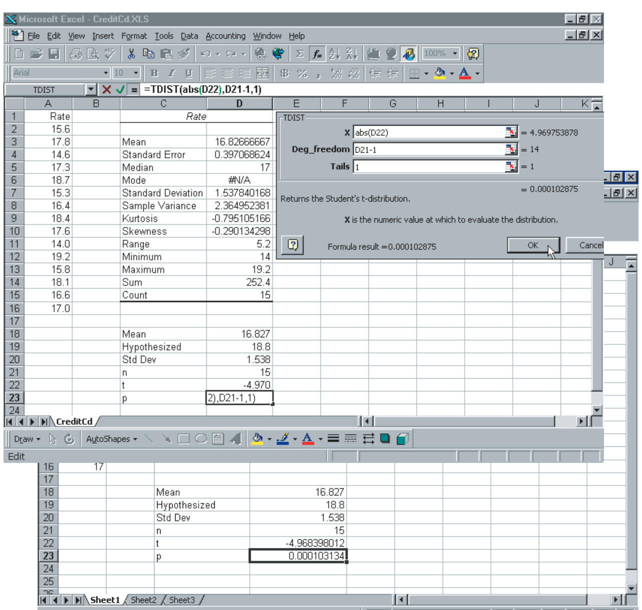


Hypothesis test for a population mean in Figure 8.10(b) on page 325 (data file: CreditCd.xls): The Data Analysis ToolPak in Excel does not explicitly provide for one-sample tests of hypotheses. A one-sample test can be conducted using the Descriptive Statistics component of the Analysis ToolPak and a few additional computations using Excel.

Descriptive statistics:

- Enter the interest rate data from Table 8.3 (page 324) into cells A2.A16 with the label Rate in cell A1.
- Select **Tools : Data Analysis : Descriptive Statistics.**
- Click OK in the Data Analysis dialog box.
- Enter A1.A16 into the Input Range box.
- Click the "Labels in First Row" check box.
- Click Output Range and enter C1 in the Output Range box.
- Click the Summary Statistics check box.
- Click OK in the Descriptive Statistics dialog box.

The resulting block of descriptive statistics is displayed in the lower right panel in the range C1.D15, and values used in the test computations have been carried down into the range C18.C21 (data file: CreditCdSol.xls).



Computation of test statistic and *p*-value:

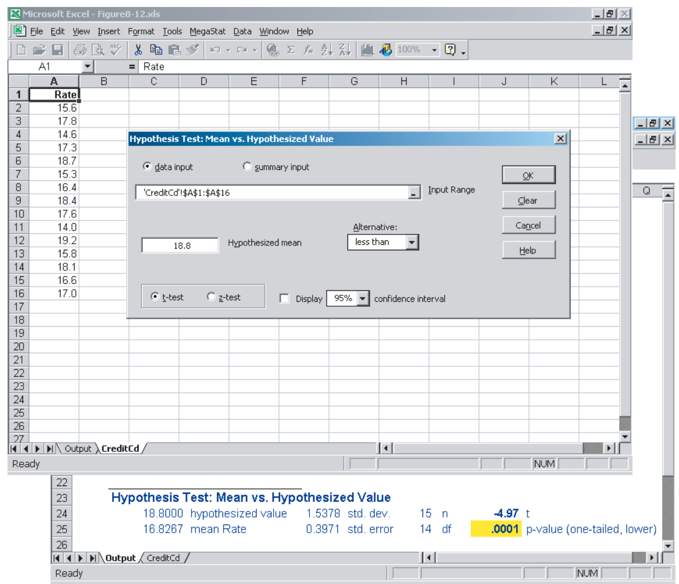
- In cell D22, use the formula =(D18-D19)/(D20/SQRT(D21)) to compute the test statistic *t*.
- Click on cell D23 and then click on the Paste Function button f_x on the Excel toolbar.
- Select Statistical from the Function Category menu, and select TDIST from the Function Name menu; click OK in the Paste Function dialog box.
- In the TDIST dialog box, enter abs(D22) in the X box.
- Enter D21 − 1 in the Deg_freedom box.
- Enter 1 in the Tails box to select a one-tailed test.
- Click OK in the TDIST dialog box.
- The *p*-value will appear in cell D23.

Appendix 8.3 ■ One-Sample Hypothesis Testing Using MegaStat

The instructions in this section begin by describing the entry of data into an Excel worksheet. Alternatively, the data may be loaded directly from the data disk included with the text. The appropriate data file name is given at the top of each instruction block. Please refer to Appendix 1.2 for further information about entering data and saving and printing results in Excel. Please refer to Appendix 1.3 for more information about using MegaStat.

Hypothesis test for a population mean in Figure 8.11 on page 325 (data file:CreditCD.xls):

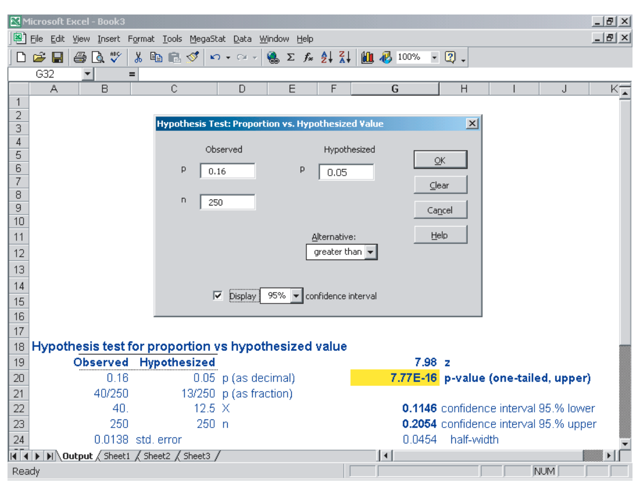
- Enter the interest rate data from Table 8.3 (page 324) into cells A2.A16 with the label Rate in cell A1.
- Select **MegaStat : Hypothesis Tests : Mean vs. Hypothesized Value.**
- In the "Hypothesis Test : Mean vs. Hypothesized Value" dialog box, click on "data input" and use the AutoExpand feature to enter the range A1.A16 into the Input Range box.
- Enter the hypothesized value (here equal to 18.8) into the Hypothesized mean box.
- Select the desired alternative (here "less than") from the drop-down menu in the Alternative box.
- Click on t-test and click OK in the "Hypothesis Test : Mean vs. Hypothesized Value" dialog box.
- A hypothesis test employing summary data can be carried out by clicking on "summary input" and by entering a range into the Input Range box that contains the following—label; sample mean; sample standard deviation; sample size *n.*

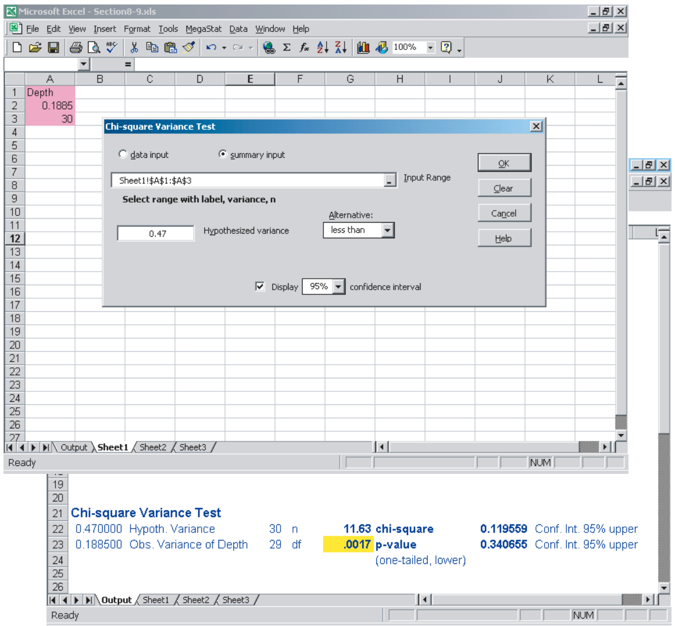


A *z* test can be carried out for large samples (or in the unlikely event that the population standard deviation is known) by clicking on "z-test".

Hypothesis test for a population proportion in the electronic article surveillance situation of Example 8.8 on page 331:

- Select **MegaStat : Hypothesis Tests : Proportion vs. Hypothesized Value.**
- In the "Hypothesis Test : Proportion vs. Hypothesized Value" dialog box, enter the hypothesized value (here equal to 0.05) into the "Hypothesized p" box.
- Enter the observed sample proportion (here equal to 0.16) into the "Observed p" box.
- Enter the sample size (here equal to 250) into the "n" box.
- Select the desired alternative (here "greater than") from the drop-down menu in the Alternative box.
- Check the "Display confidence interval" check box (if desired), and select or type the appropriate level of confidence.
- Click OK in the "Hypothesis Test: Proportion vs. Hypothesized Value" dialog box.





Hypothesis test for a population variance in the camshaft situation of Section 8.8 on page 342:

- Select **MegaStat : Hypothesis Tests : Chi-square Variance Test.**
- Enter a label (in this case Depth) into cell A1, the sample variance (here equal to .1855) into cell A2, and the sample size (here equal to 30) into cell A3.
- Click on "summary input."
- Enter the range A1.A3 into the Input Range box—that is, enter the range containing the data label, the sample variance, and the sample size.
- Enter the hypothesized value (here equal to 0.47) into the "Hypothesized variance" box.
- Select the desired alternative (in this case "less than") from the drop-down menu in the Alternative box.
- Check the "Display confidence interval" check box (if desired) and select or type the appropriate level of confidence.
- Click OK in the "Chi-square Variance Test" dialog box.
- A chi-square variance test may be carried out using "data input" by entering the observed sample values into a column in the Excel worksheet, and by then using the AutoExpand feature to enter the range containing the label and sample values into the Input Range box.