

Douglas C. Montgomery • George C. Runger

APPLIED STATISTICS AND PROBABILITY FOR ENGINEERS

7th Edition

WILEY

Statistical Intervals for a Single Sample

© jkullander / iStockphoto



LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Construct confidence intervals on the mean of a normal distribution, using either the normal distribution or the t distribution method
 2. Construct confidence intervals on the variance and standard deviation of a normal distribution
 3. Construct confidence intervals on a population proportion
 4. Use a general method for constructing an approximate confidence interval on a parameter
 5. Construct a prediction interval for a future observation
 6. Construct a tolerance interval for a normal distribution
 7. Explain the three types of interval estimates: confidence intervals, prediction intervals, and tolerance intervals
-

CHAPTER OUTLINE

8.1 Confidence Interval on the Mean of a Normal Distribution, Variance Known**8.1.1** Development of the Confidence Interval and Its Basic Properties**8.1.2** Choice of Sample Size**8.1.3** One-Sided Confidence Bounds**8.1.4** General Method to Derive a Confidence Interval**8.1.5** Large-Sample Confidence Interval for μ **8.2 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown****8.2.1** t Distribution**8.2.2** t Confidence Interval on μ **8.3 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution****8.4 Large-Sample Confidence Interval for a Population Proportion****8.5 Guidelines for Constructing Confidence Intervals****8.6 Bootstrap Confidence Interval****8.7 Tolerance and Prediction Intervals****8.7.1** Prediction Interval for a Future Observation**8.7.2** Tolerance Interval for a Normal Distribution

Introduction

Engineers are often involved in estimating parameters. For example, there is an ASTM Standard E23 that defines a technique called the *Charpy V-notch method for notched bar impact testing* of metallic materials. The impact energy is often used to determine whether the material experiences a ductile-to-brittle transition as the temperature decreases. Suppose that we have tested a sample of 10 specimens of a particular material with this procedure. We know that we can use the sample average \bar{X} to estimate the true mean impact energy μ . However, we also know that the true mean impact energy is unlikely to be exactly equal to your estimate. Reporting the results of your test as a single number is unappealing because nothing inherent in \bar{X} provides any information about how close it is to the true value of the mean μ . Our estimate could be very close, or it could be considerably far from the true mean. A way to avoid this is to report the estimate in terms of a range of plausible values called a **confidence interval**. A confidence interval always specifies a confidence level, usually 90%, 95%, or 99%, which is a measure of the reliability of the procedure. So if a 95% confidence interval on the impact energy based on the data from our 10 specimens has a lower limit of 63.84J and an upper limit of 65.08J, then we can say that at the 95% level of confidence any value of *mean* impact energy between 63.84J and 65.08J is a plausible value. By *reliability*, we mean that if we repeated this experiment over and over again, 95% of all samples would produce a confidence interval that contains the true mean impact energy, and only 5% of the time would the interval be in error. In this chapter, you learn how to construct confidence intervals and other useful types of statistical intervals for many important types of problem situations.

In the previous chapter, we illustrated how a point estimate of a parameter can be estimated from sample data. However, it is important to understand how good the estimate obtained is. For example, suppose that we estimate the mean viscosity of a chemical product to be $\hat{\mu} = \bar{x} = 1000$. Now because of sampling variability, it is almost never the case that the true mean μ is exactly equal to the estimate \bar{x} . The point estimate says nothing about how close $\hat{\mu}$ is to μ . Is the process mean likely to be between 900 and 1100? Or is it likely to be between 990 and 1010? The answer to these questions affects our decisions regarding this process. Bounds that represent an interval of plausible values for a parameter are examples of an interval estimate. Surprisingly, it is easy to determine such intervals in many cases, and the same data that provided the point estimate are typically used.

An interval estimate for a population parameter is called a **confidence interval**. Information about the precision of estimation is conveyed by the length of the interval. A short interval implies precise estimation. We cannot be certain that the interval contains the true, unknown population

parameter—we use only a sample from the full population to compute the point estimate and the interval. However, the confidence interval is constructed so that we have high confidence that it does contain the unknown population parameter. Confidence intervals are widely used in engineering and the sciences.

A **tolerance interval** is another important type of interval estimate. For example, the chemical product viscosity data might be assumed to be normally distributed. We might like to calculate limits that bound 95% of the viscosity values. For a normal distribution, we know that 95% of the distribution is in the interval

$$\mu - 1.96\sigma, \mu + 1.96\sigma$$

However, this is not a useful tolerance interval because the parameters μ and σ are unknown. Point estimates such as \bar{x} and s can be used in the preceding equation for μ and σ . However, we need to account for the potential error in each point estimate to form a tolerance interval for the distribution. The result is an interval of the form

$$\bar{x} - ks, \bar{x} + ks$$

where k is an appropriate constant (that is larger than 1.96 to account for the estimation error). As in the case of a confidence interval, it is not certain that the tolerance interval bounds 95% of the distribution, but the interval is constructed so that we have high confidence that it does. Tolerance intervals are widely used and, as we will subsequently see, they are easy to calculate for normal distributions.

Confidence and tolerance intervals bound unknown elements of a distribution. In this chapter, you learn to appreciate the value of these intervals. A **prediction interval** provides bounds on one (or more) *future observations* from the population. For example, a prediction interval could be used to bound a single, new measurement of viscosity—another useful interval. With a large sample size, the prediction interval for normally distributed data tends to the tolerance interval, but for more modest sample sizes, the prediction and tolerance intervals are different.

Keep the purpose of the three types of interval estimates clear:

- A confidence interval bounds population or distribution parameters (such as the mean viscosity).
- A tolerance interval bounds a selected proportion of a distribution.
- A prediction interval bounds future observations from the population or distribution.

Our experience has been that it is easy to confuse the three types of intervals. For example, a confidence interval is often reported when the problem situation calls for a prediction interval. Confidence intervals are covered in this chapter, while tolerance and prediction intervals are presented in the online material.

8.1

Confidence Interval on the Mean of a Normal Distribution, Variance Known

The basic ideas of a confidence interval (CI) are most easily understood by initially considering a simple situation. Suppose that we have a normal population with unknown mean μ and known variance σ^2 . This is a somewhat unrealistic scenario because typically both the mean and variance are unknown. However, in subsequent sections, we present confidence intervals for more general situations.

8.1.1

Development of the Confidence Interval and Its Basic Properties

Suppose that X_1, X_2, \dots, X_n is a random sample from a normal distribution with unknown mean μ and known variance σ^2 . From the results of Chapter 5, we know that the sample mean \bar{X} is

normally distributed with mean μ and variance σ^2/n . We may *standardize* \bar{X} by subtracting the mean and dividing by the standard deviation, which results in the variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (8.1)$$

The random variable Z has a standard normal distribution.

A **confidence interval** estimate for μ is an interval of the form $l \leq \mu \leq u$, where the end-points l and u are computed from the sample data. Because different samples will produce different values of l and u , these end-points are values of random variables L and U , respectively. Suppose that we can determine values of L and U such that the following probability statement is true:

$$P\{L \leq \mu \leq U\} = 1 - \alpha \quad (8.2)$$

where $0 \leq \alpha \leq 1$. There is a probability of $1 - \alpha$ of selecting a sample for which the CI will contain the true value of μ . Once we have selected the sample, so that $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, and computed l and u , the resulting confidence interval for μ is

$$l \leq \mu \leq u \quad (8.3)$$

The end-points or bounds l and u are called the **lower-** and **upper-confidence limits (bounds)**, respectively, and $1 - \alpha$ is called the **confidence coefficient**.

In our problem situation, because $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has a standard normal distribution, we may write

$$P\left\{-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right\} = 1 - \alpha$$

Now manipulate the quantities inside the brackets by (1) multiplying through by σ/\sqrt{n} , (2) subtracting \bar{X} from each term, and (3) multiplying through by -1 . This results in

$$P\left\{\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha \quad (8.4)$$

This is a *random interval* because the end-points $\bar{X} \pm Z_{\alpha/2}\sigma/\sqrt{n}$ involve the random variable \bar{X} . From consideration of Equation 8.4, the lower and upper end-points or limits of the inequalities in Equation 8.4 are the lower- and upper-confidence limits L and U , respectively. This leads to the following definition.

Confidence Interval on the Mean, Variance Known

If \bar{x} is the sample mean of a random sample of size n from a normal population with known variance σ^2 , a $100(1 - \alpha)\%$ confidence interval on μ is given by

$$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{n} \quad (8.5)$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution.

The development of this CI assumed that we are sampling from a normal population. The CI is quite robust to this assumption. That is, moderate departures from normality are of no serious concern. From a practical viewpoint, this implies that an *advertised* 95% CI might have actual confidence of 93% or 94%.

EXAMPLE 8.1 | Metallic Material Transition

ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at 60°C are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3. Assume that impact energy is normally distributed with $\sigma = 1 J$. We want to find a 95% CI for μ , the mean impact energy. The required quantities are $z_{\alpha/2} = z_{0.025} = 1.96$, $n = 10$, $\sigma = 1$, and $\bar{x} = 64.46$.

The resulting 95% CI is found from Equation 8.5 as follows:

$$\begin{aligned}\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ 64.46 - 1.96 \frac{1}{\sqrt{10}} &\leq \mu \leq 64.46 + 1.96 \frac{1}{\sqrt{10}} \\ 63.84 &\leq \mu \leq 65.08\end{aligned}$$

Practical Interpretation: Based on the sample data, a range of highly plausible values for mean impact energy for A238 steel at 60°C is $63.84 J \leq \mu \leq 65.08 J$.

Interpreting a Confidence Interval How does one interpret a confidence interval? In the impact energy estimation problem in Example 8.1, the 95% CI is $63.84 \leq \mu \leq 65.08$, so it is tempting to conclude that μ is within this interval with probability 0.95. However, with a little reflection, it is easy to see that this cannot be correct; the true value of μ is unknown, and the statement $63.84 \leq \mu \leq 65.08$ is either correct (true with probability 1) or incorrect (false with probability 1). The correct interpretation lies in the realization that a CI is a *random interval* because in the probability statement defining the end-points of the interval (Equation 8.2), L and U are random variables. Consequently, the correct interpretation of a $100(1 - \alpha)\%$ CI depends on the relative frequency view of probability. Specifically, if an infinite number of random samples are collected and a $100(1 - \alpha)\%$ confidence interval for μ is computed from each sample, $100(1 - \alpha)\%$ of these intervals will contain the true value of μ .

The situation is illustrated in Figure 8.1, which shows several $100(1 - \alpha)\%$ confidence intervals for the mean μ of a normal distribution. The dots at the center of the intervals indicate the point estimate of μ (that is, \bar{x}). Notice that one of the intervals fails to contain the true value of μ . If this were a 95% confidence interval, in the long run only 5% of the intervals would fail to contain μ .

Now in practice, we obtain only one random sample and calculate one confidence interval. Because this interval either will or will not contain the true value of μ , it is not reasonable to attach a probability level to this specific event. The appropriate statement is that the observed interval $[l, u]$ brackets the true value of μ with *confidence* $100(1 - \alpha)$. This statement has a frequency interpretation; that is, we do not know whether the statement is true for this specific sample, but the *method* used to obtain the interval $[l, u]$ yields correct statements $100(1 - \alpha)\%$ of the time.

Confidence Level and Precision of Estimation Notice that in Example 8.1, our choice of the 95% level of confidence was essentially arbitrary. What would have happened if we had chosen a higher level of confidence, say, 99%? In fact, is it not reasonable that we would want the

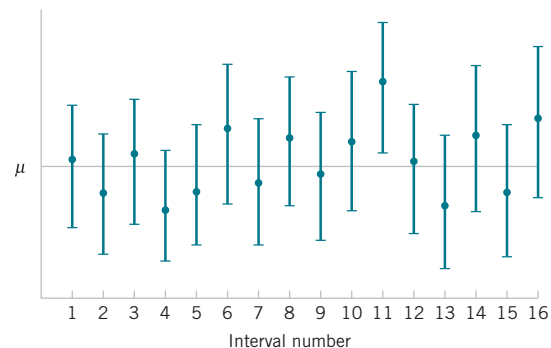


FIGURE 8.1

Repeated construction of a confidence interval for μ .

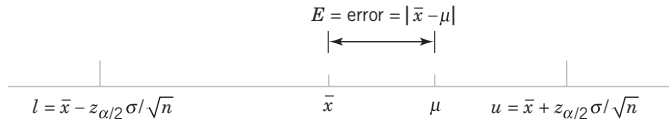


FIGURE 8.2

Error in estimating μ with \bar{x} .

higher level of confidence? At $\alpha = 0.01$, we find $z_{\alpha/2} = z_{0.01/2} = z_{0.005} = 2.58$, while for $\alpha = 0.05$, $z_{0.025} = 1.96$. Thus, the *length* of the 95% confidence interval is

$$2 \left(1.96\sigma/\sqrt{n} \right) = 3.92\sigma/\sqrt{n}$$

whereas the length of the 99% CI is

$$2 \left(2.58\sigma/\sqrt{n} \right) = 5.16\sigma/\sqrt{n}$$

Thus, the 99% CI is longer than the 95% CI. This is why we have a higher level of confidence in the 99% confidence interval. Generally, for a fixed sample size n and standard deviation σ , the higher the confidence level, the longer the resulting CI.

The length of a confidence interval is a measure of the *precision* of estimation. Many authors define the half-length of the CI (in our case $z_{\alpha/2}\sigma/\sqrt{n}$) as the bound on the error in estimation of the parameter. From the preceding discussion, we see that precision is inversely related to the confidence level. It is desirable to obtain a confidence interval that is short enough for decision-making purposes and that also has adequate confidence. One way to achieve this is by choosing the sample size n to be large enough to give a CI of specified length or precision with prescribed confidence.

8.1.2 Choice of Sample Size

The precision of the confidence interval in Equation 8.5 is $2z_{\alpha/2}\sigma/\sqrt{n}$. This means that in using \bar{x} to estimate μ , the error $E = |\bar{x} - \mu|$ is less than or equal to $z_{\alpha/2}\sigma/\sqrt{n}$ with confidence $100(1 - \alpha)\%$. This is shown graphically in Figure 8.2. In situations whose sample size can be controlled, we can choose n so that we are $100(1 - \alpha)\%$ confident that the error in estimating μ is less than a specified bound on the error E . The appropriate sample size is found by choosing n such that $z_{\alpha/2}\sigma/\sqrt{n} = E$. Solving this equation gives the following formula for n .

Sample Size for Specified Error on the Mean, Variance Known

If \bar{x} is used as an estimate of μ , we can be $100(1 - \alpha)\%$ confident that the error $|\bar{x} - \mu|$ will not exceed a specified amount E when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 \quad (8.6)$$

If the right-hand side of Equation 8.6 is not an integer, it must be rounded up. This will ensure that the level of confidence does not fall below $100(1 - \alpha)\%$. Notice that $2E$ is the length of the resulting confidence interval.

EXAMPLE 8.2 | Metallic Material Transition

To illustrate the use of this procedure, consider the CVN test described in Example 8.1 and suppose that we want to determine how many specimens must be tested to ensure that the 95% CI on μ for A238 steel cut at 60°C has a length of at most 1.0 J. Because the bound on error in estimation E is one-half of the length of the CI, to determine n , we use Equation 8.6

with $E = 0.5$, $\sigma = 1$, and $z_{\alpha/2} = 1.96$. The required sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2 = \left[\frac{(1.96)1}{0.5} \right]^2 = 15.37$$

and because n must be an integer, the required sample size is $n = 16$.

Notice the general relationship between sample size, desired length of the confidence interval $2E$, confidence level $100(1 - \alpha)$, and standard deviation σ :

- As the desired length of the interval $2E$ decreases, the required sample size n increases for a fixed value of σ and specified confidence.
- As σ increases, the required sample size n increases for a fixed desired length $2E$ and specified confidence.
- As the level of confidence increases, the required sample size n increases for fixed desired length $2E$ and standard deviation σ .

8.1.3 One-Sided Confidence Bounds

The confidence interval in Equation 8.5 gives both a lower confidence bound and an upper confidence bound for μ . Thus, it provides a two-sided CI. It is also possible to obtain one-sided confidence bounds for μ by setting either the lower bound $l = -\infty$ or the upper bound $u = \infty$ and replacing $z_{\alpha/2}$ by z_α .

One-Sided Confidence Bounds on the Mean, Variance Known

A $100(1 - \alpha)\%$ **upper-confidence bound** for μ is

$$\mu \leq \bar{x} + z_\alpha \sigma / \sqrt{n} \quad (8.7)$$

and a $100(1 - \alpha)\%$ **lower-confidence bound** for μ is

$$\bar{x} - z_\alpha \sigma / \sqrt{n} \leq \mu \quad (8.8)$$

EXAMPLE 8.3 | One-Sided Confidence Bound

The same data for impact testing from Example 8.1 are used to construct a lower, one-sided 95% confidence interval for the mean impact energy. Recall that $\bar{x} = 64.46$, $\sigma = 1J$, and $n = 10$. The interval is

$$\begin{aligned} \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} &\leq \mu \\ 64.46 - 1.64 \frac{1}{\sqrt{10}} &\leq \mu \\ 63.94 &\leq \mu \end{aligned}$$

Practical Interpretation: The lower limit for the two-sided interval in Example 8.1 was 63.84. Because $z_\alpha < z_{\alpha/2}$, the lower limit of a one-sided interval is always greater than the lower limit of a two-sided interval of equal confidence. The one-sided interval does not bound μ from above so that it still achieves 95% confidence with a slightly larger lower limit. If our interest is only in the lower limit for μ , then the one-sided interval is preferred because it provides equal confidence with a greater limit. Similarly, a one-sided upper limit is always less than a two-sided upper limit of equal confidence.

8.1.4 General Method to Derive a Confidence Interval

It is easy to give a general method for finding a confidence interval for an unknown parameter θ . Let X_1, X_2, \dots, X_n be a random sample of n observations. Suppose that we can find a statistic $g(X_1, X_2, \dots, X_n; \theta)$ with the following properties:

1. $g(X_1, X_2, \dots, X_n; \theta)$ depends on both the sample and θ .
2. The probability distribution of $g(X_1, X_2, \dots, X_n; \theta)$ does not depend on θ or any other unknown parameter.

In the case considered in this section, the parameter was $\theta = \mu$. The random variable $g(X_1, X_2, \dots, X_n; \mu) = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ satisfies both conditions; the random variable depends on the sample and on μ , and it has a standard normal distribution because σ is known. Now we must find constants C_L and C_U so that

$$P[C_L \leq g(X_1, X_2, \dots, X_n; \theta) \leq C_U] = 1 - \alpha \quad (8.9)$$

Because of property 2, C_L and C_U do not depend on θ . In our example, $C_L = -z_{\alpha/2}$ and $C_U = z_{\alpha/2}$. Finally, we must manipulate the inequalities in the probability statement so that

$$P[L(X_1, X_2, \dots, X_n) \leq \theta \leq U(X_1, X_2, \dots, X_n)] = 1 - \alpha \quad (8.10)$$

This gives $L(X_1, X_2, \dots, X_n)$ and $U(X_1, X_2, \dots, X_n)$ as the lower and upper confidence limits defining the $100(1 - \alpha)$ confidence interval for θ . The quantity $g(X_1, X_2, \dots, X_n; \theta)$ is often called a *pivotal quantity* because we pivot on this quantity in Equation 8.9 to produce Equation 8.10. In our example, we manipulated the pivotal quantity $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ to obtain $L(X_1, X_2, \dots, X_n) = \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$ and $U(X_1, X_2, \dots, X_n) = \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$.

8.1.5 Large-Sample Confidence Interval for μ

We have assumed that the population distribution is normal with unknown mean and known standard deviation σ . We now present a **large-sample CI** for μ that does not require these assumptions. Let X_1, X_2, \dots, X_n be a random sample from a population with unknown mean μ and variance σ^2 . Now if the sample size n is large, the central limit theorem implies that \bar{X} has approximately a normal distribution with mean μ and variance σ^2/n . Therefore, $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution. This ratio could be used as a pivotal quantity and manipulated as in Section 8.1.1 to produce an approximate CI for μ . However, the standard deviation σ is unknown. It turns out that when n is large, replacing σ by the sample standard deviation S has little effect on the distribution of Z . This leads to the following useful result.

Large-Sample Confidence Interval on the Mean

When n is large, the quantity

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has an approximate standard normal distribution. Consequently,

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \quad (8.11)$$

is a large-sample confidence interval for μ , with confidence level of approximately $100(1 - \alpha)\%$.

Equation 8.11 holds regardless of the shape of the population distribution. Generally, n should be at least 40 to use this result reliably. The central limit theorem generally holds for $n \geq 30$, but the larger sample size is recommended here because replacing s with S in Z results in additional variability.

EXAMPLE 8.4 | Mercury Contamination

An article in the 1993 volume of the *Transactions of the American Fisheries Society* reports the results of a study to investigate the mercury contamination in largemouth bass. A sample of fish was selected from 53 Florida lakes, and mercury concentration in the muscle tissue was measured (ppm). The mercury concentration values were

1.230	1.330	0.040	0.044	1.200	0.270
0.490	0.190	0.830	0.810	0.710	0.500
0.490	1.160	0.050	0.150	0.190	0.770
1.080	0.980	0.630	0.560	0.410	0.730
0.590	0.340	0.340	0.840	0.500	0.340
0.280	0.340	0.750	0.870	0.560	0.170
0.180	0.190	0.040	0.490	1.100	0.160
0.100	0.210	0.860	0.520	0.650	0.270
0.940	0.400	0.430	0.250	0.270	

The summary statistics for these data are as follows:

Variable	N	Mean	Median	StDev
Concentration	53	0.5250	0.4900	0.3486

Minimum	Maximum	Q1	Q3
0.0400	1.3300	0.2300	0.7900

Figure 8.3 presents the histogram and normal probability plot of the mercury concentration data. Both plots indicate that the distribution of mercury concentration is not normal and is positively skewed. We want to find an approximate 95% CI on μ . Because $n > 40$, the assumption of normality is not necessary to use in Equation 8.11. The required quantities are $n = 53$, $\bar{x} = 0.5250$, $s = 0.3486$, and $z_{0.025} = 1.96$. The approximate 95% CI on μ is

$$\bar{x} - z_{0.025} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{s}{\sqrt{n}}$$

$$0.5250 - 1.96 \frac{0.3486}{\sqrt{53}} \leq \mu \leq 0.5250 + 1.96 \frac{0.3486}{\sqrt{53}}$$

$$0.4311 \leq \mu \leq 0.6189$$

Practical Interpretation: This interval is fairly wide because there is substantial variability in the mercury concentration measurements. A larger sample size would have produced a shorter interval.

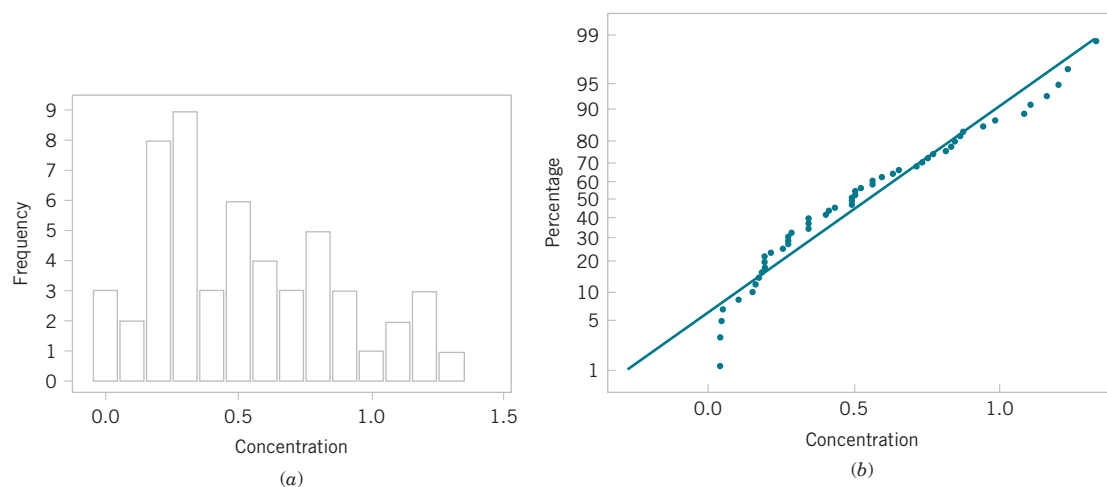


FIGURE 8.3

Mercury concentration in largemouth bass. (a) Histogram. (b) Normal probability plot.

Large-Sample Confidence Interval for a Parameter The large-sample confidence interval for μ in Equation 8.11 is a special case of a more general result. Suppose that θ is a parameter of a probability distribution, and let $\hat{\theta}$ be an estimator of θ . If $\hat{\theta}$ (1) has an approximate normal distribution, (2) is an approximately unbiased estimator for θ , and (3) has standard deviation $\sigma_{\hat{\theta}}$ that can be estimated from the sample data, the quantity $(\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$ has an approximate standard normal distribution. Then a large-sample approximate CI for θ is given by

Large-Sample Approximate Confidence Interval

$$\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}} \quad (8.12)$$

Maximum likelihood estimators usually satisfy the three conditions just listed, so Equation 8.12 is often used when $\hat{\theta}$ is the maximum likelihood estimator of θ . Finally, note that Equation 8.12 can be used even when $\sigma_{\hat{\theta}}$ is a function of other unknown parameters (or of θ). Essentially, we simply use the sample data to compute estimates of the unknown parameters and substitute those estimates into the expression for $\sigma_{\hat{\theta}}$.

8.2 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

When we are constructing confidence intervals on the mean μ of a normal population when σ^2 is known, we can use the procedure in Section 8.1.1. This CI is also approximately valid (because of the central limit theorem) regardless of whether or not the underlying population is normal so long as n is reasonably large ($n \geq 40$, say). As noted in Section 8.1.5, we can even handle the case of unknown variance for the large-sample-size situation. However, when the sample is small and σ^2 is unknown, we must make an assumption about the form of the underlying distribution to obtain a valid CI procedure. A reasonable assumption in many cases is that the underlying distribution is normal.

Many populations encountered in practice are well approximated by the normal distribution, so this assumption will lead to confidence interval procedures of wide applicability. In fact, moderate departure from normality will have little effect on validity. When the assumption is unreasonable, an alternative is to use nonparametric statistical procedures that are valid for any underlying distribution.

Suppose that the population of interest has a normal distribution with unknown mean μ and unknown variance σ^2 . Assume that a random sample of size n , say, X_1, X_2, \dots, X_n , is available, and let \bar{X} and S^2 be the sample mean and variance, respectively.

We wish to construct a two-sided CI on μ . If the variance σ^2 is known, we know that $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has a standard normal distribution. When σ^2 is unknown, a logical procedure is to replace σ with the sample standard deviation S . The random variable Z now becomes $T = (\bar{X} - \mu)/(S/\sqrt{n})$. A logical question is what effect replacing σ with S has on the distribution of the random variable T . If n is large, the answer to this question is “very little,” and we can proceed to use the confidence interval based on the normal distribution from Section 8.1.5. However, n is usually small in most engineering problems, and in this situation, a different distribution must be employed to construct the CI.

8.2.1 t Distribution **t Distribution**

Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2 . The random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (8.13)$$

has a t distribution with $n - 1$ degrees of freedom.

The t probability density function is

$$f(x) = \frac{\Gamma[(k+1)/2]}{\sqrt{k}\Gamma(k/2)} \cdot \frac{1}{[(x^2/k) + 1]^{(k+1)/2}} \quad -\infty < x < \infty \quad (8.14)$$

where k is the number of degrees of freedom. The mean and variance of the t distribution are zero and $k/(k-2)$ (for $k > 2$), respectively.

Several t distributions are shown in Figure 8.4. The general appearance of the t distribution is similar to the standard normal distribution in that both distributions are symmetric and unimodal, and the maximum ordinate value is reached when the mean $\mu = 0$. However, the t distribution has heavier tails than the normal; that is, it has more probability in the tails than does the normal distribution. As the number of degrees of freedom $k \rightarrow \infty$, the limiting form of the t distribution is the standard normal distribution. Generally, the number of degrees of freedom for t is the number of degrees of freedom associated with the estimated standard deviation.

Appendix Table V provides *percentage points* of the t distribution. We let $t_{\alpha,k}$ be the value of the random variable T with k degrees of freedom above which we find an area (or probability) α . Thus, $t_{\alpha,k}$ is an upper-tailed 100α percentage point of the t distribution with k degrees of freedom. This percentage point is shown in Figure 8.5. In Appendix Table V, the α values are the column headings, and the degrees of freedom are listed in the left column. To illustrate the use of the table, note that the t -value with 10 degrees of freedom having an area of 0.05 to the right is $t_{0.05,10} = 1.812$. That is,

$$P(T_{10} > t_{0.05,10}) = P(T_{10} > 1.812) = 0.05$$

Because the t distribution is symmetric about zero, we have $t_{1-\alpha,k} = -t_{\alpha,k}$; that is, the t -value having an area of $1 - \alpha$ to the right (and therefore an area of α to the left) is equal to the negative of the t -value that has area α in the right tail of the distribution. Therefore, $t_{0.95,10} = -t_{0.05,10} = -1.812$. Finally, because $t_{\alpha,\infty}$ is the standard normal distribution, the familiar z_α values appear in the last row of Appendix Table V.

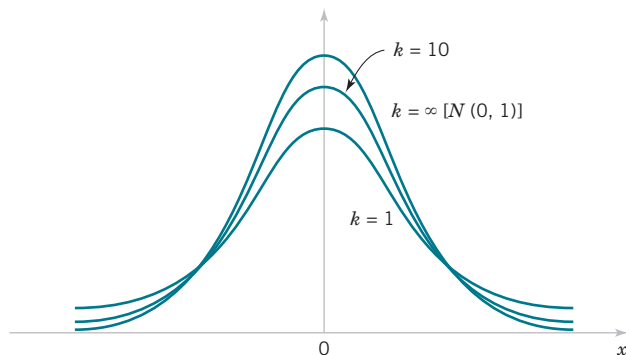


FIGURE 8.4

Probability density functions of several t distributions.

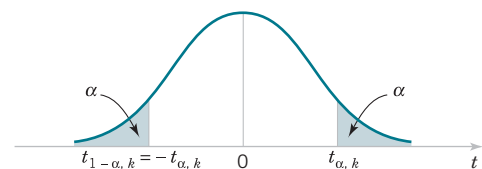


FIGURE 8.5

Percentage points of the t distribution.

8.2.2 t Confidence Interval on μ

It is easy to find a $100(1 - \alpha)\%$ confidence interval on the mean of a normal distribution with unknown variance by proceeding essentially as we did in Section 8.1.1. We know that the distribution of $T = (\bar{X} - \mu)/(S/\sqrt{n})$ is t with $n - 1$ degrees of freedom. Letting $t_{\alpha/2, n-1}$ be the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom, we may write

$$P(-t_{\alpha/2, n-1} \leq T \leq t_{\alpha/2, n-1}) = 1 - \alpha$$

or

$$P\left(-t_{\alpha/2, n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2, n-1}\right) = 1 - \alpha$$

Rearranging this last equation yields

$$P(\bar{X} - t_{\alpha/2, n-1}S/\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2, n-1}S/\sqrt{n}) = 1 - \alpha \quad (8.15)$$

This leads to the following definition of the $100(1 - \alpha)\%$ two-sided confidence interval on μ .

Confidence Interval on the Mean, Variance Unknown

If \bar{x} and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2 , a $100(1 - \alpha)\%$ confidence interval on μ is given by

$$\bar{x} - t_{\alpha/2, n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1}s/\sqrt{n} \quad (8.16)$$

where $t_{\alpha/2, n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

The assumption underlying this CI is that we are sampling from a normal population. However, the t distribution-based CI is relatively insensitive or robust to this assumption. Checking the normality assumption by constructing a normal probability plot of the data is a good general practice. Small to moderate departures from normality are not a cause for concern.

One-sided confidence bounds on the mean of a normal distribution are also of interest and are easy to find. Simply use only the appropriate lower or upper confidence limit from Equation 8.16 and replace $t_{\alpha/2, n-1}$ by $t_{\alpha, n-1}$.

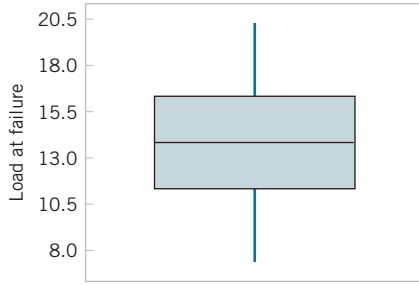
EXAMPLE 8.5 | Alloy Adhesion

An article in the *Journal of Materials Engineering* ["Instrumented Tensile Adhesion Tests on Plasma Sprayed Thermal Barrier Coatings" (1989, Vol. 11(4), pp. 275–282)] describes the results of tensile adhesion tests on 22 U-700 alloy specimens. The load at specimen failure is as follows (in megapascals):

19.8	10.1	14.9	7.5	15.4	15.4
15.4	18.5	7.9	12.7	11.9	11.4
11.4	14.1	17.6	16.7	15.8	
19.5	8.8	13.6	11.9	11.4	

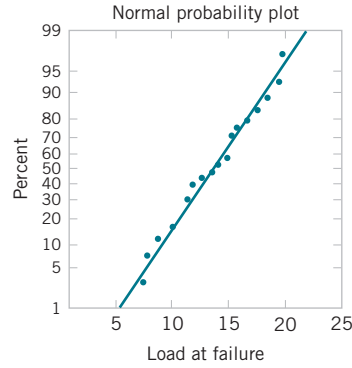
The sample mean is $\bar{x} = 13.71$, and the sample standard deviation is $s = 3.55$. Figures 8.6 and 8.7 show a box plot and a normal probability plot of the tensile adhesion test data, respectively. These displays provide good support for the assumption that the population is normally distributed. We want to find a 95% CI on μ . Since $n = 22$, we have $n - 1 = 21$ degrees of freedom for t , so $t_{0.025, 21} = 2.080$. The resulting CI is

$$\begin{aligned} \bar{x} - t_{\alpha/2, n-1}s/\sqrt{n} &\leq \mu \leq \bar{x} + t_{\alpha/2, n-1}s/\sqrt{n} \\ 13.71 - 2.080(3.55)/\sqrt{22} &\leq \mu \leq 13.71 + 2.080(3.55)/\sqrt{22} \\ 13.71 - 1.57 &\leq \mu \leq 13.71 + 1.57 \\ 12.14 &\leq \mu \leq 15.28 \end{aligned}$$

**FIGURE 8.6**

Box-and-whisker plot for the load at failure data.

Practical Interpretation: The CI is fairly wide because there is a lot of variability in the tensile adhesion test measurements. A larger sample size would have led to a shorter interval.

**FIGURE 8.7**

Normal probability plot of the load at failure data.

It is not as easy to select a sample size n to obtain a specified length (or precision of estimation) for this CI as it was in the known- σ case because the length of the interval involves s (which is unknown before the data are collected), n , and $t_{\alpha/2, n-1}$. Note that the t -percentile depends on the sample size n . Consequently, an appropriate n can only be obtained through trial and error. The results of this will, of course, also depend on the reliability of our prior “guess” for σ .

8.3

Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Sometimes confidence intervals on the population variance or standard deviation are needed. When the population is modeled by a normal distribution, the tests and intervals described in this section are applicable. The following result provides the basis of constructing these confidence intervals.

χ^2 Distribution

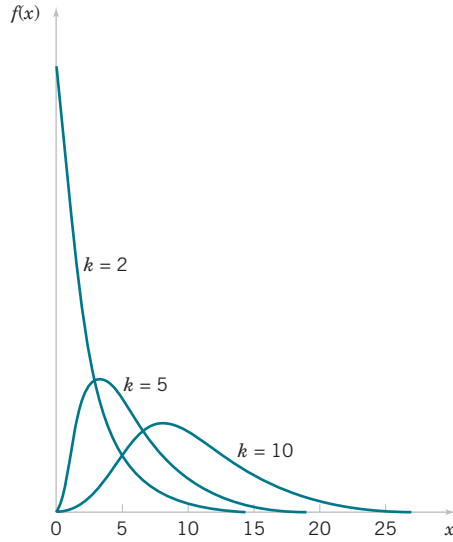
Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with mean μ and variance σ^2 , and let S^2 be the sample variance. Then the random variable

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \quad (8.17)$$

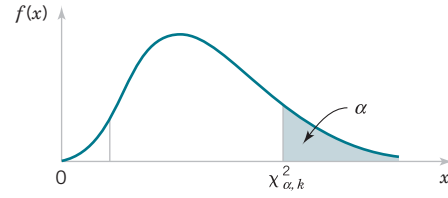
has a chi-square (χ^2) distribution with $n - 1$ degrees of freedom.

The probability density function of a χ^2 random variable is

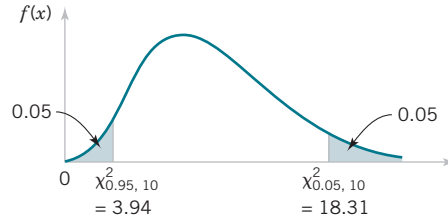
$$f(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad x > 0 \quad (8.18)$$

**FIGURE 8.8**

Probability density functions of several χ^2 distributions.



(a)



(b)

FIGURE 8.9

Percentage point of the χ^2 distribution. (a) The percentage point $\chi^2_{\alpha, k}$. (b) The upper percentage point $\chi^2_{0.05, 10} = 18.31$ and the lower percentage point $\chi^2_{0.95, 10} = 3.94$.

where k is the number of degrees of freedom. The mean and variance of the χ^2 distribution are k and $2k$, respectively. Several chi-square distributions are shown in Figure 8.8. Note that the chi-square random variable is non-negative and that the probability distribution is skewed to the right. However, as k increases, the distribution becomes more symmetric. As $k \rightarrow \infty$, the limiting form of the chi-square distribution is the normal distribution.

The *percentage points* of the χ^2 distribution are given in Table IV of the Appendix. Define $\chi^2_{\alpha, k}$ as the percentage point or value of the chi-square random variable with k degrees of freedom such that the probability that X^2 exceeds this value is α . That is,

$$P(X^2 > \chi^2_{\alpha, k}) = \int_{\chi^2_{\alpha, k}}^{\infty} f(u) du = \alpha$$

This probability is shown as the shaded area in Figure 8.9(a). To illustrate the use of Table IV, note that the areas α are the column headings and the degrees of freedom k are given in the left column. Therefore, the value with 10 degrees of freedom having an area (probability) of 0.05 to the right is $\chi^2_{0.05, 10} = 18.31$. This value is often called an *upper 5%* point of chi-square with 10 degrees of freedom. We may write this as a probability statement as follows:

$$P(X^2 > \chi^2_{0.05, 10}) = P(X^2 > 18.31) = 0.05$$

Conversely, a *lower 5%* point of chi-square with 10 degrees of freedom would be $\chi^2_{0.95, 10} = 3.94$ (from Appendix A). Both of these percentage points are shown in Figure 8.9(b).

The construction of the $100(1 - \alpha)\%$ CI for σ^2 is straightforward. Because

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

is chi-square with $n - 1$ degrees of freedom, we may write

$$P(\chi^2_{1-\alpha/2, n-1} \leq X^2 \leq \chi^2_{\alpha/2, n-1}) = 1 - \alpha$$

so that

$$P\left(\chi_{1-\alpha/2,n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2,n-1}^2\right) = 1 - \alpha$$

This last equation can be rearranged as

$$P\left(\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}^2}\right) = 1 - \alpha$$

This leads to the following definition of the confidence interval for σ^2 .

Confidence Interval on the Variance

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2 , then a $100(1 - \alpha)\%$ confidence interval on σ^2 is

$$\frac{(n-1)s^2}{\chi_{\alpha/2,n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2,n-1}^2} \quad (8.19)$$

where $\chi_{\alpha/2,n-1}^2$ and $\chi_{1-\alpha/2,n-1}^2$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. A confidence interval for σ has lower and upper limits that are the square roots of the corresponding limits in Equation 8.19.

It is also possible to find a $100(1 - \alpha)\%$ lower confidence bound or upper confidence bound on σ^2 .

One-Sided Confidence Bounds on the Variance

The $100(1 - \alpha)\%$ lower and upper confidence bounds on σ^2 are

$$\frac{(n-1)s^2}{\chi_{\alpha,n-1}^2} \leq \sigma^2 \quad \text{and} \quad \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha,n-1}^2} \quad (8.20)$$

respectively.

The CIs given in Equations 8.19 and 8.20 are less robust to the normality assumption. The distribution of $(n-1)S^2/\sigma^2$ can be very different from the chi-square if the underlying population is not normal.

EXAMPLE 8.6 | Detergent Filling

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153^2$ (fluid ounce). If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume

that the fill volume is approximately normally distributed. A 95% upper confidence bound is found from Equation 8.20 as follows:

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi_{0.95,19}^2}$$

or

$$\sigma^2 \leq \frac{(19)0.0153}{10.117} = 0.0287 \text{ (fluid ounce)}^2$$

This last expression may be converted into a confidence interval on the standard deviation σ by taking the square root of both sides, resulting in

$$\sigma \leq 0.17$$

Practical Interpretation: Therefore, at the 95% level of confidence, the data indicate that the process standard deviation could be as large as 0.17 fluid ounce. The process engineer or manager now needs to determine whether a standard deviation this large could lead to an operational problem with under- or over-filled bottles.

8.4 Large-Sample Confidence Interval for a Population Proportion

It is often necessary to construct confidence intervals on a population proportion. For example, suppose that a random sample of size n has been taken from a large (possibly infinite) population and that $X(\leq n)$ observations in this sample belong to a class of interest. Then $\hat{P} = X/n$ is a point estimator of the proportion of the population p that belongs to this class. Note that n and p are the parameters of a binomial distribution. Furthermore, from Chapter 4 we know that the sampling distribution of \hat{P} is approximately normal with mean p and variance $p(1-p)/n$, if p is not too close to either 0 or 1 and if n is relatively large. Typically, to apply this approximation we require that np and $n(1-p)$ be greater than or equal to 5. We use the normal approximation in this section.

Normal Approximation for a Binomial Proportion

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal.

To construct the confidence interval on p , note that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \simeq 1 - \alpha$$

so

$$P\left(-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) \simeq 1 - \alpha$$

This may be rearranged as

$$P\left(\hat{P} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}\right) \simeq 1 - \alpha \quad (8.21)$$

The quantity $\sqrt{p(1-p)/n}$ in Equation 8.21 is called the *standard error of the point estimator* \hat{P} . This was discussed in Chapter 7. Unfortunately, the upper and lower limits of the confidence interval obtained from Equation 8.21 contain the unknown parameter p . However, as suggested

at the end of Section 8.1.5, a solution that is often satisfactory is to replace p by \hat{P} in the standard error, which results in

$$P\left(\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \simeq 1 - \alpha \quad (8.22)$$

This leads to the approximate $100(1 - \alpha)\%$ confidence interval on p .

Approximate Confidence Interval on a Binomial Proportion

If \hat{p} is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate $100(1 - \alpha)\%$ confidence interval on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (8.23)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

This procedure depends on the adequacy of the normal approximation to the binomial. To be reasonably conservative, this requires that np and $n(1 - p)$ be greater than or equal to 5. In situations when this approximation is inappropriate, particularly in cases when n is small, other methods must be used. Tables of the binomial distribution could be used to obtain a confidence interval for p . However, we could also use numerical methods that are implemented on the binomial probability mass function in some computer programs.

EXAMPLE 8.7 | Crankshaft Bearings

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Therefore, a point estimate of the proportion of bearings in the population that exceeds the roughness specification is $\hat{p} = x/n = 10/85 = 0.12$. A 95% two-sided confidence interval for p is computed from Equation 8.23 as

$$\hat{p} - z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

or

$$0.12 - 1.96 \sqrt{\frac{0.12(0.88)}{85}} \leq p \leq 0.12 + 1.96 \sqrt{\frac{0.12(0.88)}{85}}$$

which simplifies to

$$0.0509 \leq p \leq 0.2243$$

Practical Interpretation: This is a wide CI. Although the sample size does not appear to be small ($n = 85$), the value of \hat{p} is fairly small, which leads to a large standard error for \hat{p} contributing to the wide CI.

Choice of Sample Size Because \hat{P} is the point estimator of p , we can define the error in estimating p by \hat{P} as $E = |p - \hat{P}|$. Note that we are approximately $100(1 - \alpha)\%$ confident that this error is less than $z_{\alpha/2} \sqrt{p(1-p)/n}$. For instance, in Example 8.7, we are 95% confident that the sample proportion $\hat{p} = 0.12$ differs from the true proportion p by an amount not exceeding 0.07.

In situations when the sample size can be selected, we may choose n to be $100(1 - \alpha)\%$ confident that the error is less than some specified value E . If we set $E = z_{\alpha/2} \sqrt{p(1-p)/n}$ and solve for n , the appropriate sample size is

Sample Size for a Specified Error on a Binomial Proportion

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 p(1-p) \quad (8.24)$$

An estimate of p is required to use Equation 8.24. If an estimate \hat{p} from a previous sample is available, it can be substituted for p in Equation 8.24, or perhaps a subjective estimate can be made. If these alternatives are unsatisfactory, a preliminary sample can be taken, \hat{p} computed, and then Equation 8.24 used to determine how many additional observations are required to estimate p with the desired accuracy. Another approach to choosing n uses the fact that the sample size from Equation 8.24 will always be a maximum for $p = 0.5$ [that is, $p(1-p) \leq 0.25$ with equality for $p = 0.5$], and this can be used to obtain an upper bound on n . In other words, we are at least $100(1-\alpha)\%$ confident that the error in estimating p by \hat{P} is less than E if the sample size is

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 (0.25) \quad (8.25)$$

EXAMPLE 8.8 | Crankshaft Bearings

Consider the situation in Example 8.7. How large a sample is required if we want to be 95% confident that the error in using \hat{p} to estimate p is less than 0.05? Using $\hat{p} = 0.12$ as an initial estimate of p , we find from Equation 8.24 that the required sample size is

$$n = \left(\frac{z_{0.025}}{E} \right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.96}{0.05} \right)^2 0.12(0.88) \cong 163$$

If we wanted to be *at least* 95% confident that our estimate \hat{p} of the true proportion p was within 0.05 regardless of the value of p , we would use Equation 8.25 to find the sample size

$$n = \left(\frac{z_{0.025}}{E} \right)^2 (0.25) = \left(\frac{1.96}{0.05} \right)^2 (0.25) \cong 385$$

Practical Interpretation: Notice that if we have information concerning the value of p , either from a preliminary sample or from past experience, we could use a smaller sample while maintaining both the desired precision of estimation and the level of confidence.

One-Sided Confidence Bounds We may find approximate one-sided confidence bounds on p by using a simple modification of Equation 8.23.

Approximate One-Sided Confidence Bounds on a Binomial Proportion

The approximate $100(1-\alpha)\%$ lower and upper confidence bounds are

$$\hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \quad \text{and} \quad p \leq \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (8.26)$$

respectively.

A Different Confidence Interval on the Binomial Proportion There is a different way to construct a CI on a binomial proportion than the traditional approach in Equation 8.23. Starting with Equation 8.22 and replacing the inequalities with an equality and solving the resulting quadratic equation for p results in

$$p = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n}$$

This implies that a two-sided CI on a proportion p is as follows:

$$\begin{aligned} UCL &= \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n} \\ LCL &= \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n} \end{aligned} \quad (8.27)$$

The article by Agresti and Coull in *The American Statistician* (“Approximate Better Than ‘Exact’ for Interval Estimation of a Binomial Proportion,” 1998, Vol. 52, pp. 119–126) reports that the actual confidence level for the CI in Equation 8.27 is closer to the “advertised” or nominal level for almost all values of α and p than for the traditional CI in Equation 8.23. The authors also report that this new interval can be used with nearly all sample sizes. So the requirements that $n\hat{p} \geq 5$ or 10 or $n(1-\hat{p}) \geq 5$ or 10 are not too important. If the sample size is large, the quantity $z_{\alpha/2}^2/(2n)$ will be small relative to \hat{p} , $z_{\alpha/2}^2/(4n^2)$ will be small relative to $\hat{p}(1-\hat{p})/n$, and $z_{\alpha/2}^2/n$ will be small, so as a result the **Agresti-Coull CI** in Equation 8.27 will reduce to the traditional CI given in Equation 8.23.

8.5 Guidelines for Constructing Confidence Intervals

The most difficult step in constructing a confidence interval is often the match of the appropriate calculation to the objective of the study. Common cases are listed in Table 8.1 along with the reference to the section that covers the appropriate calculation for a confidence interval test. Table 8.1 provides a simple road map to help select the appropriate analysis. Two primary comments can help identify the analysis:

1. Determine the parameter (and the distribution of the data) that will be bounded by the confidence interval or tested by the hypothesis.
2. Check if other parameters are known or need to be estimated.

In Chapter 9, we study a procedure closely related to confidence intervals called *hypothesis testing*. Table 8.1 can be used for those procedures also. This road map is extended to more cases in Chapter 10.

TABLE 8.1 The Roadmap for Constructing Confidence Intervals and Performing Hypothesis Tests, One-Sample Case

Parameter to Be Bounded by the Confidence Interval or Tested with a Hypothesis?	Symbol	Other Parameters?	Confidence Interval Section	Hypothesis Test Section	Comments
Mean of normal distribution	μ	Standard deviation σ known	8.1	9.2	Large sample size is often taken to be $n \geq 40$
Mean of arbitrary distribution with large sample size	μ	Sample size large enough that central limit theorem applies and σ is essentially known	8.1.5	9.2.3	
Mean of normal distribution	μ	Standard deviation σ unknown and estimated	8.2	9.3	
Variance (or standard deviation) of normal distribution	σ^2	Mean μ unknown and estimated	8.3	9.4	
Population proportion	p	None	8.4	9.5	

8.6 Bootstrap Confidence Interval

In Section 7.3.4, we saw how a computer-intensive technique called the *bootstrap* could be used to find the estimated standard error of a statistic, say $\hat{\theta}$. The bootstrap technique can also be used to find confidence intervals. These techniques can be useful in situations in which a “standard” CI is not readily available. To illustrate the general approach, let’s consider a case for which there is a standard CI, the $100(1 - \alpha)\%$ CI on the mean of a normal distribution with known variance. Here the parameter of interest is the population mean μ , and the statistic that estimates μ is the sample average \bar{X} . The quantity $z_{\alpha/2}\sigma/\sqrt{n}$ is the $100(1 - \alpha/2)$ percentile of the distribution of $\hat{\theta}_i^B - \bar{\theta}^B$, $i = 1, 2, \dots, n_B$ and by the same logic, the quantity $-z_{\alpha/2}\sigma/\sqrt{n}$ is the $100(\alpha/2)$ percentile of the distribution of $\bar{X} - \mu$. Therefore, the $100(1 - \alpha/2)\%$ CI can be written as:

$$P(\alpha/2\text{th percentile} \leq \bar{X} - \mu \leq (1 - \alpha/2)\text{th percentile}) = 1 - \alpha/2$$

This can be rearranged as

$$P(\bar{X} - (1 - \alpha/2)\text{th percentile} \leq \mu \leq \bar{X} + \alpha/2\text{th percentile}) = 1 - \alpha/2$$

So the lower confidence bound is $\bar{X} - (1 - \alpha/2)\text{th percentile}$ of the distribution of $\bar{X} - \mu$ and the upper confidence bound is $\bar{X} + \alpha/2\text{th percentile}$ of the distribution of $\bar{X} - \mu$. When these percentiles cannot be easily determined for some arbitrary parameter θ , they can often be estimated by using bootstrap samples. The procedure would consist of taking n_B bootstrap samples, calculating the bootstrap estimates $\hat{\theta}_1^B, \hat{\theta}_2^B, \dots, \hat{\theta}_{n_B}^B$ and $\bar{\theta}^B$, and then computing the differences $\hat{\theta}_i^B - \bar{\theta}^B$, $i = 1, 2, \dots, n_B$. The $\alpha/2$ smallest and largest of these differences are the estimates of the percentiles required to construct the bootstrap CI.

8.7 Tolerance and Prediction Intervals

8.7.1 Prediction Interval for a Future Observation

In some problem situations, we may be interested in predicting a future observation of a variable. This is a different problem than estimating the mean of that variable, so a confidence interval is

not appropriate. In this section, we show how to obtain a $100(1 - \alpha)\%$ **prediction interval** on a future value of a normal random variable.

Suppose that X_1, X_2, \dots, X_n is a random sample from a normal population. We wish to predict the value X_{n+1} , a single **future** observation. A point prediction of X_{n+1} is \bar{X} , the sample mean. The prediction error is $X_{n+1} - \bar{X}$. The expected value of the prediction error is

$$E(X_{n+1} - \bar{X}) = \mu - \mu = 0$$

and the variance of the prediction error is

$$V(X_{n+1} - \bar{X}) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2 \left(1 + \frac{1}{n}\right)$$

because the future observation X_{n+1} is independent of the mean of the current sample \bar{X} . The prediction error $X_{n+1} - \bar{X}$ is normally distributed. Therefore,

$$Z = \frac{X_{n+1} - \bar{X}}{\sigma \sqrt{1 + \frac{1}{n}}}$$

has a standard normal distribution. Replacing σ with S results in

$$T = \frac{X_{n+1} - \bar{X}}{S \sqrt{1 + \frac{1}{n}}}$$

which has a t distribution with $n - 1$ degrees of freedom. Manipulating T as we have done previously in the development of a CI leads to a prediction interval on the future observation X_{n+1} .

Prediction Interval

A $100(1 - \alpha)\%$ **prediction interval (PI)** on a single future observation from a normal distribution is given by

$$\bar{x} - t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \quad (8.28)$$

The prediction interval for X_{n+1} will always be longer than the confidence interval for μ because more variability is associated with the prediction error than with the error of estimation. This is easy to see because the prediction error is the difference between two random variables ($X_{n+1} - \bar{X}$), and the estimation error in the CI is the difference between one random variable and a constant ($\bar{X} - \mu$). As n gets larger ($n \rightarrow \infty$), the length of the CI decreases to zero, essentially becoming the single value μ , but the length of the PI approaches $2z_{\alpha/2}\sigma$. So as n increases, the uncertainty in estimating μ goes to zero, although there will always be uncertainty about the future value X_{n+1} , even when estimating any of the distribution parameters is not necessary.

We noted in Section 8.2 that the t distribution-based CI for μ was robust to the normality assumption when n is small. The practical implication of this is that although we have computed a 95% CI, the actual confidence level will not be exactly 95%, but it will be very close—maybe 93% or 94%. Prediction intervals, on the other hand, are very sensitive to the normality assumption, and Equation 8.28 should not be used unless we are very comfortable with the normality assumption.

EXAMPLE 8.9 | Alloy Adhesion

Reconsider the tensile adhesion tests on specimens of U-700 alloy described in Example 8.5. The load at failure for $n = 22$ specimens was observed, and we found that $\bar{x} = 13.71$ and $s = 3.55$. The 95% confidence interval on μ was $12.14 \leq \mu \leq 15.28$. We plan to test a 23rd specimen. A 95% prediction interval on the load at failure for this specimen is

$$\bar{x} - t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}}$$

$$\begin{aligned} 13.71 - (2.080)3.55 \sqrt{1 + \frac{1}{22}} &\leq X_{23} \leq 13.71 \\ &\quad + (2.080)3.55 \sqrt{1 + \frac{1}{22}} \\ 6.16 &\leq X_{23} \leq 21.26 \end{aligned}$$

Practical Interpretation: Notice that the prediction interval is considerably longer than the CI. This is because the CI is an estimate of a parameter, but the PI is an interval estimate of a single future observation.

8.7.2 Tolerance Interval for a Normal Distribution

Consider a population of semiconductor processors. Suppose that the speed of these processors has a normal distribution with mean $\mu = 600$ megahertz and standard deviation $\sigma = 30$ megahertz. Then the interval from $600 - 1.96(30) = 541.2$ to $600 + 1.96(30) = 658.8$ megahertz captures the speed of 95% of the processors in this population because the interval from -1.96 to 1.96 captures 95% of the area under the standard normal curve. The interval from $\mu - z_{\alpha/2}\sigma$ to $\mu + z_{\alpha/2}\sigma$ is called a **tolerance interval**.

If μ and σ are unknown, we can use the data from a random sample of size n to compute \bar{x} and s and then form the interval $(\bar{x} - 1.96s, \bar{x} + 1.96s)$. However, because of sampling variability in \bar{x} and s , it is likely that this interval will contain less than 95% of the values in the population. The solution to this problem is to replace 1.96 with some value that will make the proportion of the distribution contained in the interval 95% with some level of confidence. Fortunately, it is easy to do this.

Tolerance Interval

A **tolerance interval** for capturing at least $\gamma\%$ of the values in a normal distribution with confidence level $100(1 - \alpha)\%$ is

$$\bar{x} - ks, \quad \bar{x} + ks$$

where k is a tolerance interval factor found in Appendix Table XII. Values are given for $\gamma = 90\%$, 95% , and 99% , and for 90% , 95% , and 99% confidence.

This interval is very sensitive to the normality assumption. One-sided tolerance bounds can also be computed. The tolerance factors for these bounds are also given in Appendix Table XII.

EXAMPLE 8.10 | Alloy Adhesion

Let's reconsider the tensile adhesion tests originally described in Example 8.5. The load at failure for $n = 22$ specimens was observed, and we found that $\bar{x} = 13.71$ and $s = 3.55$. We want to find a tolerance interval for the load at failure that includes 90% of the values in the population with 95% confidence. From Appendix Table XII, the tolerance factor k for $n = 22$, $\gamma = 0.90$, and 95% confidence is $k = 2.264$. The desired tolerance interval is

$$(\bar{x} - ks, \bar{x} + ks)$$

or

$$[13.71 - (2.264)3.55, 13.71 + (2.264)3.55]$$

which reduces to (5.67, 21.74).

Practical Interpretation: We can be 95% confident that at least 90% of the values of load at failure for this particular alloy lie between 5.67 and 21.74 megapascals.

From Appendix Table XII, we note that as $n \rightarrow \infty$, the value of k goes to the z -value associated with the desired level of containment for the normal distribution. For example, if we want 90% of the population to fall in the two-sided tolerance interval, k approaches $z_{0.05} = 1.645$ as $n \rightarrow \infty$. Note that as $n \rightarrow \infty$, a $100(1 - \alpha)\%$ prediction interval on a future value approaches a tolerance interval that contains $100(1 - \alpha)\%$ of the distribution.

Important Terms and Concepts

Agresti-Coull confidence interval on a population proportion	Confidence interval on the variance of a normal distribution	One-sided confidence bound
Chi-square distribution	Confidence interval on the mean of a normal distribution	Prediction interval
Confidence coefficient	Confidence level	t distribution
Confidence interval	Error in estimation	Tolerance interval
Confidence interval on a population proportion	Large-sample confidence interval	Two-sided confidence interval

Tests of Hypotheses for a Single Sample

Alain Nogues/Sygma/Sygma/Getty Images



LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Structure engineering decision-making problems as hypothesis tests
 2. Test hypotheses on the mean of a normal distribution using either a Z -test or a t -test procedure
 3. Test hypotheses on the variance or standard deviation of a normal distribution
 4. Test hypotheses on a population proportion
 5. Use the P -value approach for making decisions in hypothesis tests
 6. Compute power and type II error probability, and make sample size selection decisions for tests on means, variances, and proportions
 7. Explain and use the relationship between confidence intervals and hypothesis tests
 8. Use the chi-square goodness-of-fit test to check distributional assumptions
 9. Apply contingency table tests
 10. Apply nonparametric tests
 11. Use equivalence testing
 12. Combine P -values
-

CHAPTER OUTLINE

9.1 Hypothesis Testing

- 9.1.1 Statistical Hypotheses
- 9.1.2 Tests of Statistical Hypotheses
- 9.1.3 One-Sided and Two-Sided Hypotheses
- 9.1.4 *P*-Values in Hypothesis Tests
- 9.1.5 Connection between Hypothesis Tests and Confidence Intervals
- 9.1.6 General Procedure for Hypothesis Tests

9.2 Tests on the Mean of a Normal Distribution, Variance Known

- 9.2.1 Hypothesis Tests on the Mean
- 9.2.2 Type II Error and Choice of Sample Size
- 9.2.3 Large-Sample Test

9.3 Tests on the Mean of a Normal Distribution, Variance Unknown

- 9.3.1 Hypothesis Tests on the Mean
- 9.3.2 Type II Error and Choice of Sample Size

9.4 Tests on the Variance and Standard Deviation of a Normal Distribution

- 9.4.1 Hypothesis Tests on the Variance
- 9.4.2 Type II Error and Choice of Sample Size

9.5 Tests on a Population Proportion

- 9.5.1 Large-Sample Tests on a Proportion
- 9.5.2 Type II Error and Choice of Sample Size

9.6 Summary Table of Inference Procedures for a Single Sample**9.7 Testing for Goodness of Fit****9.8 Contingency Table Tests****9.9 Nonparametric Procedures**

- 9.9.1 The Sign Test
- 9.9.2 The Wilcoxon Signed-Rank Test
- 9.9.3 Comparison to the *t*-Test

9.10 Equivalence Testing**9.11 Combining *P*-Values**

Introduction

In the previous two chapters, we showed how a parameter of a population can be estimated from sample data, using either a *point estimate* (Chapter 7) or an interval of likely values called a *confidence interval* (Chapter 8). In many situations, a different type of problem is of interest; there are two competing claims about the value of a parameter, and the engineer must determine which claim is correct. For example, suppose that an engineer is designing an air crew escape system that consists of an ejection seat and a rocket motor that powers the seat. The rocket motor contains a propellant, and for the ejection seat to function properly, the propellant should have a mean burning rate of 50 cm/sec. If the burning rate is too low, the ejection seat may not function properly, leading to an unsafe ejection and possible injury of the pilot. Higher burning rates may imply instability in the propellant or an ejection seat that is too powerful, again leading to possible pilot injury. So the practical engineering question that must be answered is: Does the mean burning rate of the propellant equal 50 cm/sec, or is it some other value (either higher or lower)? This type of question can be answered using a statistical technique called **hypothesis testing**. This chapter focuses on the basic principles of hypothesis testing and provides techniques for solving the most common types of hypothesis testing problems involving a single sample of data.

9.1 Hypothesis Testing

9.1.1 Statistical Hypotheses

In the previous chapter, we illustrated how to construct a confidence interval estimate of a parameter from sample data. However, many problems in engineering require that we decide which of two competing claims or statements about some parameter is true. The statements are called

hypotheses, and the decision-making procedure is called **hypothesis testing**. This is one of the most useful aspects of statistical inference, because many types of decision-making problems, tests, or experiments in the engineering world can be formulated as hypothesis-testing problems. Furthermore, as we will see, a very close connection exists between hypothesis testing and confidence intervals.

Statistical hypothesis testing and confidence interval estimation of parameters are the fundamental methods used at the data analysis stage of a *comparative experiment* in which the engineer is interested, for example, in comparing the mean of a population to a specified value. These simple comparative experiments are frequently encountered in practice and provide a good foundation for the more complex experimental design problems that we discuss in Chapters 13 and 14. In this chapter, we discuss comparative experiments involving a single population, and our focus is on testing hypotheses concerning the parameters of the population.

We now give a formal definition of a statistical hypothesis.

Statistical Hypothesis

A **statistical hypothesis** is a statement about the parameters of one or more populations.

Because we use probability distributions to represent populations, a statistical hypothesis may also be thought of as a statement about the probability distribution of a random variable. The hypothesis will usually involve one or more parameters of this distribution.

For example, consider the air crew escape system described in the introduction. Suppose that we are interested in the burning rate of the solid propellant. Burning rate is a random variable that can be described by a probability distribution. Suppose that our interest focuses on the mean burning rate (a parameter of this distribution). Specifically, we are interested in deciding whether or not the mean burning rate is 50 centimeters per second. We may express this formally as

$$H_0: \mu = 50 \text{ centimeters per second} \quad H_1: \mu \neq 50 \text{ centimeters per second} \quad (9.1)$$

The statement $H_0: \mu = 50$ centimeters per second in Equation 9.1 is called the **null hypothesis**. This is a claim that is initially assumed to be true. The statement $H_1: \mu \neq 50$ centimeters per second is called the **alternative hypothesis** and it is a statement that contradicts the null hypothesis. Because the alternative hypothesis specifies values of μ that could be either greater or less than 50 centimeters per second, it is called a **two-sided alternative hypothesis**. In some situations, we may wish to formulate a **one-sided alternative hypothesis**, as in

$$\begin{aligned} H_0: \mu = 50 \text{ centimeters per second} \quad H_1: \mu < 50 \text{ centimeters per second} \\ \text{or} \\ H_0: \mu = 50 \text{ centimeters per second} \quad H_1: \mu > 50 \text{ centimeters per second} \end{aligned} \quad (9.2)$$

We will always state the null hypothesis as an equality claim. However, when the alternative hypothesis is stated with the $<$ sign, the implicit claim in the null hypothesis can be taken as \geq and when the alternative hypothesis is stated with the $>$ sign, the implicit claim in the null hypothesis can be taken as \leq .

It is important to remember that hypotheses are always statements about the population or distribution under study, not statements about the sample. The value of the population parameter specified in the null hypothesis (50 centimeters per second in the preceding example) is usually determined in one of three ways. First, it may result from past experience or knowledge of the process or even from previous tests or experiments. The objective of hypothesis testing, then, is usually to determine whether the parameter value has changed. Second, this value may be determined from some theory or model regarding the process under study. Here the objective

of hypothesis testing is to verify the theory or model. A third situation arises when the value of the population parameter results from external considerations, such as design or engineering specifications, or from contractual obligations. In this situation, the usual objective of hypothesis testing is conformance testing.

A procedure leading to a decision about the null hypothesis is called a **test of a hypothesis**. Hypothesis-testing procedures rely on using the information in a random sample from the population of interest. If this information is consistent with the null hypothesis, we will not reject it; however, if this information is inconsistent with the null hypothesis, we will conclude that the null hypothesis is false and reject it in favor of the alternative. We emphasize that the truth or falsity of a particular hypothesis can never be known with certainty unless we can examine the entire population. This is usually impossible in most practical situations. Therefore, a hypothesis-testing procedure should be developed with the probability of reaching a wrong conclusion in mind. Testing the hypothesis involves taking a random sample, computing a **test statistic** from the sample data, and then using the test statistic to make a decision about the null hypothesis.

9.1.2 Tests of Statistical Hypotheses

To illustrate the general concepts, consider the propellant burning rate problem introduced earlier. The null hypothesis is that the mean burning rate is 50 centimeters per second, and the alternate is that it is not equal to 50 centimeters per second. That is, we wish to test

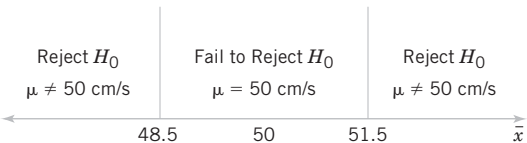
$$\begin{aligned} H_0: \mu &= 50 \text{ centimeters per second} \\ H_1: \mu &\neq 50 \text{ centimeters per second} \end{aligned}$$

Suppose that a sample of $n = 10$ specimens is tested and that the sample mean burning rate \bar{x} is observed. The sample mean is an estimate of the true population mean μ . A value of the sample mean \bar{x} that falls close to the hypothesized value of $\mu = 50$ centimeters per second does not conflict with the null hypothesis that the true mean μ is really 50 centimeters per second. On the other hand, a sample mean that is considerably different from 50 centimeters per second is evidence in support of the alternative hypothesis H_1 . Thus, the sample mean is the test statistic in this case.

The sample mean can take on many different values. Suppose that if $48.5 \leq \bar{x} \leq 51.5$, we will not reject the null hypothesis $H_0: \mu = 50$, and if either $\bar{x} < 48.5$ or $\bar{x} > 51.5$, we will reject the null hypothesis in favor of the alternative hypothesis $H_1: \mu \neq 50$. This is illustrated in Figure 9.1. The values of \bar{x} that are less than 48.5 and greater than 51.5 constitute the **critical region** for the test; all values that are in the interval $48.5 \leq \bar{x} \leq 51.5$ form a region for which we will fail to reject the null hypothesis. By convention, this is usually called the **acceptance region**. The boundaries between the critical regions and the acceptance region are called the **critical values**. In our example, the critical values are 48.5 and 51.5. It is customary to state conclusions relative to the null hypothesis H_0 . Therefore, we reject H_0 in favor of H_1 if the test statistic falls in the critical region and fails to reject H_0 otherwise.

This decision procedure can lead to either of two wrong conclusions. For example, the true mean burning rate of the propellant could be equal to 50 centimeters per second. However, for the randomly selected propellant specimens that are tested, we could observe a value of the test statistic \bar{x} that falls into the critical region. We would then reject the null hypothesis H_0 in favor of the alternate H_1 when, in fact, H_0 is really true. This type of wrong conclusion is called a **type I error**.

FIGURE 9.1
Decision criteria for testing $H_0: \mu = 50$ centimeters per second versus $H_1: \mu \neq 50$ centimeters per second.



Type I Error

Rejecting the null hypothesis H_0 when it is true is defined as a **type I error**.

Now suppose that the true mean burning rate is different from 50 centimeters per second, yet the sample mean \bar{x} falls in the acceptance region. In this case, we would fail to reject H_0 when it is false, and this leads to the other type of error.

Type II Error

Failing to reject the null hypothesis when it is false is defined as a **type II error**.

Thus, in testing any statistical hypothesis, four different situations determine whether the final decision is correct or in error. These situations are presented in Table 9.1.

Because our decision is based on random variables, probabilities can be associated with the type I and type II errors in Table 9.1. The probability of making a type I error is denoted by the Greek letter α .

Probability of Type I Error

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \quad (9.3)$$

Sometimes the type I error probability is called the **significance level**, the **α -error**, or the **size of the test**. In the propellant burning rate example, a type I error will occur when either $\bar{x} > 51.5$ or $\bar{x} < 48.5$ when the true mean burning rate really is $\mu = 50$ centimeters per second. Suppose that the standard deviation of the burning rate is $\sigma = 2.5$ centimeters per second and that the burning rate has a distribution for which the conditions of the central limit theorem apply, so the distribution of the sample mean is approximately normal with mean $\mu = 50$ and standard deviation $\sigma/\sqrt{n} = 2.5/\sqrt{10} = 0.79$. The probability of making a type I error (or the significance level of our test) is equal to the sum of the areas that have been shaded in the tails of the normal distribution in Figure 9.2. We may find this probability as

$$\alpha = P(\bar{X} < 48.5 \text{ when } \mu = 50) + P(\bar{X} > 51.5 \text{ when } \mu = 50)$$

The z -values that correspond to the critical values 48.5 and 51.5 are

$$z_1 = \frac{48.5 - 50}{0.79} = -1.90 \quad \text{and} \quad z_2 = \frac{51.5 - 50}{0.79} = 1.90$$

Therefore,

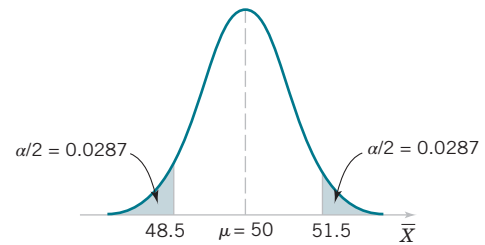
$$\alpha = P(z < -1.90) + P(z > 1.90) = 0.0287 + 0.0287 = 0.0574$$

TABLE 9.1 Decisions in Hypothesis Testing

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

FIGURE 9.2

The critical region for $H_0: \mu = 50$ versus $H_1: \mu \neq 50$ and $n = 10$.



This is the type I error probability. This implies that 5.74% of all random samples would lead to rejection of the hypothesis $H_0: \mu = 50$ centimeters per second when the true mean burning rate is really 50 centimeters per second.

From an inspection of Figure 9.2, notice that we can reduce α by widening the acceptance region. For example, if we make the critical values 48 and 52, the value of α is

$$\begin{aligned}\alpha &= P\left(z < -\frac{48 - 50}{0.79}\right) + P\left(z > \frac{52 - 50}{0.79}\right) = P(z < -2.53) + P(z > 2.53) \\ &= 0.0057 + 0.0057 = 0.0114\end{aligned}$$

We could also reduce α by increasing the sample size. If $n = 16$, $\sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$ and using the original critical region from Figure 9.1, we find

$$z_1 = \frac{48.5 - 50}{0.625} = -2.40 \quad \text{and} \quad z_2 = \frac{51.5 - 50}{0.625} = 2.40$$

Therefore,

$$\alpha = P(Z < -2.40) + P(Z > 2.40) = 0.0082 + 0.0082 = 0.0164$$

In evaluating a hypothesis-testing procedure, it is also important to examine the probability of a type II error, which we denote by β . That is,

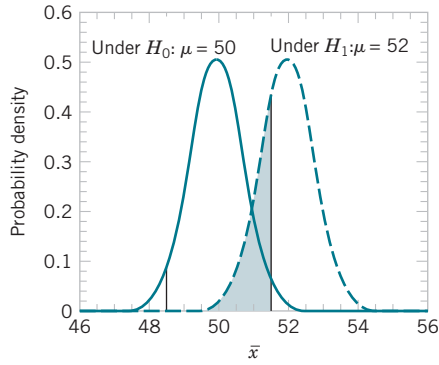
Probability of Type II Error

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) \quad (9.4)$$

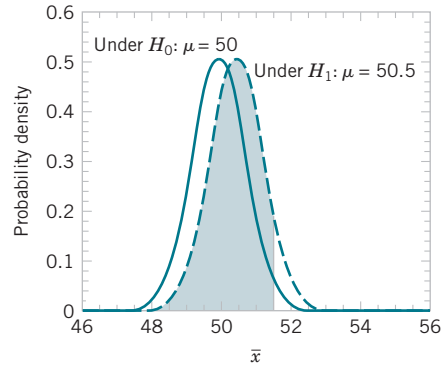
To calculate β (sometimes called the **β -error**), we must have a specific alternative hypothesis; that is, we must have a particular value of μ . For example, suppose that it is important to reject the null hypothesis $H_0: \mu = 50$ whenever the mean burning rate μ is greater than 52 centimeters per second or less than 48 centimeters per second. We could calculate the probability of a type II error β for the values $\mu = 52$ and $\mu = 48$ and use this result to tell us something about how the test procedure would perform. Specifically, how will the test procedure work if we wish to detect, that is, reject H_0 , for a mean value of $\mu = 52$ or $\mu = 48$? Because of symmetry, it is necessary to evaluate only one of the two cases—say, find the probability of failing to reject the null hypothesis $H_0: \mu = 50$ centimeters per second when the true mean is $\mu = 52$ centimeters per second.

Figure 9.3 helps us calculate the probability of type II error β . The normal distribution on the left in Figure 9.3 is the distribution of the test statistic \bar{X} when the null hypothesis $H_0: \mu = 50$ is true (this is what is meant by the expression “under $H_0: \mu = 50$ ”), and the normal distribution on the right is the distribution of \bar{X} when the alternative hypothesis is true and the value of the mean is 52 (or “under $H_1: \mu = 52$ ”). A type II error will be committed if the sample mean \bar{X} falls between 48.5 and 51.5 (the critical region boundaries) when $\mu = 52$. As seen in Figure 9.3, this is just the probability that $48.5 \leq \bar{X} \leq 51.5$ when the true mean is $\mu = 52$, or the shaded area under the normal distribution centered at $\mu = 52$. Therefore, referring to Figure 9.3, we find that

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ when } \mu = 52)$$

**FIGURE 9.3**

The probability of type II error when $\mu = 52$ and $n = 10$.

**FIGURE 9.4**

The probability of type II error when $\mu = 50.5$ and $n = 10$.

The z -values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$z_1 = \frac{48.5 - 52}{0.79} = -4.43 \quad \text{and} \quad z_2 = \frac{51.5 - 52}{0.79} = -0.63$$

Therefore,

$$\begin{aligned} \beta &= P(-4.43 \leq Z \leq -0.63) = P(Z \leq -0.63) - P(Z \leq -4.43) \\ &= 0.2643 - 0.0000 = 0.2643 \end{aligned}$$

Thus, if we are testing $H_0: \mu = 50$ against $H_1: \mu \neq 50$ with $n = 10$ and the true value of the mean is $\mu = 52$, the probability that we will fail to reject the false null hypothesis is 0.2643. By symmetry, if the true value of the mean is $\mu = 48$, the value of β will also be 0.2643.

The probability of making a type II error β increases rapidly as the true value of μ approaches the hypothesized value. For example, see Figure 9.4, where the true value of the mean is $\mu = 50.5$ and the hypothesized value is $H_0: \mu = 50$. The true value of μ is very close to 50, and the value for β is

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ when } \mu = 50.5)$$

As shown in Figure 9.4, the z -values corresponding to 48.5 and 51.5 when $\mu = 50.5$ are

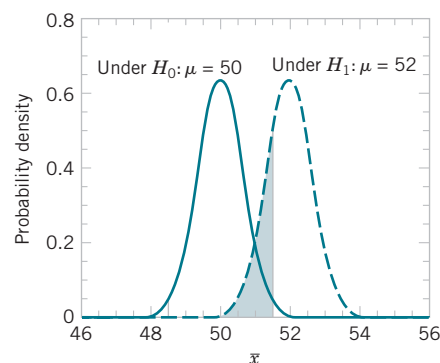
$$z_1 = \frac{48.5 - 50.5}{0.79} = -2.53 \quad \text{and} \quad z_2 = \frac{51.5 - 50.5}{0.79} = 1.27$$

Therefore,

$$\begin{aligned} \beta &= P(-2.53 \leq Z \leq 1.27) = P(Z \leq 1.27) - P(Z \leq -2.53) \\ &= 0.8980 - 0.0057 = 0.8923 \end{aligned}$$

Thus, the type II error probability is much higher for the case in which the true mean is 50.5 centimeters per second than for the case in which the mean is 52 centimeters per second. Of course, in many practical situations, we would not be as concerned with making a type II error if the mean were “close” to the hypothesized value. We would be much more interested in detecting large differences between the true mean and the value specified in the null hypothesis.

The type II error probability also depends on the sample size n . Suppose that the null hypothesis is $H_0: \mu = 50$ centimeters per second and that the true value of the mean is $\mu = 52$. If the sample size is increased from $n = 10$ to $n = 16$, the situation of Figure 9.5 results. The normal distribution on the left is the distribution of \bar{X} when the mean $\mu = 50$, and the normal distribution

**FIGURE 9.5**

The probability of type II error when $\mu = 52$ and $n = 16$.

on the right is the distribution of \bar{X} when $\mu = 52$. As shown in Figure 9.5, the type II error probability is

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ when } \mu = 52)$$

When $n = 16$, the standard deviation of \bar{X} is $\sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$, and the z -values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$z_1 = \frac{48.5 - 52}{0.625} = -5.60 \quad \text{and} \quad z_2 = \frac{51.5 - 52}{0.625} = -0.80$$

Therefore,

$$\begin{aligned} \beta &= P(-5.60 \leq Z \leq -0.80) = P(Z \leq -0.80) - P(Z \leq -5.60) \\ &= 0.2119 - 0.0000 = 0.2119 \end{aligned}$$

Recall that when $n = 10$ and $\mu = 52$, we found that $\beta = 0.2643$; therefore, increasing the sample size results in a decrease in the probability of type II error.

The results from this section and a few other similar calculations are summarized in the following table. The critical values are adjusted to maintain equal α for $n = 10$ and $n = 16$. This type of calculation is discussed later in the chapter.

Acceptance Region	Sample Size	α	β at $\mu = 52$	β at $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.81 < \bar{x} < 51.19$	16	0.0576	0.0966	0.8606
$48.42 < \bar{x} < 51.58$	16	0.0114	0.2515	0.9578

The results in the boxes were not calculated in the text but the reader can easily verify them. This display and the discussion above reveal four important points:

1. The size of the critical region, and consequently the probability of a type I error α , can always be reduced by appropriate selection of the critical values.
2. Type I and type II errors are related. A decrease in the probability of one type of error always results in an increase in the probability of the other provided that the sample size n does not change.
3. An increase in sample size reduces β provided that α is held constant.
4. When the null hypothesis is false, β increases as the true value of the parameter approaches the value hypothesized in the null hypothesis. The value of β decreases as the difference between the true mean and the hypothesized value increases.

Generally, the analyst controls the type I error probability α when he or she selects the critical values. Thus, it is usually easy for the analyst to set the type I error probability at (or near) any desired value. Because the analyst can directly control the probability of wrongly rejecting H_0 , we always think of rejection of the null hypothesis H_0 as a *strong conclusion*.

Because we can control the probability of making a type I error (or significance level), a logical question is what value should be used. The type I error probability is a measure of risk, specifically, the risk of concluding that the null hypothesis is false when it really is not. So, the value of α should be chosen to reflect the consequences (economic, social, etc.) of incorrectly rejecting the null hypothesis. Smaller values of α would reflect more serious consequences and larger values of α would be consistent with less severe consequences. This is often hard to do, so what has evolved in much of scientific and engineering practice is to use the value $\alpha = 0.05$ in most situations unless information is available that this is an inappropriate choice. In the rocket propellant problem with $n = 10$, this would correspond to critical values of 48.45 and 51.55.

A widely used procedure in hypothesis testing is to use a type I error or significance level of $\alpha = 0.05$. This value has evolved through experience and may not be appropriate for all situations.

On the other hand, the probability of type II error β is not a constant but depends on the true value of the parameter. It also depends on the sample size that we have selected. Because the type II error probability β is a function of both the sample size and the extent to which the null hypothesis H_0 is false, it is customary to think of the decision to accept H_0 as a *weak conclusion* unless we know that β is acceptably small. Therefore, rather than saying we “accept H_0 ,” we prefer the terminology “fail to reject H_0 .” Failing to reject H_0 implies that we have not found sufficient evidence to reject H_0 , that is, to make a strong statement. Failing to reject H_0 does not necessarily mean that there is a high probability that H_0 is true. It may simply mean that more data are required to reach a strong conclusion. This can have important implications for the formulation of hypotheses.

A useful analog exists between hypothesis testing and a jury trial. In a trial, the defendant is assumed innocent (this is like assuming the null hypothesis to be true). If strong evidence is found to the contrary, the defendant is declared to be guilty (we reject the null hypothesis). If evidence is insufficient, the defendant is declared to be not guilty. This is not the same as proving the defendant innocent and so, like failing to reject the null hypothesis, it is a weak conclusion.

An important concept that we use is the power of a statistical test.

Power

The **power** of a statistical test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis is true.

The power is computed as $1 - \beta$, and power can be interpreted as *the probability of correctly rejecting a false null hypothesis*. We often compare statistical tests by comparing their power properties. For example, consider the propellant burning rate problem when we are testing $H_0: \mu = 50$ centimeters per second against $H_1: \mu \neq 50$ centimeters per second. Suppose that the true value of the mean is $\mu = 52$. When $n = 10$, we found that $\beta = 0.2643$, so the power of this test is $1 - \beta = 1 - 0.2643 = 0.7357$ when $\mu = 52$.

Power is a very descriptive and concise measure of the *sensitivity* of a statistical test when by sensitivity we mean the ability of the test to detect differences. In this case, the sensitivity of the test for detecting the difference between a mean burning rate of 50 centimeters per second and

52 centimeters per second is 0.7357. That is, if the true mean is really 52 centimeters per second, this test will correctly reject $H_0: \mu = 50$ and “detect” this difference 73.57% of the time. If this value of power is judged to be too low, the analyst can increase either α or the sample size n .

9.1.3 One-Sided and Two-Sided Hypotheses

In constructing hypotheses, we always state the null hypothesis as an equality so that the probability of type I error α can be controlled at a specific value. The alternative hypothesis might be either one-sided or two-sided, depending on the conclusion to be drawn if H_0 is rejected. If the objective is to make a claim involving statements such as greater than, less than, superior to, exceeds, at least, and so forth, a one-sided alternative is appropriate. If no direction is implied by the claim, or if the claim “not equal to” is to be made, a two-sided alternative should be used.

EXAMPLE 9.1 | Propellant Burning Rate

Consider the propellant burning rate problem. Suppose that if the burning rate is less than 50 centimeters per second, we wish to show this with a strong conclusion. The hypotheses should be stated as

$$\begin{aligned} H_0: \mu &= 50 \text{ centimeters per second} \\ H_1: \mu &< 50 \text{ centimeters per second} \end{aligned}$$

Here the critical region lies in the lower tail of the distribution of \bar{X} . Because the rejection of H_0 is always a strong conclusion, this statement of the hypotheses will produce the desired outcome if H_0 is rejected. Notice that, although the null hypothesis is stated with an equals sign, it is understood to include any value of μ not specified by the alternative hypothesis (that is, $\mu \leq 50$). Therefore, failing to reject H_0 does not mean that $\mu = 50$ centimeters per second exactly, but only that we do not have strong evidence in support of H_1 .

In some real-world problems in which one-sided test procedures are indicated, selecting an appropriate formulation of the alternative hypothesis is occasionally difficult. For example, suppose that a soft-drink beverage bottler purchases 10-ounce bottles from a glass company. The bottler wants to be sure that the bottles meet the specification on mean internal pressure or bursting strength, which for 10-ounce bottles is a minimum strength of 200 psi. The bottler has decided to formulate the decision procedure for a specific lot of bottles as a hypothesis testing problem. There are two possible formulations for this problem, either

$$H_0: \mu = 200 \text{ psi} \qquad H_1: \mu > 200 \text{ psi} \qquad (9.5)$$

or

$$H_0: \mu = 200 \text{ psi} \qquad H_1: \mu < 200 \text{ psi} \qquad (9.6)$$

Consider the formulation in Equation 9.5. If the null hypothesis is rejected, the bottles will be judged satisfactory; if H_0 is not rejected, the implication is that the bottles do not conform to specifications and should not be used. Because rejecting H_0 is a strong conclusion, this formulation forces the bottle manufacturer to “demonstrate” that the mean bursting strength of the bottles exceeds the specification. Now consider the formulation in Equation 9.6. In this situation, the bottles will be judged satisfactory unless H_0 is rejected. That is, we conclude that the bottles are satisfactory unless there is strong evidence to the contrary.

Which formulation is correct, the one of Equation 9.5 or Equation 9.6? The answer is that it depends on the objective of the analysis. For Equation 9.5, there is some probability that H_0 will not be rejected (i.e., we would decide that the bottles are not satisfactory) even though the true mean is slightly greater than 200 psi. This formulation implies that we want the bottle manufacturer to demonstrate that the product meets or exceeds our specifications. Such a formulation could be appropriate if the manufacturer has experienced difficulty in meeting specifications in the past or if product safety considerations force us to hold tightly to the 200-psi specification. On the other hand, for the formulation of Equation 9.6, there is some probability that H_0 will

be accepted and the bottles judged satisfactory, even though the true mean is slightly less than 200 psi. We would conclude that the bottles are unsatisfactory only when there is strong evidence that the mean does not exceed 200 psi, that is, when $H_0: \mu = 200$ psi is rejected. This formulation assumes that we are relatively happy with the bottle manufacturer's past performance and that small deviations from the specification of $\mu \geq 200$ psi are not harmful.

In formulating one-sided alternative hypotheses, we should remember that rejecting H_0 is always a strong conclusion. Consequently, we should put the statement about which it is important to make a strong conclusion in the alternative hypothesis. In real-world problems, this will often depend on our point of view and experience with the situation.

9.1.4 *P*-Values in Hypothesis Tests

One way to report the results of a hypothesis test is to state that the null hypothesis was or was not rejected at a specified α -value or level of significance. This is called **fixed significance level** testing.

The fixed significance level approach to hypothesis testing is very nice because it leads directly to the concepts of type II error and power, which are of considerable value in determining the appropriate sample sizes to use in hypothesis testing. But the fixed significance level approach does have some disadvantages.

For example, in the propellant problem above, we can say that $H_0: \mu = 50$ was rejected at the 0.05 level of significance. This statement of conclusions may be often inadequate because it gives the decision maker no idea about whether the computed value of the test statistic was just barely in the rejection region or whether it was very far into this region. Furthermore, stating the results this way imposes the predefined level of significance on other users of the information. This approach may be unsatisfactory because some decision makers might be uncomfortable with the risks implied by $\alpha = 0.05$.

To avoid these difficulties, the ***P*-value** approach has been adopted widely in practice. The *P*-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true. Thus, a *P*-value conveys much information about the weight of evidence against H_0 , and so a decision maker can draw a conclusion at *any* specified level of significance. We now give a formal definition of a *P*-value.

***P*-Value**

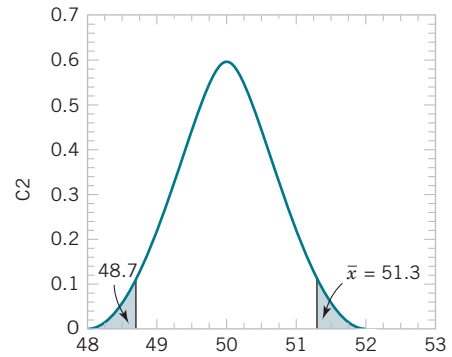
The ***P*-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

It is customary to consider the test statistic (and the data) significant when the null hypothesis H_0 is rejected; therefore, we may think of the *P*-value as the smallest level α at which the data are significant. In other words, the *P*-value is the **observed significance level**. Once the *P*-value is known, the decision maker can determine how significant the data are without the data analyst formally imposing a preselected level of significance.

Consider the two-sided hypothesis test for burning rate

$$H_0: \mu = 50 \quad H_1: \mu \neq 50$$

with $n = 16$ and $\sigma = 2.5$. Suppose that the observed sample mean is $\bar{x} = 51.3$ centimeters per second. Figure 9.6 is a critical region for this test with the value of $\bar{x} = 51.3$ and the symmetric

**FIGURE 9.6**

***P*-value is the area of the shaded region when $\bar{x} = 51.3$.**

value 48.7. The *P*-value of the test is the probability above 51.3 plus the probability below 48.7. The *P*-value is easy to compute after the test statistic is observed. In this example,

$$\begin{aligned}
 P\text{-value} &= 1 - P(48.7 < \bar{X} < 51.3) \\
 &= 1 - P\left(\frac{48.7 - 50}{2.5/\sqrt{16}} < Z < \frac{51.3 - 50}{2.5/\sqrt{16}}\right) \\
 &= 1 - P(-2.08 < Z < 2.08) \\
 &= 1 - 0.962 = 0.038
 \end{aligned}$$

The *P*-value tells us that if the null hypothesis $H_0 = 50$ is true, the probability of obtaining a random sample whose mean is at least as far from 50 as 51.3 (or 48.7) is 0.038. Therefore, an observed sample mean of 51.3 is a fairly rare event if the null hypothesis $H_0 = 50$ is really true. Compared to the “standard” level of significance 0.05, our observed *P*-value is smaller, so if we were using a fixed significance level of 0.05, the null hypothesis would be rejected. In fact, the null hypothesis $H_0 = 50$ would be rejected at *any* level of significance greater than or equal to 0.038. This illustrates the previous boxed definition; the *P*-value is the smallest level of significance that would lead to rejection of $H_0 = 50$.

Operationally, once a *P*-value is computed, we typically compare it to a predefined significance level to make a decision. Often this predefined significance level is 0.05. However, in presenting results and conclusions, it is standard practice to report the observed *P*-value along with the decision that is made regarding the null hypothesis.

Clearly, the *P*-value provides a measure of the credibility of the null hypothesis. Specifically, it is the risk that we have made an incorrect decision if we reject the null hypothesis H_0 . The *P*-value is *not* the probability that the null hypothesis is false, nor is $1 - P$ the probability that the null hypothesis is true. The null hypothesis is either true or false (there is no probability associated with this), so the proper interpretation of the *P*-value is in terms of the risk of wrongly rejecting the null hypothesis H_0 .

Computing the exact *P*-value for a statistical test is not always easy. However, most modern statistics software packages report the results of hypothesis testing problems in terms of *P*-values. We use the *P*-value approach extensively here.

More About *P*-Values We have observed that the procedure for testing a statistical hypothesis consists of drawing a random sample from the population, computing an appropriate statistic, and using the information in that statistic to make a decision regarding the null hypothesis. For example, we have used the sample average in decision making. Because the sample average is a random variable, its value will differ from sample to sample, meaning that the *P*-value associated with the test procedure is also a random variable. It also will differ from sample to sample. We are going to use a computer experiment (a simulation) to show how the *P*-value behaves when the null hypothesis is true and when it is false.

Consider testing the null hypothesis $H_0: \mu = 0$ against the alternative hypothesis $H_0: \mu \neq 0$ when we are sampling from a normal population with standard deviation $\sigma = 1$. Consider first the

case in which the null hypothesis is true and let's suppose that we are going to test the preceding hypotheses using a sample size of $n = 10$. We wrote a computer program to simulate drawing 10,000 different samples at random from a normal distribution with $\mu = 0$ and $\sigma = 1$. Then we calculated the P -values based on the values of the sample averages. Figure 9.7 is a histogram of the P -values obtained from the simulation. Notice that the histogram of the P -values is relatively uniform or flat over the interval from 0 to 1. It turns out that just slightly less than 5% of the P -values are in the interval from 0 to 0.05. It can be shown theoretically that if the null hypothesis is true, the probability distribution of the P -value is exactly uniform on the interval from 0 to 1. Because the null hypothesis is true in this situation, we have demonstrated by simulation that if a test of significance level 0.05 is used, the probability of wrongly rejecting the null hypothesis is (approximately) 0.05.

Now let's see what happens when the null hypothesis is false. We changed the mean of the normal distribution to $\mu = 1$ and repeated the previous computer simulation experiment by drawing another 10,000 samples and computing the P -values. Figure 9.8 is the histogram of the simulated P -values for this situation. Notice that this histogram looks very different from the one in Figure 9.7; there is a tendency for the P -values to stack up near the origin with many more small values between 0 and 0.05 than in the case in which the null hypothesis was true. Not all of the P -values are less than 0.05; those that exceed 0.05 represent type II errors or cases in which the null hypothesis is not rejected at the 0.05 level of significance even though the true mean is not 0.

Finally, Figure 9.9 shows the simulation results when the true value of the mean is even larger; in this case, $\mu = 2$. The simulated P -values are shifted even more toward 0 and concentrated on the left side of the histogram. Generally, as the true mean moves farther and farther away

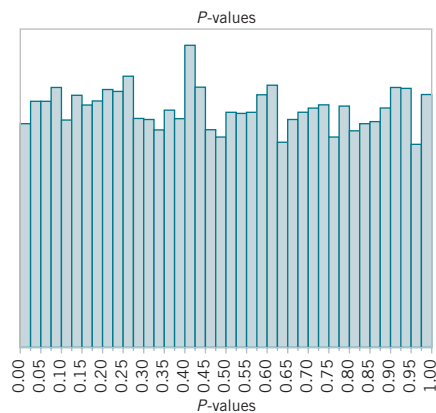


FIGURE 9.7

A P -value simulation when $H_0: \mu = 0$ is true.

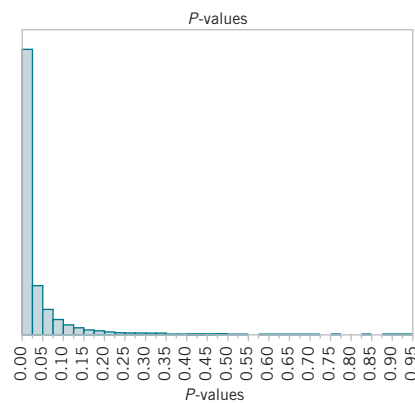


FIGURE 9.8

A P -value simulation when $\mu = 1$.

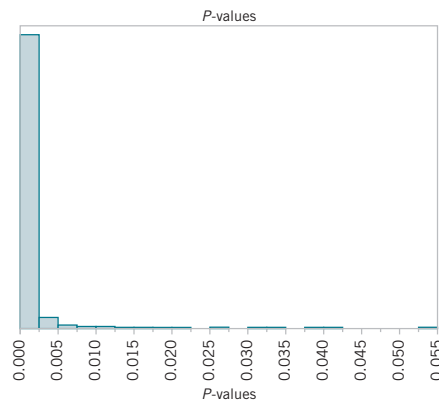


FIGURE 9.9

A P -value simulation when $\mu = 2$.

from the hypothesized value of 0 the distribution of the P -values will become more and more concentrated near 0 and fewer and fewer values will exceed 0.05. That is, the farther the mean is from the value specified in the null hypothesis, the higher is the chance that the test procedure will correctly reject the null hypothesis.

9.1.5 Connection between Hypothesis Tests and Confidence Intervals

A close relationship exists between the test of a hypothesis about any parameter, say θ , and the confidence interval for θ . If $[l, u]$ is a $100(1 - \alpha)\%$ confidence interval for the parameter θ , the test with level of significance α of the hypothesis

$$H_0: \theta = \theta_0 \quad H_1: \theta \neq \theta_0$$

will lead to rejection of H_0 if and only if θ_0 is *not* in the $100(1 - \alpha)\%$ CI $[l, u]$. As an illustration, consider the escape system propellant problem with $\bar{x} = 51.3$, $\sigma = 2.5$, and $n = 16$. The null hypothesis $H_0: \mu = 50$ was rejected, using $\alpha = 0.05$. The 95% two-sided CI on μ can be calculated using Equation 8.7. This CI is $51.3 \pm 1.96(2.5/\sqrt{16})$ and this is $50.075 \leq \mu \leq 52.525$. Because the value $\mu_0 = 50$ is not included in this interval, the null hypothesis $H_0: \mu = 50$ is rejected.

Although hypothesis tests and CIs are equivalent procedures insofar as decision making or **inference** about μ is concerned, each provides somewhat different insights. For instance, the confidence interval provides a range of likely values for μ at a stated confidence level whereas hypothesis testing is an easy framework for displaying the *risk levels* such as the P -value associated with a specific decision. We continue to illustrate the connection between the two procedures throughout the text.

9.1.6 General Procedure for Hypothesis Tests

This chapter develops hypothesis-testing procedures for many practical problems. Use of the following sequence of steps in applying hypothesis-testing methodology is recommended.

1. **Parameter of interest:** From the problem context, identify the parameter of interest.
2. **Null hypothesis, H_0 :** State the null hypothesis, H_0 .
3. **Alternative hypothesis, H_1 :** Specify an appropriate alternative hypothesis, H_1 .
4. **Test statistic:** Determine an appropriate test statistic.
5. **Reject H_0 if:** State the rejection criteria for the null hypothesis.
6. **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
7. **Draw conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context.

Steps 1–4 should be completed prior to examining the sample data. This sequence of steps is illustrated in subsequent sections.

In practice, such a formal and (seemingly) rigid procedure is not always necessary. Generally, once the experimenter (or decision maker) has decided on the question of interest and has determined the *design of the experiment* (that is, how the data are to be collected, how the measurements are to be made, and how many observations are required), only three steps are really required:

1. Specify the test statistic to be used (such as Z_0).
2. Specify the location of the critical region (two-tailed, upper-tailed, or lower-tailed).
3. Specify the criteria for rejection (typically, the value of α or the P -value at which rejection should occur).

These steps are often completed almost simultaneously in solving real-world problems, although we emphasize that it is important to think carefully about each step. That is why we present and use the seven-step process; it seems to reinforce the essentials of the correct approach. Although we may not use it every time in solving real problems, it is a helpful framework when we are first learning about hypothesis testing.

Statistical Versus Practical Significance We noted previously that reporting the results of a hypothesis test in terms of a P -value is very useful because it conveys more information than just the simple statement “reject H_0 ” or “fail to reject H_0 .” That is, rejection of H_0 at the 0.05 level of significance is much more meaningful if the value of the test statistic is well into the critical region, greatly exceeding the 5% critical value, than if it barely exceeds that value.

Even a very small P -value can be difficult to interpret from a practical viewpoint when we are making decisions because, although a small P -value indicates **statistical significance** in the sense that H_0 should be rejected in favor of H_1 , the actual departure from H_0 that has been detected may have little (if any) **practical significance** (engineers like to say “engineering significance”). This is particularly true when the sample size n is large.

For example, consider the propellant burning rate problem in Example 9.1 in which we test $H_0: \mu = 50$ centimeters per second versus $H_1: \mu \neq 50$ centimeters per second with $\sigma = 2.5$. If we suppose that the mean rate is really 50.5 centimeters per second, this is not a serious departure from $H_0: \mu = 50$ centimeters per second in the sense that if the mean really is 50.5 centimeters per second, there is no practical observable effect on the performance of the air crew escape system. In other words, concluding that $\mu = 50$ centimeters per second when it is really 50.5 centimeters per second is an inexpensive error and has no practical significance. For a reasonably large sample size, a true value of $\mu = 50.5$ will lead to a sample \bar{x} that is close to 50.5 centimeters per second, and we would not want this value of \bar{x} from the sample to result in rejection of H_0 . The following display shows the P -value for testing $H_0: \mu = 50$ when we observe $\bar{x} = 50.5$ centimeters per second and the power of the test at $\alpha = 0.05$ when the true mean is 50.5 for various sample sizes n :

Sample Size n	P -value When $\bar{x} = 50.5$	Power (at $\alpha = 0.05$) When True $\mu = 50.5$
10	0.527	0.097
25	0.317	0.170
50	0.157	0.293
100	0.046	0.516
400	6.3×10^{-5}	0.979
1000	2.5×10^{-10}	1.000

The P -value column in this display indicates that for large sample sizes, the observed sample value of $\bar{x} = 50.5$ would strongly suggest that $H_0: \mu = 50$ should be rejected, even though the observed sample results imply that from a practical viewpoint, the true mean does not differ much at all from the hypothesized value $\mu_0 = 50$. The power column indicates that if we test a hypothesis at a fixed significance level α , and even if there is little practical difference between the true mean and the hypothesized value, a large sample size will almost always lead to rejection of H_0 . The moral of this demonstration is clear:

Be careful when interpreting the results from hypothesis testing when the sample size is large because any small departure from the hypothesized value μ_0 will probably be detected, even when the difference is of little or no practical significance.

9.2 Tests on the Mean of a Normal Distribution, Variance Known

In this section, we consider hypothesis testing about the mean μ of a single normal population where the variance of the population σ^2 is known. We assume that a random sample X_1, X_2, \dots, X_n has been taken from the population. Based on our previous discussion, the sample mean \bar{X} is an unbiased point estimator of μ with variance σ^2/n .

9.2.1 Hypothesis Tests on the Mean

Suppose that we wish to test the hypotheses

$$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0 \quad (9.7)$$

where μ_0 is a specified constant. We have a random sample X_1, X_2, \dots, X_n from a normal population. Because \bar{X} has a normal distribution (i.e., the *sampling distribution* of \bar{X} is normal) with mean μ_0 and standard deviation σ/\sqrt{n} if the null hypothesis is true, we could calculate a *P*-value or construct a critical region based on the computed value of the sample mean \bar{X} , as in Section 9.1.2.

It is usually more convenient to *standardize* the sample mean and use a test statistic based on the standard normal distribution. That is, the test procedure for $H_0: \mu = \mu_0$ uses the *test statistic*:

Test Statistic

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (9.8)$$

If the null hypothesis $H_0: \mu = \mu_0$ is true, $E(\bar{X}) = \mu_0$, and it follows that the distribution of Z_0 is the standard normal distribution [denoted $N(0, 1)$].

The hypothesis testing procedure is as follows. Take a random sample of size n and compute the value of the sample mean \bar{x} . To test the null hypothesis using the *P*-value approach, we would find the probability of observing a value of the sample mean that is at least as extreme as \bar{x} , given that the null hypothesis is true. The standard normal *z*-value that corresponds to \bar{x} is found from the test statistic in Equation 9.8:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

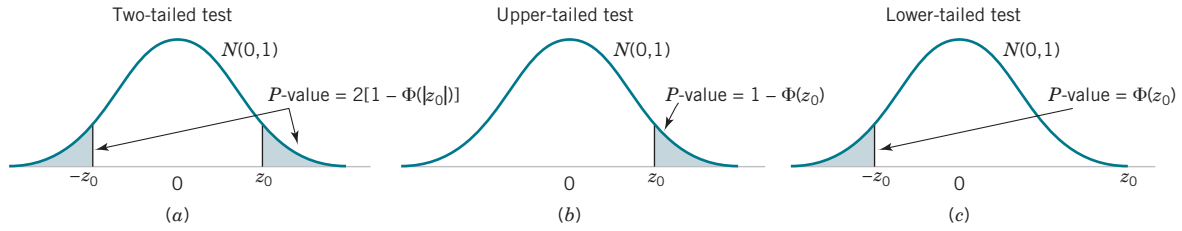
In terms of the standard normal cumulative distribution function (CDF), the probability we are seeking is $1 - \Phi(|z_0|)$. The reason that the argument of the standard normal cdf is $|z_0|$ is that the value of z_0 could be either positive or negative, depending on the observed sample mean. Because this is a two-tailed test, this is only one-half of the *P*-value. Therefore, for the two-sided alternative hypothesis, the *P*-value is

$$P = 2 [1 - \Phi(|z_0|)] \quad (9.9)$$

This is illustrated in Figure 9.10(a).

Now let's consider the one-sided alternatives. Suppose that we are testing

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \quad (9.10)$$

**FIGURE 9.10**

The P -value for a z -test. (a) The two-sided alternative $H_1: \mu \neq \mu_0$. (b) The one-sided alternative $H_1: \mu > \mu_0$. (c) The one-sided alternative $H_1: \mu < \mu_0$.

Once again, suppose that we have a random sample of size n and that the sample mean is \bar{x} . We compute the test statistic from Equation 9.8 and obtain z_0 . Because the test is an upper-tailed test, only values of \bar{x} that are greater than μ_0 are consistent with the alternative hypothesis. Therefore, the P -value would be the probability that the standard normal random variable is greater than the value of the test statistic z_0 . This P -value is computed as

$$P = 1 - \Phi(z_0) \quad (9.11)$$

This P -value is shown in Figure 9.10(b).

The lower-tailed test involves the hypotheses

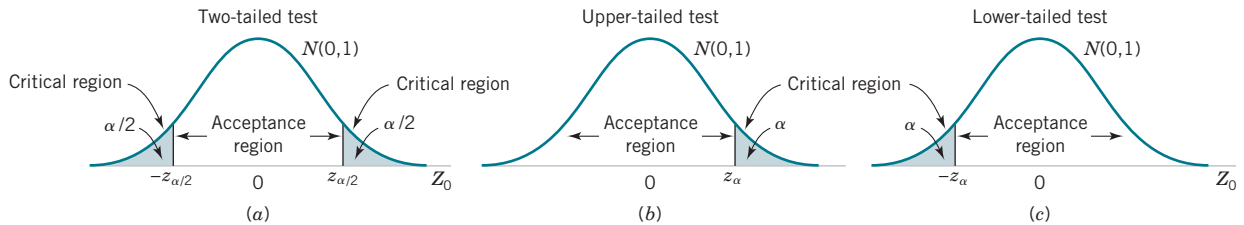
$$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \quad (9.12)$$

Suppose that we have a random sample of size n and that the sample mean is \bar{x} . We compute the test statistic from Equation 9.8 and obtain z_0 . Because the test is a lower-tailed test, only values of \bar{x} that are less than μ_0 are consistent with the alternative hypothesis. Therefore, the P -value would be the probability that the standard normal random variable is less than the value of the test statistic z_0 . This P -value is computed as

$$P = \Phi(z_0) \quad (9.13)$$

and shown in Figure 9.10(c). The **reference distribution** for this test is the standard normal distribution. The test is usually called a **z -test**.

We can also use the fixed significance level approach with the z -test. The only thing we have to do is determine where to place the critical regions for the two-sided and one-sided alternative hypotheses. First, consider the two-sided alternative in Equation 9.10. Now if $H_0: \mu = \mu_0$ is true, the probability is $1 - \alpha$ that the test statistic Z_0 falls between $-z_{\alpha/2}$ and $z_{\alpha/2}$ where $z_{\alpha/2}$ is the 100($\alpha/2$) percentage point of the standard normal distribution. The regions associated with $z_{\alpha/2}$ and $-z_{\alpha/2}$ are illustrated in Figure 9.11(a). Note that the probability is α that the test statistic Z_0 will fall in the region $Z_0 > z_{\alpha/2}$ or $Z_0 < -z_{\alpha/2}$, when $H_0: \mu = \mu_0$ is true. Clearly, a sample producing

**FIGURE 9.11**

The distribution of Z_0 when $H_0: \mu = \mu_0$ is true with critical region for (a) the two-sided alternative $H_1: \mu \neq \mu_0$, (b) the one-sided alternative $H_1: \mu > \mu_0$, and (c) the one-sided alternative $H_1: \mu < \mu_0$.

a value of the test statistic that falls in the tails of the distribution of Z_0 would be unusual if $H_0: \mu = \mu_0$ is true; therefore, it is an indication that H_0 is false. Thus, we should reject H_0 if either

$$z_0 > z_{\alpha/2} \quad (9.14)$$

or

$$z_0 < -z_{\alpha/2} \quad (9.15)$$

and we should fail to reject H_0 if

$$-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2} \quad (9.16)$$

Equations 9.14 and 9.15 define the **critical region** or **rejection region** for the test. The type I error probability for this test procedure is α .

We may also develop fixed significance level testing procedures for the one-sided alternatives. Consider the upper-tailed case in Equation 9.10.

In defining the critical region for this test, we observe that a negative value of the test statistic Z_0 would never lead us to conclude that $H_0: \mu = \mu_0$ is false. Therefore, we would place the critical region in the upper tail of the standard normal distribution and reject H_0 if the computed value z_0 is too large. Refer to Figure 9.11(b). That is, we would reject H_0 if

$$z_0 > z_{\alpha} \quad (9.17)$$

Similarly, to test the lower-tailed case in Equation 9.12, we would calculate the test statistic Z_0 and reject H_0 if the value of Z_0 is too small. That is, the critical region is in the lower tail of the standard normal distribution as in Figure 9.11(c), and we reject H_0 if

$$z_0 < -z_{\alpha} \quad (9.18)$$

Summary of Tests on the Mean, Variance Known

Testing Hypotheses on the Mean, Variance Known (Z-Tests)

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above $ z_0 $ and probability below $- z_0 $, $P = 2[1 - \Phi(z_0)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu > \mu_0$	Probability above z_0 , $P = 1 - \Phi(z_0)$	$z_0 > z_{\alpha}$
$H_1: \mu < \mu_0$	Probability below z_0 , $P = \Phi(z_0)$	$z_0 < -z_{\alpha}$

The P -values and critical regions for these situations are shown in Figures 9.10 and 9.11.

In general, understanding the critical region and the test procedure is easier when the test statistic is Z_0 rather than \bar{X} . However, the same critical region can always be written in terms of the computed value of the sample mean \bar{x} . A procedure identical to the preceding fixed significance level test is as follows:

Reject $H_0: \mu = \mu_0$ if either $\bar{x} > a$ or $\bar{x} < b$

where

$$a = \mu_0 + z_{\alpha/2}\sigma/\sqrt{n} \quad \text{and} \quad b = \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$$

EXAMPLE 9.2 | Propellant Burning Rate

Air crew escape systems are powered by a solid propellant. The burning rate of this propellant is an important product characteristic. Specifications require that the mean burning rate must be 50 centimeters per second. We know that the standard deviation of burning rate is $\sigma = 2$ centimeters per second. The experimenter decides to specify a type I error probability or significance level of $\alpha = 0.05$ and selects a random sample of $n = 25$ and obtains a sample average burning rate of $\bar{x} = 51.3$ centimeters per second. What conclusions should be drawn?

We may solve this problem by following the seven-step procedure outlined in Section 9.1.6. This results in

1. **Parameter of interest:** The parameter of interest is μ , the mean burning rate.
2. **Null hypothesis:** $H_0: \mu = 50$ centimeters per second
3. **Alternative hypothesis:** $H_1: \mu \neq 50$ centimeters per second

4. **Test statistic:** The test statistic is

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

5. **Reject H_0 if:** Reject H_0 if the P -value is less than 0.05. To use a fixed significance level test, the boundaries of the critical region would be $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.

6. **Computations:** Because $\bar{x} = 51.3$ and $\sigma = 2$,

$$z_0 = \frac{51.3 - 50}{2/\sqrt{25}} = 3.25$$

7. **Conclusion:** Because the P -value $= 2[1 - \Phi(3.25)] = 0.0012$ we reject $H_0: \mu = 50$ at the 0.05 level of significance.

Practical Interpretation: We conclude that the mean burning rate differs from 50 centimeters per second, based on a sample of 25 measurements. In fact, there is strong evidence that the mean burning rate exceeds 50 centimeters per second.

9.2.2 Type II Error and Choice of Sample Size

In testing hypotheses, the analyst directly selects the type I error probability. However, the probability of type II error β depends on the choice of sample size. In this section, we show how to calculate the probability of type II error β . We also show how to select the sample size to obtain a specified value of β .

Finding the Probability of Type II Error β Consider the two-sided hypotheses

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

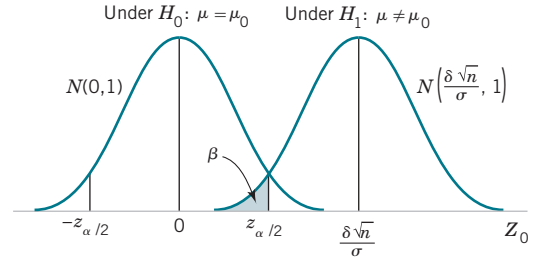
Suppose that the null hypothesis is false and that the true value of the mean is $\mu = \mu_0 + \delta$, say, where $\delta > 0$. The expected value of the test statistic Z_0 is

$$E(Z_0) = \frac{E(\bar{X}) - \mu_0}{\sigma/\sqrt{n}} = \frac{(\mu_0 + \delta) - \mu_0}{\sigma/\sqrt{n}} = \frac{\delta\sqrt{n}}{\sigma}$$

Therefore, the distribution of Z_0 when H_1 is true is

$$Z_0 \sim N\left(\frac{\delta\sqrt{n}}{\sigma}, 1\right) \quad (9.19)$$

The distribution of the test statistic Z_0 under both the null hypothesis H_0 and the alternate hypothesis H_1 is shown in Figure 9.9. From examining this figure, we note that if H_1 is true, a type II error will be made only if $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ where $Z_0 \sim N(\delta\sqrt{n}/\sigma, 1)$. That is, the probability of the type II error β is the probability that Z_0 falls between $-z_{\alpha/2}$ and $z_{\alpha/2}$ given that H_1 is true.

**FIGURE 9.12**

The distribution of Z_0 under H_0 and H_1 .

This probability is shown as the shaded portion of Figure 9.12. Expressed mathematically, this probability is

Probability of a Type II Error for a Two-Sided Test on the Mean, Variance Known

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \quad (9.20)$$

where $\Phi(z)$ denotes the probability to the left of z in the standard normal distribution. Note that Equation 9.20 was obtained by evaluating the probability that Z_0 falls in the interval $[-z_{\alpha/2}, z_{\alpha/2}]$ when H_1 is true. Furthermore, note that Equation 9.20 also holds if $\delta < 0$ because of the symmetry of the normal distribution. It is also possible to derive an equation similar to Equation 9.20 for a one-sided alternative hypothesis.

Sample Size Formulas One may easily obtain formulas that determine the appropriate sample size to obtain a particular value of β for a given Δ and α . For the two-sided alternative hypothesis, we know from Equation 9.20 that

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

or, if $\delta > 0$,

$$\beta \simeq \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \quad (9.21)$$

because $\Phi(-z_{\alpha/2} - \delta\sqrt{n}/\sigma) \simeq 0$ when δ is positive. Let z_β be the 100β upper percentile of the standard normal distribution. Then, $\beta = \Phi(-z_\beta)$. From Equation 9.21,

$$-z_\beta \simeq z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}$$

or

Sample Size for a Two-Sided Test on the Mean, Variance Known

$$n \simeq \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0 \quad (9.22)$$

If n is not an integer, the convention is to round the sample size up to the next integer. This approximation is good when $\Phi(-z_{\alpha/2} - \delta\sqrt{n}/\sigma)$ is small compared to β . For either of the one-sided alternative hypotheses, the sample size required to produce a specified type II error with probability β given δ and α is

Sample Size for a One-Sided Test on the Mean, Variance Known

$$n \simeq \frac{(z_{\alpha} + z_{\beta})^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0 \quad (9.23)$$

EXAMPLE 9.3 | Propellant Burning Rate Type II Error

Consider the rocket propellant problem in Example 9.2. Suppose that the true burning rate is 49 centimeters per second. What is β for the two-sided test with $\alpha = 0.05$, $\sigma = 2$, and $n = 25$?

Here $\delta = 1$ and $z_{\alpha/2} = 1.96$. From Equation 9.20,

$$\begin{aligned} \beta &= \Phi\left(1.96 - \frac{\sqrt{25}}{\sigma}\right) - \Phi\left(-1.96 - \frac{\sqrt{25}}{\sigma}\right) \\ &= \Phi(-0.54) - \Phi(-4.46) = 0.295 \end{aligned}$$

The probability is about 0.3 that this difference from 50 centimeters per second will not be detected. That is, the probability is about 0.3 that the test will fail to reject the null hypothesis when the true burning rate is 49 centimeters per second.

Practical Interpretation: A sample size of $n = 25$ results in reasonable, but not great, power = $1 - \beta = 1 - 0.3 = 0.70$.

Suppose that the analyst wishes to design the test so that if the true mean burning rate differs from 50 centimeters per second by as much as 1 centimeter per second, the test will detect this (i.e., reject $H_0: \mu = 50$) with a high probability, say, 0.90. Now we note that $\sigma = 2$, $\delta = 51 - 50 = 1$, $\alpha = 0.05$, and $\beta = 0.10$. Because $z_{\alpha/2} = z_{0.025} = 1.96$ and $z_{\beta} = z_{0.10} = 1.28$, the sample size required to detect this departure from $H_0: \mu = 50$ is found by Equation 9.22 as

$$n \simeq \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2} = \frac{(1.96 + 1.28)^2 2^2}{(1^2)} \simeq 42$$

The approximation is good here because $\Phi(-z_{\alpha/2} - \delta\sqrt{n}/\sigma) = \Phi(-1.96 - (1)\sqrt{42}/2) = \Phi(-5.20) \simeq 0$, which is small relative to β .

Practical Interpretation: To achieve a much higher power of 0.90, you will need a considerably large sample size, $n = 42$ instead of $n = 25$.

Using Operating Characteristic Curves When performing sample size or type II error calculations, it is sometimes more convenient to use the **operating characteristic (OC) curves** in Appendix Charts VIIa & b. These curves plot β as calculated from Equation 9.20 against a parameter d for various sample sizes n . Curves are provided for both $\alpha = 0.05$ and $\alpha = 0.01$. The parameter d is defined as

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} \quad (9.24)$$

so one set of operating characteristic curves can be used for all problems regardless of the values of μ_0 and σ . From examining the operating characteristic curves or from Equation 9.20 and Figure 9.9, we note that

1. The farther the true value of the mean μ is from μ_0 , the smaller the probability of type II error β for a given n and α . That is, we see that for a specified sample size and α , large differences in the mean are easier to detect than small ones.
2. For a given δ and α , the probability of type II error β decreases as n increases. That is, to detect a specified difference δ in the mean, we may make the test more powerful by increasing the sample size.

EXAMPLE 9.4 | Propellant Burning Rate Type II Error from OC Curve

Consider the propellant problem in Example 9.2. Suppose that the analyst is concerned about the probability of type II error if the true mean burning rate is $\mu = 51$ centimeters per second. We may use the operating characteristic curves to find β . Note that $\delta = 51 - 50 = 1$, $n = 25$, $\sigma = 2$, and $\alpha = 0.05$. Then using Equation 9.24 gives

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} = \frac{1}{2}$$

and from Appendix Chart VIIa with $n = 25$, we find that $\beta = 0.30$. That is, if the true mean burning rate is $\mu = 51$ centimeters per second, there is approximately a 30% chance that this will not be detected by the test with $n = 25$.

EXAMPLE 9.5 | Propellant Burning Rate Sample Size from OC Curve

Once again, consider the propellant problem in Example 9.2. Suppose that the analyst would like to design the test so that if the true mean burning rate differs from 50 centimeters per second by as much as 1 centimeter per second, the test will detect this (i.e., reject $H_0: \mu = 50$) with a high probability, say, 0.90. This is exactly the same requirement as in Example 9.3

in which we used Equation 9.22 to find the required sample size to be $n = 42$. The operating characteristic curves can also be used to find the sample size for this test. Because $d = |\mu - \mu_0|/\sigma = 1/2$, $\alpha = 0.05$, and $\beta = 0.10$, we find from Appendix Chart VIIa that the required sample size is approximately $n = 40$. This closely agrees with the sample size calculated from Equation 9.22.

In general, the operating characteristic curves involve three parameters: β , d , and n . Given any two of these parameters, the value of the third can be determined. There are two typical applications of these curves:

1. For a given n and d , find β (as illustrated in Example 9.4). Analysts often encounter this kind of problem when they are concerned about the sensitivity of an experiment already performed, or when sample size is restricted by economic or other factors.
2. For a given β and d , find n . This was illustrated in Example 9.5. Analysts usually encounter this kind of problem when they have the opportunity to select the sample size at the outset of the experiment.

Operating characteristic curves are given in Appendix Charts VIIc and VIId for the one-sided alternatives. If the alternative hypothesis is either $H_1: \mu > \mu_0$ or $H_1: \mu < \mu_0$, the abscissa scale on these charts is

$$d = \frac{|\mu - \mu_0|}{\sigma} \tag{9.25}$$

Using the Computer Many statistics software packages can calculate sample sizes and type II error probabilities. To illustrate, here are some typical computer calculations for the propellant burning rate problem:

Power and Sample Size

1-Sample Z-Test

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 2

Difference	Sample Size	Target Power	Actual Power
1	43	0.9000	0.9064

Power and Sample Size

1-Sample Z-Test

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 2

Difference	Sample Size	Target Power	Actual Power
1	28	0.7500	0.7536

Power and Sample Size

1-Sample Z-Test

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 2

Difference	Sample Size	Power
1	25	0.7054

In the first part of the boxed display, we worked Example 9.3, that is, to find the sample size n that would allow detection of a difference from $\mu_0 = 50$ of 1 centimeter per second with power of 0.9 and $\alpha = 0.05$. The answer, $n = 43$, agrees closely with the calculated value from Equation 9.22 in Example 9.3, which was $n = 42$. The difference is due to the software's use of a value of z_β that has more than two decimal places. The second part of the computer output relaxes the power requirement to 0.75. Note that the effect is to reduce the required sample size to $n = 28$. The third part of the output is the solution to Example 9.4 for which we wish to determine the type II error probability of (β) or the power = $1 - \beta$ for the sample size $n = 25$. Note that software computes the power to be 0.7054, which agrees closely with the answer obtained from the OC curve in Example 9.4. Generally, however, the computer calculations will be more accurate than visually reading values from an OC curve.

9.2.3 Large-Sample Test

We have developed the test procedure for the null hypothesis $H_0: \mu = \mu_0$ assuming that the population is normally distributed and that σ^2 is known. In many if not most practical situations, σ^2 will be unknown. Furthermore, we may not be certain that the population is well modeled by a normal distribution. In these situations, if n is large (say, $n > 40$), the sample standard deviation s can be substituted for σ in the test procedures with little effect. Thus, although we have given a test for the mean of a normal distribution with known σ^2 , it can be easily converted into a *large-sample test procedure for unknown σ^2* that is valid regardless of the form of the distribution of the population. This large-sample test relies on the central limit theorem just as the large-sample confidence interval on μ that was presented in the previous chapter did. Exact treatment of the case in which the population is normal, σ^2 is unknown, and n is small involves use of the t distribution and is deferred until Section 9.3.

9.3 Tests on the Mean of a Normal Distribution, Variance Unknown

9.3.1 Hypothesis Tests on the Mean

We now consider the case of **hypothesis testing** on the mean of a population with *unknown variance* σ^2 . The situation is analogous to the one in Section 8.2 where we considered a *confidence*

interval on the mean for the same situation. As in that section, the validity of the test procedure we describe rests on the assumption that the population distribution is at least approximately normal. The important result on which the test procedure relies is that if X_1, X_2, \dots, X_n is a random sample from a normal distribution with mean μ and variance σ^2 , the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom. Recall that we used this result in Section 8.2 to devise the t -confidence interval for μ . Now consider testing the hypotheses

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

We use the test statistic:

Test Statistic

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad (9.26)$$

If the null hypothesis is true, T_0 has a t distribution with $n - 1$ degrees of freedom. When we know the distribution of the test statistic when H_0 is true (this is often called the **reference distribution** or the **null distribution**), we can calculate the P -value from this distribution, or, if we use a fixed significance level approach, we can locate the critical region to control the type I error probability at the desired level.

To test $H_0: \mu = \mu_0$ against the two-sided alternative $H_1: \mu \neq \mu_0$, the value of the test statistic t_0 in Equation 9.26 is calculated, and the P -value is found from the t distribution with $n - 1$ degrees of freedom (denoted by T_{n-1}). Because the test is two-tailed, the P -value is the sum of the probabilities in the two tails of the t distribution. Refer to Figure 9.13(a). The P -value is the probability above $|t_0|$ plus the probability below. Because the t distribution is symmetric around zero, a simple way to write this is

$$P = 2P(T_{n-1} > |t_0|) \quad (9.27)$$

A small P -value is evidence against H_0 , so if P is of sufficiently small value (typically < 0.05), reject the null hypothesis.

For the one-sided alternative hypotheses,

$$H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \quad (9.28)$$

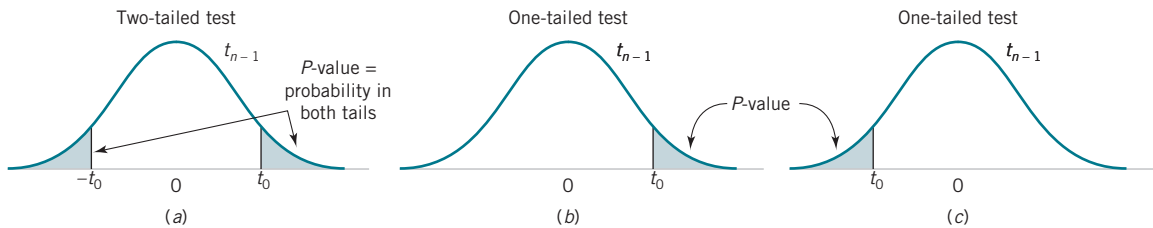


FIGURE 9.13

Calculating the P -value for a t -test: (a) $H_1: \mu \neq \mu_0$, (b) $H_1: \mu > \mu_0$, (c) $H_1: \mu < \mu_0$.

we calculate the test statistic t_0 from Equation 9.26 and calculate the P -value as

$$P = P(T_{n-1} > t_0) \quad (9.29)$$

For the other one-sided alternative,

$$H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \quad (9.30)$$

we calculate the P -value as

$$P = P(T_{n-1} < t_0) \quad (9.31)$$

Figure 9.13(b) and (c) show how these P -values are calculated.

Statistics software packages calculate and display P -values. However, in working problems by hand, it is useful to be able to find the P -value for a ***t*-test**. Because the t -table in Appendix A Table V contains only 10 critical values for each t distribution, determining the exact P -value from this table is usually impossible. Fortunately, it is easy to find lower and upper bounds on the P -value by using this table.

To illustrate, suppose that we are conducting an upper-tailed t -test (so $H_1: \mu > \mu_0$) with 14 degrees of freedom. The relevant critical values from Appendix A Table II are as follows:

Critical value:	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
Tail area:	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005

After calculating the test statistic, we find that $t_0 = 2.8$. Now $t_0 = 2.8$ is between two tabulated values, 2.624 and 2.977. Therefore, the P -value must be between 0.01 and 0.005. Refer to Figure 9.14. These are effectively the upper and lower bounds on the P -value.

This illustrates the procedure for an upper-tailed test. If the test is lower-tailed, just change the sign on the lower and upper bounds for t_0 and proceed in the same way. Remember that for a two-tailed test, the level of significance associated with a particular critical value is twice the corresponding tail area in the column heading. This consideration must be taken into account when we compute the bound on the P -value. For example, suppose that $t_0 = 2.8$ for a two-tailed alternative based on 14 degrees of freedom. The value of the test statistic $t_0 > 2.624$ (corresponding to $\alpha = 2 \times 0.01 = 0.02$) and $t_0 < 2.977$ (corresponding to $\alpha = 2 \times 0.005 = 0.01$), so the lower and upper bounds on the P -value would be $0.01 < P < 0.02$ for this case.

Some statistics software packages can calculate P -values. For example, many software packages have the capability to find cumulative probabilities from many standard probability

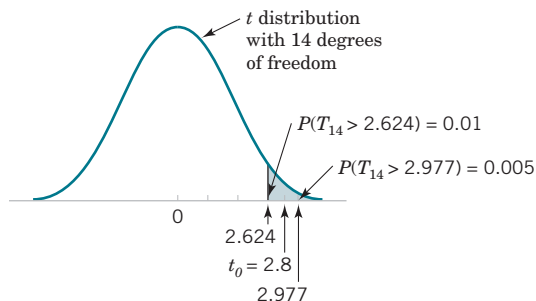


FIGURE 9.14

P -value for $t_0 = 2.8$; an upper-tailed test is shown to be between 0.005 and 0.01.

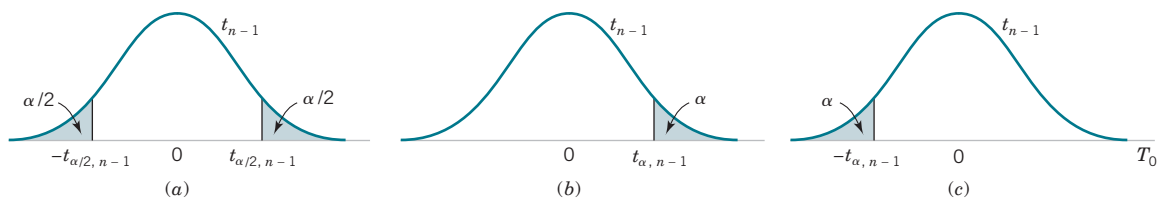


FIGURE 9.15

The distribution of T_0 when $H_0: \mu = \mu_0$ is true with critical region for (a) $H_1: \mu \neq \mu_0$, (b) $H_1: \mu > \mu_0$, and (c) $H_1: \mu < \mu_0$.

distributions, including the t distribution. Simply enter the value of the test statistic t_0 along with the appropriate number of degrees of freedom. Then the software will display the probability $P(T_\nu \leq t_0)$ where ν is the degrees of freedom for the test statistic t_0 . From the cumulative probability, the P -value can be determined.

The single-sample t -test we have just described can also be conducted using the **fixed significance level** approach. Consider the two-sided alternative hypothesis. The null hypothesis would be rejected if the value of the test statistic t_0 falls in the critical region defined by the lower and upper $\alpha/2$ percentage points of the t distribution with $n - 1$ degrees of freedom. That is, reject H_0 if

$$t_0 > t_{\alpha/2, n-1} \quad \text{or} \quad t_0 < -t_{\alpha/2, n-1}$$

For the one-tailed tests, the location of the critical region is determined by the direction to which the inequality in the alternative hypothesis “points.” So, if the alternative is $H_1: \mu > \mu_0$, reject H_0 if

$$t_0 > t_{\alpha, n-1}$$

and if the alternative is $H_1: \mu < \mu_0$, reject H_0 if

$$t_0 < -t_{\alpha, n-1}$$

Figure 9.15 provides the locations of these critical regions.

Summary for the One-Sample t -test

Testing Hypotheses on the Mean of a Normal Distribution, Variance Unknown

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above $ t_0 $ and probability below $- t_0 $	$t_0 > t_{\alpha/2, n-1}$ or $t_0 < -t_{\alpha/2, n-1}$
$H_1: \mu > \mu_0$	Probability above t_0	$t_0 > t_{\alpha, n-1}$
$H_1: \mu < \mu_0$	Probability below t_0	$t_0 < -t_{\alpha, n-1}$

The calculations of the P -values and the locations of the critical regions for these situations are shown in Figures 9.13 and 9.15, respectively.

EXAMPLE 9.6 | Golf Club Design

The increased availability of light materials with high strength has revolutionized the design and manufacture of golf clubs, particularly drivers. Clubs with hollow heads and very thin faces can result in much longer tee shots, especially for players of modest skills. This is due partly to the “spring-like effect” that the thin face imparts to the ball. Firing a golf ball at the head of the club and measuring the ratio of the ball’s outgoing velocity to the incoming velocity can quantify this spring-like effect. The ratio of velocities is called the *coefficient of restitution of the club*. An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. In the experiment, the golf balls were fired from an air cannon so that the incoming velocity and spin rate of the ball could be precisely controlled. It is of interest to determine whether there is evidence (with $\alpha = 0.05$) to support a claim that the mean coefficient of restitution exceeds 0.82. The observations follow:

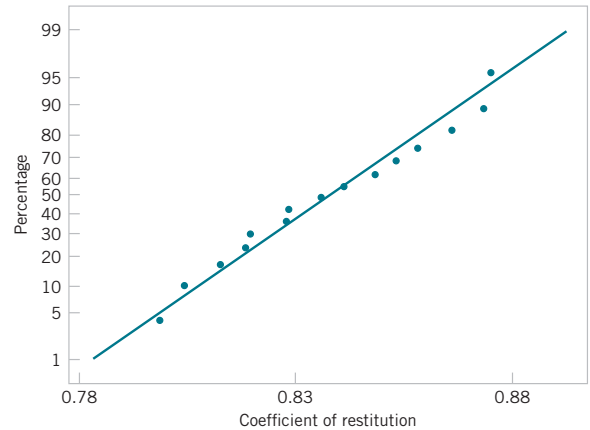
0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

The sample mean and sample standard deviation are $\bar{x} = 0.83725$ and $s = 0.02456$. The normal probability plot of the data in Figure 9.16 supports the assumption that the coefficient of restitution is normally distributed. Because the experiment’s objective is to demonstrate that the mean coefficient of restitution exceeds 0.82, a one-sided alternative hypothesis is appropriate.

The solution using the seven-step procedure for hypothesis testing is as follows:

- Parameter of interest:** The parameter of interest is the mean coefficient of restitution, μ .
- Null hypothesis:** $H_0: \mu = 0.82$
- Alternative hypothesis:** $H_1: \mu > 0.82$. We want to reject H_0 if the mean coefficient of restitution exceeds 0.82.
- Test statistic:** The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

**FIGURE 9.16**

Normal probability plot of the coefficient of restitution data.

- Reject H_0 if:** Reject H_0 if the P -value is less than 0.05.
- Computations:** Because $\bar{x} = 0.83725$, $s = 0.02456$, $\mu_0 = 0.82$, and $n = 15$, we have

$$t_0 = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72$$

- Conclusions:** From Appendix A Table II we find for a t distribution with 14 degrees of freedom that $t_0 = 2.72$ falls between two values: 2.624, for which $\alpha = 0.01$, and 2.977, for which $\alpha = 0.005$. Because this is a one-tailed test, we know that the P -value is between those two values, that is, $0.005 < P < 0.01$. Therefore, because $P < 0.05$, we reject H_0 and conclude that the mean coefficient of restitution exceeds 0.82.

Practical Interpretation: There is strong evidence to conclude that the mean coefficient of restitution exceeds 0.82.

Normality and the t -Test The development of the t -test assumes that the population from which the random sample is drawn is normal. This assumption is required to formally derive the t distribution as the reference distribution for the test statistic in Equation 9.26. Because it can be difficult to identify the form of a distribution based on a small sample, a logical question to ask is how important this assumption is. Studies have investigated this. Fortunately, studies have found that the t -test is relatively insensitive to the normality assumption. If the underlying population is reasonably symmetric and unimodal, the t -test will work satisfactorily. The exact significance level will not match the “advertised” level; for instance, the results may be significant at the 6%

or 7% level instead of the 5% level. This is usually not a serious problem in practice. A normal probability plot of the sample data as illustrated for the golf club data in Figure 9.16 is usually a good way to verify the adequacy of the normality assumption. Only severe departures from normality that are evident in the plot should be a cause for concern.

Many software packages conduct the one-sample t -test. Typical computer output for Example 9.6 is shown in the following display:

One-Sample T

Test of $\mu = 0.82$ vs $\mu > 0.82$							
Variable	N	Mean	StDev	SE Mean	95.0% Lower confidence bound	T	P-value
COR	15	0.83725	0.02456	0.00634	0.82608	2.72	0.008

Notice that the software computes both the test statistic T_0 and a 95% lower confidence bound for the coefficient of restitution. The reported P -value is 0.008. Because the 95% lower confidence bound exceeds 0.82, we would reject the hypothesis that $H_0: \mu = 0.82$ and conclude that the alternative hypothesis $H_1: \mu > 0.82$ is true.

9.3.2 Type II Error and Choice of Sample Size

The type II error probability for the t -test depends on the distribution of the test statistic in Equation 9.26 when the null hypothesis $H_0: \mu = \mu_0$ is false. When the true value of the mean is $\mu = \mu_0 + \delta$, the distribution for T_0 is called the noncentral t distribution with $n - 1$ degrees of freedom and noncentrality parameter $\delta\sqrt{n/\sigma}$. Note that if $\delta = 0$, the noncentral t distribution reduces to the usual *central t distribution*. Therefore, the type II error of the two-sided alternative (for example) would be

$$\begin{aligned}\beta &= P(-t_{\alpha/2, n-1} \leq T_0 \leq t_{\alpha/2, n-1} \mid \delta \neq 0) \\ &= P(-t'_{\alpha/2, n-1} \leq T'_0 \leq t'_{\alpha/2, n-1})\end{aligned}$$

where T'_0 denotes the noncentral t random variable. Finding the type II error probability β for the t -test involves finding the probability contained between two points of the noncentral t distribution. Because the noncentral t -random variable has a messy density function, this integration must be done numerically.

Fortunately, this ugly task has already been done, and the results are summarized in a series of O.C. curves in Appendix Charts VIIe, VIIf, VIIg, and VIIh that plot β for the t -test against a parameter d for various sample sizes n . Curves are provided for two-sided alternatives on Charts VIIe and VIIf. The abscissa scale factor d on these charts is defined as

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} \quad (9.32)$$

For the one-sided alternative $\mu > \mu_0$ or $\mu < \mu_0$, we use charts VIIg and VIIh with

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} \quad (9.33)$$

We note that d depends on the unknown parameter σ^2 . We can avoid this difficulty in several ways. In some cases, we may use the results of a previous experiment or prior information to make a rough initial estimate of σ^2 . If we are interested in evaluating test performance after the data have been collected, we could use the sample variance s^2 to estimate σ^2 . If there is no previous experience on which to draw in estimating σ^2 , we then define the difference in the mean d that we wish to detect relative to σ . For example, if we wish to detect a small difference in the mean, we

might use a value of $d = |\delta|/\sigma \leq 1$ (for example), whereas if we are interested in detecting only moderately large differences in the mean, we might select $d = |\delta|/\sigma = 2$ (for example). That is, the value of the ratio $|\delta|/\sigma$ is important in determining sample size, and if it is possible to specify the relative size of the difference in means that we are interested in detecting, then a proper value of d can usually be selected.

EXAMPLE 9.7 | Golf Club Design Sample Size

Consider the golf club testing problem from Example 9.6. If the mean coefficient of restitution exceeds 0.82 by as much as 0.02, is the sample size $n = 15$ adequate to ensure that $H_0: \mu = 0.82$ will be rejected with probability at least 0.8?

To solve this problem, we use the sample standard deviation $s = 0.02456$ to estimate σ . Then $d = |\delta|/\sigma = 0.02/0.02456$

$= 0.81$. By referring to the operating characteristic curves in Appendix Chart VIIg (for $\alpha = 0.05$) with $d = 0.81$ and $n = 15$, we find that $\beta = 0.10$, approximately. Thus, the probability of rejecting $H_0: \mu = 0.82$ if the true mean exceeds this by 0.02 is approximately $1 - \beta = 1 - 0.10 = 0.90$, and we conclude that a sample size of $n = 15$ is adequate to provide the desired sensitivity.

Some software packages can also perform power and sample size computations for the one-sample t -test. Several calculations based on the golf club testing problem follow:

Power and Sample Size

1-Sample t -test

Testing mean = null (versus > null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 0.02456

Difference	Sample Size	Power
0.02	15	0.9117

Power and Sample Size

1-Sample t -test

Testing mean = null (versus > null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 0.02456

Difference	Sample Size	Power
0.01	15	0.4425

Power and Sample Size

1-Sample t -test

Testing mean = null (versus > null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 0.02456

Difference	Sample Size	Target Power	Actual Power
0.01	39	0.8000	0.8029

In the first portion of the computer output, the software reproduces the solution to Example 9.7, verifying that a sample size of $n = 15$ is adequate to give power of at least 0.8 if the mean coefficient of restitution exceeds 0.82 by at least 0.02. In the middle section of the output, we used the software to compute the power to detect the difference between μ and $\mu_0 = 0.82$ of 0.01. Notice that with $n = 15$, the power drops considerably to 0.4425. The final portion of the output is the

sample size required for a power of at least 0.8 if the difference between μ and μ_0 of interest is actually 0.01. A much larger n is required to detect this smaller difference.

9.4 Tests on the Variance and Standard Deviation of a Normal Distribution

Sometimes hypothesis tests on the population variance or standard deviation are needed. When the population is modeled by a normal distribution, the tests and intervals described in this section are applicable.

9.4.1 Hypothesis Tests on the Variance

Suppose that we wish to test the hypothesis that the variance of a normal population σ^2 equals a specified value, say σ_0^2 , or equivalently, that the standard deviation σ is equal to σ_0 . Let X_1, X_2, \dots, X_n be a random sample of n observations from this population. To test

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 \neq \sigma_0^2 \quad (9.34)$$

we use the test statistic:

Test Statistic

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2} \quad (9.35)$$

If the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ is true, the test statistic χ_0^2 defined in Equation 9.35 follows the chi-square distribution with $n-1$ degrees of freedom. This is the reference distribution for this test procedure. To perform a fixed significance level test, we would take a random sample from the population of interest, calculate χ_0^2 , the value of the test statistic χ_0^2 , and the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ would be rejected if

$$\chi_0^2 > \chi_{\alpha/2, n-1}^2 \quad \text{or if} \quad \chi_0^2 < \chi_{1-\alpha/2, n-1}^2$$

where $\chi_{\alpha/2, n-1}^2$ and $\chi_{1-\alpha/2, n-1}^2$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively. Figure 9.17(a) shows the critical region.

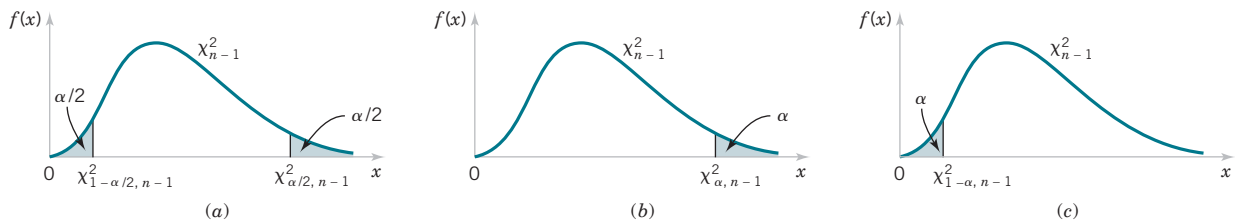


FIGURE 9.17

Reference distribution for the test of $H_0: \sigma^2 = \sigma_0^2$ with critical region values for (a) $H_1: \sigma^2 \neq \sigma_0^2$, (b) $H_1: \sigma^2 > \sigma_0^2$, and (c) $H_1: \sigma^2 < \sigma_0^2$.

The same test statistic is used for one-sided alternative hypotheses. For the one-sided hypotheses

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 > \sigma_0^2 \quad (9.36)$$

we would reject H_0 if $\chi_0^2 > \chi_{\alpha, n-1}^2$, whereas for the other one-sided hypotheses

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 < \sigma_0^2 \quad (9.37)$$

we would reject H_0 if $\chi_0^2 < \chi_{1-\alpha, n-1}^2$. The one-sided critical regions are shown in Figure 9.17(b) and (c).

Tests on the Variance of a Normal Distribution

Null hypothesis: $H_0: \sigma^2 = \sigma_0^2$

Test statistic: $\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$

Alternative Hypothesis	Rejection Criteria
$H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi_{\alpha/2, n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2, n-1}^2$
$H_1: \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{\alpha, n-1}^2$
$H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi_{1-\alpha, n-1}^2$

EXAMPLE 9.8 | Automated Filling

An automated filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ (fluid ounces)². If the variance of fill volume exceeds 0.01 (fluid ounces)², an unacceptable proportion of bottles will be underfilled or overfilled. Is there evidence in the sample data to suggest that the manufacturer has a problem with underfilled or overfilled bottles? Use $\alpha = 0.05$, and assume that fill volume has a normal distribution.

Using the seven-step procedure results in the following:

- Parameter of interest:** The parameter of interest is the population variance σ^2 .
- Null hypothesis:** $H_0: \sigma^2 = 0.01$

- Alternative hypothesis:** $H_1: \sigma^2 > 0.01$

- Test statistic:** The test statistic is $\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$

- Reject H_0 if:** Use $\alpha = 0.05$, and reject H_0 if $\chi_0^2 > \chi_{0.05, 19}^2 = 30.14$

- Computations:** $\chi_0^2 = \frac{19(0.0153)}{0.01} = 29.07$

- Conclusions:** Because $\chi_0^2 = 29.07 < \chi_{0.05, 19}^2 = 30.14$, we conclude that there is no strong evidence that the variance of fill volume exceeds 0.01 (fluid ounces)². So there is no strong evidence of a problem with incorrectly filled bottles.

We can also use the P -value approach. Using Appendix Table III, it is easy to place bounds on the P -value of a **chi-square test**. From inspection of the table, we find that $\chi_{0.10, 19}^2 = 27.20$ and $\chi_{0.05, 19}^2 = 30.14$. Because $27.20 < 29.07 < 30.14$, we conclude that the P -value for the test in Example 9.8 is in the interval $0.05 < P\text{-value} < 0.10$.

The P -value for a lower-tailed test would be found as the area (probability) in the lower tail of the chi-square distribution to the left of (or below) the computed value of the test statistic χ_0^2 . For the two-sided alternative, find the tail area associated with the computed value of the test statistic and double it to obtain the P -value.

Some software packages perform the test on a variance of a normal distribution described in this section. Typical computer output for Example 9.8 is as follows:

Test and CI for One Variance		
Null hypothesis		Sigma-squared = 0.01
Alternative hypothesis		Sigma-squared > 0.01
Statistics		
N	StDev	Variance
20	0.124	0.0153
95% One-Sided Confidence Intervals		
Lower Confidence Bound for StDev		Lower Confidence Bound for Variance
0.098		0.0096
Tests		
Chi-Square	DF	P-Value
29.07	19	0.065

Recall that we said that the t -test is relatively robust to the assumption that we are sampling from a normal distribution. The same is not true for the chi-square test on variance. Even moderate departures from normality can result in the test statistic in Equation 9.35 having a distribution that is very different from chi-square.

9.4.2

Type II Error and Choice of Sample Size

Operating characteristic curves for the chi-square tests in Section 9.4.1 are in Appendix Charts VII through VI*n* for $\alpha = 0.05$ and $\alpha = 0.01$. For the two-sided alternative hypothesis of Equation 9.34, Charts VII*i* and VII*j* plot β against an abscissa parameter

$$\lambda = \frac{\sigma}{\sigma_0}$$

(9.38)

for various sample sizes n , where σ denotes the true value of the standard deviation. Charts VII*k* and VII*l* are for the one-sided alternative $H_1: \sigma^2 > \sigma_0^2$, and Charts VII*m* and VII*n* are for the other one-sided alternative $H_1: \sigma^2 < \sigma_0^2$. In using these charts, we think of σ as the value of the standard deviation that we want to detect.

These curves can be used to evaluate the β -error (or power) associated with a particular test. Alternatively, they can be used to *design* a test—that is, to determine what sample size is necessary to detect a particular value of σ that differs from the hypothesized value σ_0 .

EXAMPLE 9.9

Automated Filling Sample Size

Consider the bottle-filling problem from Example 9.8. If the variance of the filling process exceeds 0.01 (fluid ounces)², too many bottles will be underfilled. Thus, the hypothesized value of the standard deviation is $\sigma_0 = 0.10$. Suppose that if the true standard deviation of the filling process exceeds this value by 25%, we would like to detect this with probability at least 0.8. Is the sample size of $n = 20$ adequate?

To solve this problem, note that we require

$$\lambda = \frac{\sigma}{\sigma_0} = \frac{0.125}{0.10} = 1.25$$

This is the abscissa parameter for Chart VII*k*. From this chart, with $n = 20$ and $\lambda = 1.25$, we find that $\beta \approx 0.6$. Therefore, there is only about a 40% chance that the null hypothesis will be rejected if the true standard deviation is really as large as $\sigma = 0.125$ fluid ounce.

To reduce the β -error, a larger sample size must be used. From the operating characteristic curve with $\beta = 0.20$ and $\lambda = 1.25$, we find that $n = 75$, approximately. Thus, if we want the test to perform as required, the sample size must be at least 75 bottles.

9.5 Tests on a Population Proportion

It is often necessary to test hypotheses on a population proportion. For example, suppose that a random sample of size n has been taken from a large (possibly infinite) population and that $X(\leq n)$ observations in this sample belong to a class of interest. Then $\hat{P} = X/n$ is a point estimator of the proportion of the population p that belongs to this class. Note that n and p are the parameters of a binomial distribution. Furthermore, from Chapter 7, we know that the sampling distribution of \hat{P} is approximately normal with mean p and variance $p(1-p)/n$ if p is not too close to either 0 or 1 and if n is relatively large. Typically, to apply this approximation we require that np and $n(1-p)$ be greater than or equal to 5. We provide a large-sample test that uses the normal approximation to the binomial distribution.

9.5.1 Large-Sample Tests on a Proportion

Many engineering problems concern a random variable that follows the binomial distribution. For example, consider a production process that manufactures items that are classified as either acceptable or defective. Modelling the occurrence of defectives with the binomial distribution is usually reasonable when the binomial parameter p represents the proportion of defective items produced. Consequently, many engineering decision problems involve hypothesis testing about p .

We consider testing

$$H_0: p = p_0 \qquad H_1: p \neq p_0 \qquad (9.39)$$

An approximate test based on the normal approximation to the binomial is given. As noted earlier, this approximate procedure will be valid as long as p is not extremely close to 0 or 1, and if the sample size is relatively large. Let X be the number of observations in a random sample of size n that belongs to the class associated with p . Then if the null hypothesis $H_0: p = p_0$ is true, we have $X \sim N[np_0, np_0(1-p_0)]$, approximately. To test $H_0: p = p_0$, calculate the test statistic

Test Statistic

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \qquad (9.40)$$

and determine the P -value. Because the test statistic follows a standard normal distribution if H_0 is true, the P -value is calculated exactly like the P -value for the z -tests in Section 9.2. So for the two-sided alternative hypothesis, the P -value is the sum of the probability in the standard normal distribution above $|z_0|$ and the probability below the negative value $-|z_0|$, or

$$P = 2 [1 - \Phi(|z_0|)]$$

For the one-sided alternative hypothesis $H_1: p > p_0$, the P -value is the probability above z_0 , or

$$P = 1 - \Phi(z_0)$$

and for the one-sided alternative hypothesis $H_1: p < p_0$, the P -value is the probability below z_0 , or

$$P = \Phi(z_0)$$

We can also perform a fixed-significance-level test. For the two-sided alternative hypothesis, we would reject $H_0: p = p_0$ if

$$z_0 > z_{\alpha/2} \quad \text{or} \quad z_0 < -z_{\alpha/2}$$

Critical regions for the one-sided alternative hypotheses would be constructed in the usual manner.

Summary of Approximate Tests on a Binomial Proportion

Testing Hypotheses on a Binomial Proportion

Null hypotheses: $H_0: p = p_0$

Test statistic:
$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: p \neq p_0$	Probability above $ z_0 $ and probability below $- z_0 $, $P = 2[1 - \Phi(z_0)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p > p_0$	Probability above z_0 , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p < p_0$	Probability below z_0 , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

EXAMPLE 9.10 | Automobile Engine Controller

A semiconductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout or fraction defective at a critical manufacturing step not exceed 0.05 and that the manufacturer demonstrate process capability at this level of quality using $\alpha = 0.05$. The semiconductor manufacturer takes a random sample of 200 devices and finds that four of them are defective. Can the manufacturer demonstrate process capability for the customer?

We may solve this problem using the seven-step hypothesis-testing procedure as follows:

- Parameter of interest:** The parameter of interest is the process fraction defective p .
- Null hypothesis:** $H_0: p = 0.05$
- Alternative hypothesis:** $H_1: p < 0.05$

This formulation of the problem will allow the manufacturer to make a strong claim about process

capability if the null hypothesis $H_0: p = 0.05$ is rejected.

- Test statistic:** The test statistic is (from Equation 9.40):

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

where $x = 4$, $n = 200$, and $p_0 = 0.05$.

- Reject H_0 if:** Reject $H_0: p = 0.05$ if the p -value is less than 0.05.

- Computation:** The test statistic is

$$z_0 = \frac{4 - 200(0.05)}{\sqrt{200(0.05)(0.95)}} = -1.95$$

- Conclusions:** Because $z_0 = -1.95$, the P -value is $\Phi(-1.95) = 0.0256$, so we reject H_0 and conclude that the process fraction defective p is less than 0.05.

Practical Interpretation: We conclude that the process is capable.

Another form of the test statistic Z_0 in Equation 9.40 is occasionally encountered. Note that if X is the number of observations in a random sample of size n that belongs to a class of interest, then $\hat{P} = X/n$ is the sample proportion that belongs to that class. Now divide both numerator and denominator of Z_0 in Equation 9.40 by n , giving

$$Z_0 = \frac{X/n - p_0}{\sqrt{p_0(1 - p_0)/n}} \quad \text{or} \quad Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)/n}} \quad (9.41)$$

This presents the test statistic in terms of the sample proportion instead of the number of items X in the sample that belongs to the class of interest.

Computer software packages can be used to perform the test on a binomial proportion. The following output shows typical results for Example 9.10.

Test and CI for One Proportion

Test of $p = 0.05$ vs $p < 0.05$

Sample	X	N	Sample p	95% Upper Confidence Bound	Z-Value	P-Value
1	4	200	0.020000	0.036283	-1.95	0.026

This output also shows a 95% one-sided upper-confidence bound on P . In Section 8.4, we showed how CIs on a binomial proportion are computed. This display shows the result of using the normal approximation for tests and CIs. When the sample size is small, this may be inappropriate.

Small Sample Tests on a Binomial Proportion Tests on a proportion when the sample size n is small are based on the binomial distribution, not the normal approximation to the binomial. To illustrate, suppose that we wish to test $H_0: p = p_0$ versus $H_0: p < p_0$. Let X be the number of successes in the sample. The P -value for this test would be found from the lower tail of a binomial distribution with parameters n and p_0 . Specifically, the P -value would be the probability that a binomial random variable with parameters n and p_0 is less than or equal to X . P -values for the upper-tailed one-sided test and the two-sided alternative are computed similarly.

Many software packages calculate the exact P -value for a binomial test. The following output contains the exact P -value results for Example 9.10.

Test of $p = 0.05$ vs $p < 0.05$

Sample	X	N	Sample p	95% Upper Confidence Bound	Exact P-Value
1	4	200	0.020000	0.045180	0.026

The P -value is the same as that reported for the normal approximation because the sample size is fairly large. Notice that the CI is different from the one found using the normal approximation.

9.5.2 Type II Error and Choice of Sample Size

It is possible to obtain closed-form equations for the approximate β -error for the tests in Section 9.5.1. Suppose that p is the true value of the population proportion. The approximate β -error for the two-sided alternative $H_1: p \neq p_0$ is

$$\beta = \Phi\left(\frac{p_0 - p + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) - \Phi\left(\frac{p_0 - p - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) \quad (9.42)$$

If the alternative is $H_1: p < p_0$,

$$\beta = 1 - \Phi\left(\frac{p_0 - p - z_{\alpha}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) \quad (9.43)$$

whereas if the alternative is $H_1: p > p_0$,

$$\beta = \Phi \left(\frac{p_0 - p + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}} \right) \quad (9.44)$$

These equations can be solved to find the approximate sample size n that gives a test of level α that has a specified β risk. The sample size equations are

Approximate Sample Size for a Two-Sided Test on a Binomial Proportion

$$n = \left[\frac{z_{\alpha/2} \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{p - p_0} \right]^2 \quad (9.45)$$

for a two-sided alternative and for a one-sided alternative:

Approximate Sample Size for a One-Sided Test on a Binomial Proportion

$$n = \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{p - p_0} \right]^2 \quad (9.46)$$

EXAMPLE 9.11 | Automobile Engine Controller Type II Error

Consider the semiconductor manufacturer from Example 9.10. Suppose that its process fallout is really $p = 0.03$. What is the β -error for a test of process capability that uses $n = 200$ and $\alpha = 0.05$?

The β -error can be computed using Equation 9.43 as follows:

$$\begin{aligned} \beta &= 1 - \Phi \left[\frac{0.05 - 0.03 - (1.645)\sqrt{0.05(0.95)/200}}{\sqrt{0.03(1-0.03)/200}} \right] \\ &= 1 - \Phi(-0.44) = 0.67 \end{aligned}$$

Thus, the probability is about 0.7 that the semiconductor manufacturer will fail to conclude that the process is capable if the true process fraction defective is $p = 0.03$ (3%). That is, the power of the test against this particular alternative is only

about 0.3. This appears to be a large β -error (or small power), but the difference between $p = 0.05$ and $p = 0.03$ is fairly small, and the sample size $n = 200$ is not particularly large.

Suppose that the semiconductor manufacturer was willing to accept a β -error as large as 0.10 if the true value of the process fraction defective was $p = 0.03$. If the manufacturer continues to use $\alpha = 0.05$, what sample size would be required?

The required sample size can be computed from Equation 9.46 as follows:

$$n = \left[\frac{1.645\sqrt{0.05(0.95)} + 1.28\sqrt{0.03(0.97)}}{0.03 - 0.05} \right]^2 \approx 832$$

where we have used $p = 0.03$ in Equation 9.46.

Conclusion: Note that $n = 832$ is a very large sample size. However, we are trying to detect a fairly small deviation from the null value $p_0 = 0.05$.

Some software packages also perform power and sample size calculations for the one-sample Z-test on a proportion. Typical computer output for the engine controllers tested in Example 9.10 follows.

Power and Sample Size

Test for One Proportion

Testing proportion = 0.05 (versus < 0.05)

Alpha = 0.05

Alternative Proportion	Sample Size	Power
3.00E-02	200	0.3287

Power and Sample Size

Test for One Proportion

Testing proportion = 0.05 (versus < 0.05)

Alpha = 0.05

Alternative Proportion	Sample Size	Target Power	Actual Power
3.00E-02	833	0.9000	0.9001

Power and Sample Size

Test for One Proportion

Testing proportion = 0.05 (versus < 0.05)

Alpha = 0.05

Alternative Proportion	Sample Size	Target Power	Actual Power
3.00E-02	561	0.7500	0.75030

The first part of the output shows the power calculation based on the situation described in Example 9.11 where the true proportion is really 0.03. The computer power calculation agrees with the results from Equation 9.43 in Example 9.11. The second part of the output computes the sample size necessary for a power of 0.9 ($\beta = 0.1$) if $p = 0.03$. Again, the results agree closely with those obtained from Equation 9.46. The final portion of the display shows the sample size that would be required if $p = 0.03$ and the power requirement is relaxed to 0.75. Notice that the sample size of $n = 561$ is still quite large because the difference between $p = 0.05$ and $p = 0.03$ is fairly small.

9.6 Summary Table of Inference Procedures for a Single Sample

The table in the end papers of this book (inside back cover) presents a summary of all the single-sample inference procedures from Chapters 8 and 9. The table contains the null hypothesis statement, the test statistic, the various alternative hypotheses and the criteria for rejecting H_0 , and the formulas for constructing the $100(1 - \alpha)\%$ two-sided confidence interval. It would also be helpful to refer to the roadmap table in Chapter 8 that provides guidance to match the problem type to the information inside the back cover.

9.7 Testing for Goodness of Fit

The hypothesis-testing procedures that we have discussed in previous sections are designed for problems in which the population or probability distribution is known and the hypotheses involve the parameters of the distribution. Another kind of hypothesis is often encountered: We do not know the underlying distribution of the population, and we wish to test the hypothesis that a

particular distribution will be satisfactory as a population model. For example, we might wish to test the hypothesis that the population is normal.

We have previously discussed a very useful graphical technique for this problem called **probability plotting** and illustrated how it was applied in the case of a normal distribution. In this section, we describe a formal **goodness-of-fit test** procedure based on the chi-square distribution.

The test procedure requires a random sample of size n from the population whose probability distribution is unknown. These n observations are arranged in a frequency histogram, having k bins or class intervals. Let O_i be the observed frequency in the i th class interval. From the hypothesized probability distribution, we compute the expected frequency in the i th class interval, denoted E_i . The test statistic is

Goodness-of-Fit Test Statistic

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (9.47)$$

It can be shown that, if the population follows the hypothesized distribution, χ_0^2 has, approximately, a chi-square distribution with $k - p - 1$ degrees of freedom, when p represents the number of parameters of the hypothesized distribution estimated by sample statistics. This approximation improves as n increases. We should reject the null hypothesis that the population is the hypothesized distribution if the test statistic is too large. Therefore, the P -value would be the probability under the chi-square distribution with $k - p - 1$ degrees of freedom above the computed value of the test statistic χ_0^2 or $P = P(\chi_{k-p-1}^2 > \chi_0^2)$. For a fixed-level test, we would reject the hypothesis that the distribution of the population is the hypothesized distribution if the calculated value of the test statistic $\chi_0^2 > \chi_{\alpha, k-p-1}^2$.

One point to be noted in the application of this test procedure concerns the magnitude of the expected frequencies. If these expected frequencies are too small, the test statistic χ_0^2 will not reflect the departure of observed from expected but only the small magnitude of the expected frequencies. There is no general agreement regarding the minimum value of expected frequencies, but values of 3, 4, and 5 are widely used as minimal. Some writers suggest that an expected frequency could be as small as 1 or 2 so long as most of them exceed 5. Should an expected frequency be too small, it can be combined with the expected frequency in an adjacent class interval. The corresponding observed frequencies would then also be combined, and k would be reduced by 1. Class intervals are not required to be of equal width.

We now give two examples of the test procedure.

EXAMPLE 9.12 | Printed Circuit Board Defects—Poisson Distribution

The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed circuit boards has been collected, and the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

The mean of the assumed Poisson distribution in this example is unknown and must be estimated from the sample data. The

estimate of the mean number of defects per board is the sample average, that is, $(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3)/60 = 0.75$. From the Poisson distribution with parameter 0.75, we may compute p_i , the theoretical, hypothesized probability associated with the i th class interval. Because each class interval corresponds to a particular number of defects, we may find the p_i as follows:

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

The expected frequencies are computed by multiplying the sample size $n = 60$ times the probabilities p_i . That is, $E_i = np_i$. The expected frequencies follow:

Number of Defects	Probability	Expected Frequency
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (or more)	0.041	2.46

Because the expected frequency in the last cell is less than 3, we combine the last two cells:

Number of Defects	Observed Frequency	Expected Frequency
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

The seven-step hypothesis-testing procedure may now be applied, using $\alpha = 0.05$, as follows:

- Parameter of interest:** The variable of interest is the form of the distribution of defects in printed circuit boards.

- Null hypothesis:** H_0 : The form of the distribution of defects is Poisson.

- Alternative hypothesis:** H_1 : The form of the distribution of defects is not Poisson.

- Test statistic:** The test statistic is $\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$

- Reject H_0 if:** Because the mean of the Poisson distribution was estimated, the preceding chi-square statistic will have $k - p - 1 = 3 - 1 - 1 = 1$ degree of freedom. Consider whether the P -value is less than 0.05.

- Computations:**

$$\chi_0^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94$$

- Conclusions:** We find from Appendix Table III that $\chi_{0.10,1}^2 = 2.71$ and $\chi_{0.05,1}^2 = 3.84$. Because $\chi_0^2 = 2.94$ lies between these values, we conclude that the P -value is between 0.05 and 0.10. Therefore, because the P -value exceeds 0.05, we are unable to reject the null hypothesis that the distribution of defects in printed circuit boards is Poisson. The exact P -value computed from software is 0.0864.

EXAMPLE 9.13 | Power Supply Distribution—Continuous Distribution

A manufacturing engineer is testing a power supply used in a notebook computer and, using $\alpha = 0.05$, wishes to determine whether output voltage is adequately described by a normal distribution. Sample estimates of the mean and standard deviation of $\bar{x} = 5.04$ V and $s = 0.08$ V are obtained from a random sample of $n = 100$ units.

A common practice in constructing the class intervals for the frequency distribution used in the chi-square goodness-of-fit test is to choose the cell boundaries so that the expected frequencies $E_i = np_i$ are equal for all cells. To use this method, we want to choose the cell boundaries a_0, a_1, \dots, a_k for the k cells so that all the probabilities

$$p_i = P(a_{i-1} \leq X \leq a_i) = \int_{a_{i-1}}^{a_i} f(x) dx$$

are equal. Suppose that we decide to use $k = 8$ cells. For the standard normal distribution, the intervals that divide the scale into eight equally likely segments are $(0, 0.32)$, $(0.32, 0.675)$, $(0.675, 1.15)$, $(1.15, \infty)$, and their four “mirror image” intervals on the other side of zero. For each interval $p_i = 1/8 = 0.125$, so the expected cell frequencies are $E_i = n_{pi} = 100(0.125) = 12.5$. The complete table of observed and expected frequencies is as follows:

Class Interval	Observed Frequency o_i	Expected Frequency E_i
$x < 4.948$	12	12.5
$4.948 \leq x < 4.986$	14	12.5
$4.986 \leq x < 5.014$	12	12.5
$5.014 \leq x < 5.040$	13	12.5
$5.040 \leq x < 5.066$	12	12.5
$5.066 \leq x < 5.094$	11	12.5
$5.094 \leq x < 5.132$	12	12.5
$5.132 \leq x$	14	12.5
Totals	100	100

The boundary of the first class interval is $\bar{x} - 1.15s = 4.948$. The second class interval is $[\bar{x} - 1.15s, \bar{x} - 0.675s]$ and so forth. We may apply the seven-step hypothesis-testing procedure to this problem.

- Parameter of interest:** The variable of interest is the form of the distribution of power supply voltage.
- Null hypothesis:** H_0 : The form of the distribution is normal.

3. **Alternative hypothesis:** H_1 : The form of the distribution is nonnormal.

4. **Test statistic:** The test statistic is

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

5. **Reject H_0 if:** Because two parameters in the normal distribution have been estimated, the preceding chi-square statistic will have $k - p - 1 = 8 - 2 - 1 = 5$ degrees of freedom. We use a fixed significance level test with $\alpha = 0.05$. Therefore, we will reject H_0 if $\chi_0^2 > \chi_{0.05,5}^2 = 11.07$.

6. **Computations:**

$$\begin{aligned} \chi_0^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(12 - 12.5)^2}{12.5} + \frac{(14 - 12.5)^2}{12.5} \\ &\quad + \cdots + \frac{(14 - 12.5)^2}{12.5} = 0.64 \end{aligned}$$

7. **Conclusions:** Because $\chi_0^2 = 0.64 < \chi_{0.05,5}^2 = 11.07$, we are unable to reject H_0 , and no strong evidence indicates that output voltage is not normally distributed. The P -value for the chi-square statistic $\chi_0^2 = 0.64$ is $p = 0.9861$.

9.8 Contingency Table Tests

Many times the n elements of a sample from a population may be classified according to two different criteria. It is then of interest to know whether the two methods of classification are statistically **independent**; for example, we may consider the population of graduating engineers and may wish to determine whether starting salary is independent of academic disciplines. Assume that the first method of classification has r levels and that the second method has c levels. We will let O_{ij} be the observed frequency for level i of the first classification method and level j of the second classification method. The data would, in general, appear as shown in Table 9.2. Such a table is usually called an $r \times c$ **contingency table**.

We are interested in testing the hypothesis that the row-and-column methods of classification are independent. If we reject this hypothesis, we conclude some interaction exists between the two criteria of classification. The exact test procedures are difficult to obtain, but an approximate test statistic is valid for large n . Let p_{ij} be the probability that a randomly selected element falls in the ij th cell given that the two classifications are independent. Then $p_{ij} = u_i v_j$, where u_i is the probability that a randomly selected element falls in row class i and v_j is the probability that a randomly selected element falls in column class j . Now by assuming independence, the estimators of u_i and v_j are

$$\hat{u}_i = \frac{1}{n} \sum_{j=1}^c O_{ij} \qquad \hat{v}_j = \frac{1}{n} \sum_{i=1}^r O_{ij} \qquad (9.48)$$

Therefore, the expected frequency of each cell is

$$E_{ij} = n \hat{u}_i \hat{v}_j = \frac{1}{n} \sum_{j=1}^c O_{ij} \sum_{i=1}^r O_{ij} \qquad (9.49)$$

TABLE 9.2 An $r \times c$ Contingency Table

		Columns			
		1	2	...	c
Rows	1	O_{11}	O_{12}	...	O_{1c}
	2	O_{21}	O_{22}	...	O_{2c}
	\vdots	\vdots	\vdots	\vdots	\vdots
	r	O_{r1}	O_{r2}	...	O_{rc}

Then, for large n , the statistic

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (9.50)$$

has an approximate chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom if the null hypothesis is true. We should reject the null hypothesis if the value of the test statistic χ_0^2 is too large. The P -value would be calculated as the probability beyond χ_0^2 on the $\chi_{(r-1)(c-1)}^2$ distribution, or $P = P(\chi_{(r-1)(c-1)}^2 > \chi_0^2)$. For a fixed-level test, we would reject the hypothesis of independence if the observed value of the test statistic χ_0^2 exceeded $\chi_{\alpha, (r-1)(c-1)}^2$.

EXAMPLE 9.14 | Health Insurance Plan Preference

A company has to choose among three health insurance plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$. The opinions of a random sample of 500 employees are shown in Table 9.3.

TABLE 9.3 Observed Data for Example 9.14

Health Insurance Plan				
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	<u>40</u>	<u>60</u>	<u>60</u>	<u>160</u>
Totals	200	200	100	500

To find the expected frequencies, we must first compute $\hat{u}_1 = (340/500) = 0.68$, $\hat{u}_2 = (160/500) = 0.32$, $\hat{v}_1 = (200/500) = 0.40$, $\hat{v}_2 = (200/500) = 0.40$, and $\hat{v}_3 = (100/500) = 0.20$. The expected frequencies may now be computed from Equation 9.49. For example, the expected number of salaried workers favoring health insurance plan 1 is

$$E_{11} = n\hat{u}_1\hat{v}_1 = 500(0.68)(0.40) = 136$$

The expected frequencies are shown in Table 9.4.

TABLE 9.4 Expected Frequencies for Example 9.14

Health Insurance Plan				
Job Classification	1	2	3	Totals
Salaried workers	136	136	68	340
Hourly workers	<u>64</u>	<u>64</u>	<u>32</u>	<u>160</u>
Totals	200	200	100	500

The seven-step hypothesis-testing procedure may now be applied to this problem.

- Parameter of interest:** The variable of interest is employee preference among health insurance plans.
- Null hypothesis:** H_0 : Preference is independent of salaried versus hourly job classification.
- Alternative hypothesis:** H_1 : Preference is not independent of salaried versus hourly job classification.
- Test statistic:** The test statistic is

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Reject H_0 if:** We will use a fixed-significance level test with $\alpha = 0.05$. Therefore, because $r = 2$ and $c = 3$, the degrees of freedom for chi-square are $(r - 1)(c - 1) = (1)(2) = 2$, and we would reject H_0 if $\chi_0^2 = \chi_{0.05, 2}^2 = 5.99$.
- Computations:**

$$\begin{aligned} \chi_0^2 &= \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(160 - 136)^2}{136} + \frac{(140 - 136)^2}{136} + \frac{(40 - 68)^2}{68} \\ &\quad + \frac{(40 - 64)^2}{64} + \frac{(60 - 64)^2}{64} + \frac{(60 - 32)^2}{32} \\ &= 49.63 \end{aligned}$$

- Conclusions:** Because $\chi_0^2 = 49.63 > \chi_{0.05, 2}^2 = 5.99$, we reject the hypothesis of independence and conclude that the preference for health insurance plans is not independent of job classification. The P -value for $\chi_0^2 = 49.63$ is $P = 1.671 \times 10^{-11}$. (This value was computed by computer software.) Further analysis would be necessary to explore the nature of the association between these factors. It might be helpful to examine the table of observed minus expected frequencies.

Using the two-way contingency table to test independence between two variables of classification in a sample from a single population of interest is only one application of contingency table methods. Another common situation occurs when there are r populations of interest and each population is divided into the same c categories. A sample is then taken from the i th population, and the counts are entered in the appropriate columns of the i th row. In this situation, we want to investigate whether or not the proportions in the c categories are the same for all populations. The null hypothesis in this problem states that the populations are **homogeneous** with respect to the categories. For example, with only two categories, such as success and failure, defective and non-defective, and so on, the test for homogeneity is really a test of the equality of r binomial parameters. Calculation of expected frequencies, determination of degrees of freedom, and computation of the chi-square statistic for the test for homogeneity are identical to the test for independence.

9.9 Nonparametric Procedures

Most of the hypothesis-testing and confidence interval procedures discussed previously are based on the assumption that we are working with random samples from normal populations. Traditionally, we have called these procedures **parametric** methods because they are based on a particular parametric family of distributions—in this case, the normal. Alternately, sometimes we say that these procedures are not distribution free because they depend on the assumption of normality. Fortunately, most of these procedures are relatively insensitive to moderate departures from normality. In general, the t - and F -tests and the t -confidence intervals will have actual levels of significance or confidence levels that differ from the nominal or advertised levels chosen by the experimenter, although the difference in the actual and advertised levels is usually fairly small when the underlying population is not too different from the normal.

In this section, we describe procedures called *nonparametric* and *distribution-free methods*, and we usually make no assumptions about the distribution of the underlying population other than that it is continuous. These procedures have an accurate level of significance α or confidence level $100(1 - \alpha)\%$ for many different types of distributions. These procedures have some appeal. One of their advantages is that the data need not be quantitative but can be categorical (such as yes or no, defective or nondefective) or rank data. Another advantage is that nonparametric procedures are usually very quick and easy to perform.

The procedures described in this section are alternatives to the parametric t - and F -procedures described earlier. Consequently, it is important to compare the performance of both parametric and nonparametric methods under the assumptions of both normal and nonnormal populations. In general, nonparametric procedures do not utilize all the information provided by the sample. As a result, a nonparametric procedure will be less efficient than the corresponding parametric procedure when the underlying population is normal. This loss of efficiency is reflected by a requirement of a larger sample size for the nonparametric procedure than would be required by the parametric procedure to achieve the same power. On the other hand, this loss of efficiency is usually not large, and often the difference in sample size is very small. When the underlying distributions are not close to normal, nonparametric methods may have much to offer. They often provide improvement over the normal-theory parametric methods. Generally, if both parametric and nonparametric methods are applicable to a particular problem, we should use the more efficient parametric procedure.

Another approach is to **transform** the original data, say, by taking logarithms, square roots, or a reciprocal, and then analyze the transformed data using a parametric technique. A normal probability plot often works well to see whether the transformation has been successful. When this approach is successful, it is usually preferable to using a nonparametric technique. However, sometimes transformations are not satisfactory. That is, no transformation makes the sample observations look very close to a sample from a normal distribution. One situation in which this happens is when the data are in the form of **ranks**. These situations frequently occur in practice. For instance, a panel of judges may be used to evaluate 10 different formulations of a soft-drink beverage for overall quality with the “best” formulation assigned rank 1, the

“next-best” formulation assigned rank 2, and so forth. It is unlikely that rank data satisfy the normality assumption. Transformations may not prove satisfactory either. Many nonparametric methods involve the analysis of ranks and consequently are directly suited to this type of problem.

9.9.1 The Sign Test

The **sign test** is used to test hypotheses about the **median** $\tilde{\mu}$ of a continuous distribution. The median of a distribution is a value of the random variable X such that the probability is 0.5 that an observed value of X is less than or equal to the median, and the probability is 0.5 that an observed value of X is greater than or equal to the median. That is, $P(X \leq \tilde{\mu}) = P(X \geq \tilde{\mu}) = 0.5$.

Because the normal distribution is symmetric, the mean of a normal distribution equals the median. Therefore, the sign test can be used to test hypotheses about the mean of a normal distribution. This is the same problem for which we previously used the t -test. We briefly discuss the relative merits of the two procedures in Section 9.9.3. Note that, although the t -test was designed for samples from a normal distribution, the sign test is appropriate for samples from any continuous distribution. Thus, the sign test is a nonparametric procedure.

Suppose that the hypotheses are

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \qquad H_1: \tilde{\mu} < \tilde{\mu}_0 \qquad (9.51)$$

The test procedure is easy to describe. Suppose that X_1, X_2, \dots, X_n is a random sample from the population of interest. Form the differences

$$X_i - \tilde{\mu}_0 \qquad i = 1, 2, \dots, n \qquad (9.52)$$

Now if the null hypothesis $H_0: \tilde{\mu} = \tilde{\mu}_0$ is true, any difference $X_i - \tilde{\mu}_0$ is equally likely to be positive or negative. An appropriate test statistic is the number of these differences that are positive, say, R^+ . Therefore, to test the null hypothesis, we are really testing that the number of plus signs is a value of a binomial random variable that has the parameter $p = 1/2$. A P -value for the observed number of plus signs r^+ can be calculated directly from the binomial distribution. For instance, in testing the hypotheses in Equation 9.51, we will reject H_0 in favor of H_1 only if the proportion of plus signs is sufficiently less than $1/2$ (or equivalently, when the observed number of plus signs r^+ is too small). Thus, if the computed P -value

$$P = P(R^+ \leq r^+ \text{ when } p = \frac{1}{2})$$

is less than or equal to some preselected significance level α , we will reject H_0 and conclude that H_1 is true.

To test the other one-sided hypotheses

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \qquad H_1: \tilde{\mu} > \tilde{\mu}_0 \qquad (9.53)$$

we will reject H_0 in favor of H_1 only if the observed number of plus signs, say, r^+ , is large or, equivalently, when the observed fraction of plus signs is significantly greater than $1/2$. Thus, if the computed P -value

$$P = P(R^+ \geq r^+ \text{ when } p = \frac{1}{2})$$

is less than α , we will reject H_0 and conclude that H_1 is true.

The two-sided alternative may also be tested. If the hypotheses are

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \qquad H_1: \tilde{\mu} \neq \tilde{\mu}_0 \qquad (9.54)$$

we should reject $H_0: \tilde{\mu} = \tilde{\mu}_0$ if the proportion of plus signs is significantly different from (either less than or greater than) $1/2$. This is equivalent to the observed number of plus signs r^+ being either sufficiently large or sufficiently small. Thus, if $r^+ < n/2$, the P -value is

$$P = 2P(R^+ \leq r^+ \text{ when } p = \frac{1}{2})$$

and if $r^+ > n/2$, the P -value is

$$P = 2P\left(R^+ \geq r^+ \text{ when } p = \frac{1}{2}\right)$$

If the P -value is less than some preselected level α , we will reject H_0 and conclude that H_1 is true.

EXAMPLE 9.15 | Propellant Shear Strength Sign Test

Montgomery, Peck, and Vining (2012) reported on a study in which a rocket motor is formed by binding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two propellant types is an important characteristic. The results of testing 20 randomly selected motors are shown in Table 9.5. We would like to test the hypothesis that the median shear strength is 2000 psi, using $\alpha = 0.05$.

TABLE 9.5 Propellant Shear Strength Data

Observation i	Shear Strength x_i	Differences $x_i - 2000$	Sign
1	2158.70	+158.70	+
2	1678.15	-321.85	-
3	2316.00	+316.00	+
4	2061.30	+61.30	+
5	2207.50	+207.50	+
6	1708.30	-291.70	-
7	1784.70	-215.30	-
8	2575.10	+575.10	+
9	2357.90	+357.90	+
10	2256.70	+256.70	+
11	2165.20	+165.20	+
12	2399.55	+399.55	+
13	1779.80	-220.20	-
14	2336.75	+336.75	+
15	1765.30	-234.70	-
16	2053.50	+53.50	+
17	2414.40	+414.40	+
18	2200.50	+200.50	+
19	2654.20	+654.20	+
20	1753.70	-246.30	-

This problem can be solved using the seven-step hypothesis-testing procedure:

- Parameter of interest:** The parameter of interest is the median of the distribution of propellant shear strength.
- Null hypothesis:** $H_0: \mu = 2000$ psi
- Alternative hypothesis:** $H_1: \mu \neq 2000$ psi
- Test statistic:** The test statistic is the observed number of plus differences in Table 9.5, or $r^+ = 14$.
- Reject H_0 if:** We will reject H_0 if the P -value corresponding to $r^+ = 14$ is less than or equal to $\alpha = 0.05$.
- Computations:** Because $r^+ = 14$ is greater than $n/2 = 20/2 = 10$, we calculate the P -value from

$$\begin{aligned}
 P &= 2P\left(R^+ \geq 14 \text{ when } p = \frac{1}{2}\right) \\
 &= 2 \sum_{r=14}^{20} \binom{20}{r} (0.5)^r (0.5)^{20-r} = 0.1153
 \end{aligned}$$

- Conclusions:** Because $p = 0.1153$ is not less than $\alpha = 0.05$, we cannot reject the null hypothesis that the median shear strength is 2000 psi. Another way to say this is that the observed number of plus signs $r^+ = 14$ was not large or small enough to indicate that median shear strength is different from 2000 psi at the $\alpha = 0.05$ level of significance.

It is also possible to construct a table of critical values for the sign test. This table is shown as Appendix Table VIII. Its use for the two-sided alternative hypothesis in Equation 9.54 is simple. As before, let R^+ denote the number of the differences $(X_i - \mu_0)$ that are positive and let R^-

denote the number of these differences that are negative. Let $R = \min(R^+, R^-)$. Appendix Table VIII presents critical values r_α^* for the sign test that ensure that $P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = \alpha$ for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.10$. If the observed value of the test statistic $r \leq r_\alpha^*$, the null hypothesis $H_0: \tilde{\mu} = \tilde{\mu}_0$ should be rejected.

To illustrate how this table is used, refer to the data in Table 9.5 that were used in Example 9.15. Now $r^+ = 14$ and $r^- = 6$; therefore, $r = \min(14, 6) = 6$. From Appendix Table VIII with $n = 20$ and $\alpha = 0.05$, we find that $r_{0.05}^* = 5$. Because $r = 6$ is not less than or equal to the critical value $r_{0.05}^* = 5$, we cannot reject the null hypothesis that the median shear strength is 2000 psi.

We can also use Appendix Table VIII for the sign test when a one-sided alternative hypothesis is appropriate. If the alternative is $H_1: \tilde{\mu} > \tilde{\mu}_0$, reject $H_0: \tilde{\mu} = \tilde{\mu}_0$ if $r^- \leq r_\alpha^*$; if the alternative is $H_1: \tilde{\mu} < \tilde{\mu}_0$, reject $H_0: \tilde{\mu} = \tilde{\mu}_0$ if $r^+ \leq r_\alpha^*$. The level of significance of a one-sided test is one-half the value for a two-sided test. Appendix Table VIII shows the one-sided significance levels in the column headings immediately following the two-sided levels.

Finally, note that when a test statistic has a discrete distribution such as R does in the sign test, it may be impossible to choose a critical value r_α^* that has a level of significance exactly equal to α . The approach used in Appendix Table VIII is to choose r_α^* to yield an α that is as close to the advertised significance level α as possible.

Ties in the Sign Test Because the underlying population is assumed to be continuous, there is a zero probability that we will find a “tie”—that is, a value of X_i exactly equal to $\tilde{\mu}_0$. However, this may sometimes happen in practice because of the way the data are collected. When ties occur, they should be set aside and the sign test applied to the remaining data.

The Normal Approximation When $p = 0.5$, the binomial distribution is well approximated by a normal distribution when n is at least 10. Thus, because the mean of the binomial is np and the variance is $np(1 - p)$, the distribution of R^+ is approximately normal with mean $0.5n$ and variance $0.25n$ whenever n is moderately large. Therefore, in these cases, the null hypothesis $H_0: \tilde{\mu} = \tilde{\mu}_0$ can be tested using the statistic

Normal Approximation for Sign Test Statistic

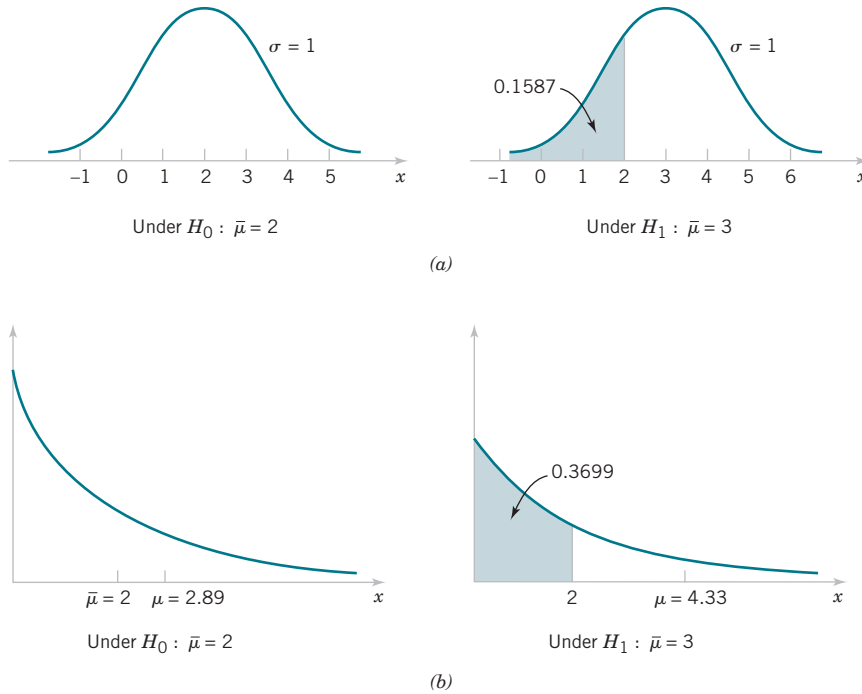
$$Z_0 = \frac{R^+ - 0.5n}{0.5\sqrt{n}} \quad (9.55)$$

A P -value approach could be used for decision making. The fixed significance level approach could also be used.

The two-sided alternative would be rejected if the observed value of the test statistic $|z_0| > z_{\alpha/2}$, and the critical regions of the one-sided alternative would be chosen to reflect the sense of the alternative. (If the alternative is $H_1: \tilde{\mu} > \tilde{\mu}_0$, reject H_0 if $z_0 > z_\alpha$, for example.)

Type II Error for the Sign Test The sign test will control the probability of a type I error at an advertised level α for testing the null hypothesis $H_0: \tilde{\mu} = \tilde{\mu}_0$ for any continuous distribution. As with any hypothesis-testing procedure, it is important to investigate the probability of a type II error, β . The test should be able to effectively detect departures from the null hypothesis, and a good measure of this effectiveness is the value of β for departures that are important. A small value of β implies an effective test procedure.

In determining β , it is important to realize not only that a particular value of $\tilde{\mu}$, say, $\tilde{\mu}_0 + \Delta$, must be used but also that the **form** of the underlying distribution will affect the calculations. To illustrate, suppose that the underlying distribution is normal with $\sigma = 1$ and we are testing the hypothesis $H_0: \tilde{\mu} = 2$ versus $H_1: \tilde{\mu} > 2$. (Because $\tilde{\mu} = \mu$ in the normal distribution, this is equivalent to testing that the mean equals 2.) Suppose that it is important to detect a departure

**FIGURE 9.18**

Calculation of β for the sign test. (a) Normal distributions. (b) Exponential distributions.

from $\tilde{\mu} = 2$ to $\tilde{\mu} = 3$. The situation is illustrated graphically in Figure 9.18(a). When the alternative hypothesis is true (H_1 : $\tilde{\mu} = 3$), the probability that the random variable X is less than or equal to the value 2 is

$$P(X \leq 2) = P(Z \leq -1) = \Phi(-1) = 0.1587$$

Suppose that we have taken a random sample of size 12. At the $\alpha = 0.05$ level, Appendix Table VIII indicates that we would reject H_0 : $\tilde{\mu} = 2$ if $r^- \leq r_{0.05}^* = 2$. Therefore, β is the probability that we do not reject H_0 : μ when in fact $\tilde{\mu} = 3$, or

$$\beta = 1 - \sum_{x=0}^2 \binom{12}{x} (0.1587)^x (0.8413)^{12-x} = 0.2944$$

If the distribution of X had been exponential rather than normal, the situation would be as shown in Figure 9.18(b), and the probability that the random variable X is less than or equal to the value $x = 2$ when $\tilde{\mu} = 3$ (note that when the median of an exponential distribution is 3, the mean is 4.33) is

$$P(X \leq 2) = \int_0^2 \frac{1}{4.33} e^{-\frac{1}{4.33}x} dx = 0.3699$$

In this case,

$$\beta = 1 - \sum_{x=0}^2 \binom{12}{x} (0.3699)^x (0.6301)^{12-x} = 0.8794$$

Thus, β for the sign test depends not only on the alternative value of $\tilde{\mu}$ but also on the area to the right of the value specified in the null hypothesis under the population probability distribution. This area depends highly on the shape of that particular probability distribution. In this example, β is large, so the ability of the test to detect this departure from the null hypothesis with the current sample size is poor.

9.9.2 The Wilcoxon Signed-Rank Test

The sign test uses only the plus and minus signs of the differences between the observations and the median $\tilde{\mu}_0$ (or the plus and minus signs of the differences between the observations in the paired case). It does not take into account the size or magnitude of these differences. Frank Wilcoxon devised a test procedure that uses both direction (sign) and magnitude. This procedure, now called the **Wilcoxon signed-rank test**, is discussed and illustrated in this section.

The Wilcoxon signed-rank test applies to the case of **symmetric continuous distributions**. Under these assumptions, the mean equals the median, and we can use this procedure to test the null hypothesis $\mu = \mu_0$.

The Test Procedure We are interested in testing $H_0: \mu = \mu_0$ against the usual alternatives. Assume that X_1, X_2, \dots, X_n is a random sample from a continuous and symmetric distribution with mean (and median) μ . Compute the differences $X_i - \mu_0, i = 1, 2, \dots, n$. Rank the absolute differences $|X_i - \mu_0|, i = 1, 2, \dots, n$ in ascending order, and then give the ranks the signs of their corresponding differences. Let W^+ be the sum of the positive ranks and W^- be the absolute value of the sum of the negative ranks, and let $W = \min(W^+, W^-)$. Appendix Table IX contains critical values of W , say, W_α^* . If the alternative hypothesis is $H_1: \mu \neq \mu_0$, then if the observed value of the statistic $w \leq w_\alpha^*$, the null hypothesis $H_0: \mu = \mu_0$ is rejected. Appendix Table IX provides significance levels of $\alpha = 0.10, \alpha = 0.05, \alpha = 0.02$, and $\alpha = 0.01$ for the two-sided test.

For one-sided tests, if the alternative is $H_1: \mu > \mu_0$, reject $H_0: \mu = \mu_0$ if $w^- \leq w_\alpha^*$; and if the alternative is $H_1: \mu < \mu_0$, reject $H_0: \mu = \mu_0$ if $w^+ \leq w_\alpha^*$. The significance levels for one-sided tests provided in Appendix Table IX are $\alpha = 0.05, 0.025, 0.01$, and 0.005 .

EXAMPLE 9.16 | Propellant Shear Strength-Wilcoxon Signed-Rank Test

We illustrate the Wilcoxon signed-rank test by applying it to the propellant shear strength data from Table 9.5. Assume that the underlying distribution is a continuous symmetric distribution. The seven-step procedure is applied as follows:

1. Parameter of interest: The parameter of interest is the mean (or median) of the distribution of propellant shear strength.

2. Null hypothesis: $H_0: \mu = 2000$ psi

3. Alternative hypothesis: $H_0: \mu \neq 2000$ psi

4. Test statistic: The test statistic is $w = \min(w^+, w^-)$

5. Reject H_0 if: We will reject H_0 if $w \leq w_{0.05}^* = 52$ from Appendix Table IX.

6. Computations: The signed ranks from Table 9.5 are shown in the following display:

The sum of the positive ranks is $w^+ = (1 + 2 + 3 + 4 + 5 + 6 + 11 + 13 + 15 + 16 + 17 + 18 + 19 + 20) = 150$, and the sum of the absolute values of the negative ranks is $w^- = (7 + 8 + 9 + 10 + 12 + 14) = 60$. Therefore,

$$w = \min(150, 60) = 60$$

Observation	Difference $x_i - 2000$	Signed Rank
16	+53.50	+1
4	+61.30	+2
1	+158.70	+3
11	+165.20	+4
18	+200.50	+5
5	+207.50	+6
7	+215.30	-7
13	-220.20	-8
15	-234.70	-9
20	-246.30	-10
10	+256.70	+11
6	-291.70	-12
3	+316.00	+13
2	-321.85	-14
14	+336.75	+15
9	+357.90	+16
12	+399.55	+17
17	+414.40	+18
8	+575.10	+19
19	+654.20	+20

7. Conclusions: Because $w = 60$ is not less than or equal to the critical value $w_{0.05}^* = 52$, we cannot reject the null hypothesis that the mean (or median, because the population is assumed to be symmetric) shear strength is 2000 psi.

Ties in the Wilcoxon Signed-Rank Test Because the underlying population is continuous, ties are theoretically impossible, although they will sometimes occur in practice. If several observations have the same absolute magnitude, they are assigned the average of the ranks that they would receive if they differed slightly from one another.

Large Sample Approximation If the sample size is moderately large, say, $n > 20$, it can be shown that W^+ (or W^-) has approximately a normal distribution with mean

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

and variance

$$\sigma_{W^+}^2 = \frac{n(n+1)(2n+1)}{24}$$

Therefore, a test of $H_0: \mu = \mu_0$ can be based on the statistic:

Normal Approximation for Wilcoxon Signed-Rank Statistic

$$Z_0 = \frac{W^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \quad (9.56)$$

An appropriate critical region for either the two-sided or one-sided alternative hypotheses can be chosen from a table of the standard normal distribution.

9.9.3 Comparison to the t -Test

If the underlying population is normal, either the sign test or the t -test could be used to test a hypothesis about the population median. The t -test is known to have the smallest value of β possible among all tests that have significance level α for the one-sided alternative and for tests with symmetric critical regions for the two-sided alternative, so it is superior to the sign test in the normal distribution case. When the population distribution is symmetric and nonnormal (but with finite mean), the t -test will have a smaller β (or a higher power) than the sign test unless the distribution has very heavy tails compared with the normal. Thus, the sign test is usually considered a test procedure for the median rather than as a serious competitor for the t -test. The Wilcoxon signed-rank test is preferable to the sign test and compares well with the t -test for symmetric distributions. It can be useful for situations in which a transformation on the observations does not produce a distribution that is reasonably close to the normal.

9.10 Equivalence Testing

Statistical hypothesis testing is one of the most useful techniques of statistical inference. However, it works in only one direction; that is, it starts with a statement that is assumed to be true (the null hypothesis H_0) and attempts to disprove this claim in favor of the alternative hypothesis H_1 . The strong statement about the alternative hypothesis is made when the null hypothesis is rejected. This procedure works well in many but not all situations.

To illustrate, consider a situation in which we are trying to qualify a new supplier of a component that we use in manufacturing our product. The current supplier produces these components with a standard mean resistance of 80 ohms. If the new supplier can provide components with the same mean resistance, we will qualify them. Having a second source for this component is considered to be important because demand for our product is expected to grow rapidly in the

near future, and the second supplier will be necessary to meet the anticipated increase in demand. The traditional formulation of the hypothesis test

$$H_0: \mu = 80 \qquad H_1: \mu \neq 80$$

really is not satisfactory. Only if we reject the null hypothesis do we have a strong conclusion. We actually want to state the hypotheses as follows:

$$H_0: \mu \neq 80 \qquad H_1: \mu = 80$$

This type of hypothesis statement is called an **equivalence test**. We assume that the new supplier is different from the standard unless we have strong evidence to reject that claim. The way that this equivalence test is carried out is to test the following two sets of one-sided alternative hypotheses:

$$H_0: \mu = 80 + \delta \qquad H_1: \mu < 80 + \delta$$

and

$$H_0: \mu = 80 - \delta \qquad H_1: \mu > 80 - \delta$$

where δ is called the **equivalence band**, which is a practical threshold or limit within which the mean performance (here the resistance) is considered to be the same as the standard. The interval $80 \pm \delta$ is called an **equivalence interval**. The first set of hypotheses is a test of the mean that shows that the difference between the mean and the standard is significantly less than the upper equivalence limit of the interval, and the second set of hypotheses is a test of the mean that shows that the difference between the mean and the standard is significantly greater than the lower equivalence limit. We are going to apply both tests to the same sample of data, leading to a test of equivalence that is sometimes called **two one-sided tests (TOST)**.

EXAMPLE 9.17

Suppose that we have a random sample of $n = 50$ components from the new supplier. Resistance is approximately normally distributed, and the sample mean and standard deviation (in ohms) are $\bar{x} = 79.98$ and $s = 0.10$. The sample mean is close to the standard of 80 ohms. Suppose that our error of measurement is approximately 0.01 ohm. We will decide that if the new supplier has a mean resistance that is within 0.05 of the standard of 80, there is no practical difference in performance. Therefore, $\delta = 0.05$. Notice that we have chosen the equivalence band to be greater than the usual or expected measurement error for the resistance. We now want to test the hypotheses

$$H_0: \mu = 80.05 \qquad H_1: \mu < 80.05$$

and

$$H_0: \mu = 79.95 \qquad H_1: \mu > 79.95$$

Consider testing the first set of hypotheses. It is straightforward to show that the value of the test statistic is $t_0 = -4.95$, and the P -value is less than 0.01. Therefore, we conclude that the mean resistance is less than 80.05. For the second set of hypotheses, the test statistic is $t_0 = 2.12$, and the P -value is less than 0.025, so the mean resistance is significantly greater than 79.95 and significantly less than 80.05. Thus, we have enough evidence to conclude that the new supplier produces components that are equivalent to those produced by the current supplier because the mean is within the ± 0.05 ohm interval.

Equivalence testing has many applications, including the supplier qualification problem illustrated here, generic drug manufacturing, and new equipment qualification. The experimenter must decide what defines equivalence. Issues that should be considered include these:

1. Specifying the equivalence band. The parameter δ should be larger than the typical measurement error. A good rule of thumb is that δ should be at least three times the typical measurement error.

2. The equivalence band should be much smaller than the usual process variation.
3. The equivalence band should be much smaller than the product or process specifications. Specifications typically define fitness for use.
4. The equivalence band should be related to actual functional performance; that is, how much of a difference can be tolerated before performance is degraded?

9.11 Combining P -Values

Testing several sets of hypotheses that relate to a problem of interest occurs fairly often in engineering and many scientific disciplines. For example, suppose that we are developing a new synthetic fiber to be used in manufacturing body armor for the military and law enforcement agencies. This fiber needs to exhibit a high breaking strength (at least 100 lb/in²) for the new product to work properly. The engineering development lab produced several batches or lots of this fiber, a random sample of three fiber specimens from each lot has been taken, and the sample specimens tested. For each lot, the hypotheses of interest are

$$H_0: \mu = 100 \qquad H_1: \mu > 100$$

The development lots are small, and the testing is destructive, so the sample sizes are also small. After six lots have been produced, the P -values from these six independent tests of hypotheses are 0.105, 0.080, 0.250, 0.026, 0.650, and 0.045. Given the size of these P -values, we suspect that the new material is going to be satisfactory, but the sample sizes are small, and it would be helpful if we could combine the results from all six tests to determine whether the new material will be acceptable. Combining results from several studies or experiments is sometimes called **meta-analysis**, a technique that has been used in many fields including public health monitoring, clinical trials of new medical devices or treatments, ecology, and genetics. One method that can be used to combine these results is to combine all of the individual P -values into a single statistic for which one P -value can be computed. This procedure was developed by R. A. Fisher.

Let P_i be the P -value for the i th set of hypotheses, $i = 1, 2, \dots, m$. The test statistic is

$$\chi_0^2 = -2 \sum_{i=1}^m \ln(P_i)$$

The test statistic χ_0^2 follows a chi-square distribution with $2m$ degrees of freedom. A P -value can be computed for the observed value of this statistic. A small P -value would lead to rejection of the shared null hypotheses and a conclusion that the combined data support the alternative.

As an example, the test statistic χ_0^2 for the six tests described is

$$\chi_0^2 = -2[\ln(0.105) + \ln(0.080) + \ln(0.250) + \ln(0.026) + \ln(0.650) + \ln(0.045)] = 26.6947$$

with $2m = 2(6) = 12$ degrees of freedom. The P -value for this statistic is $0.005 < P < 0.01$, a very small value, which leads to rejection of the null hypothesis. In other words, the combined information from all six tests provides evidence that the mean fiber strength exceeds 100 lb/in².

Fisher's method does not require all the null hypotheses be the same. Some applications involve many sets of hypotheses that do not have the same null. In these situations, the alternative hypothesis is taken to be that at least one of the null hypotheses is false. Fisher's method was developed in the 1920s. Since then, a number of other techniques has been proposed. For a good discussion of these alternative methods along with comments on their appropriateness and power, see the article by Piegorsch and Bailer ["Combining Information," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2009, Vol. 1(3), pp. 354–360].

Important Terms and Concepts

Acceptance region	Hypothesis testing	Reference distribution for a test statistic
Alternative hypothesis	Independence test	Rejection region
α and β	Inference	Sampling distribution
Chi-square tests	Nonparametric and distribution free methods	Sample size determination for hypothesis tests
Combining <i>P</i> -values	Normal approximation to nonparametric tests	Sign test
Confidence interval	Null distribution	Significance level of a test
Connection between hypothesis tests and confidence intervals	Null hypothesis	Statistical hypothesis
Contingency table	Observed significance level	Statistical versus practical significance
Critical region for a test statistic	One- and two-sided alternative hypotheses	Symmetric continuous distributions
Critical values	Operating characteristic (OC) curves	<i>t</i> -test
Equivalence testing	Parametric	Test statistic
Fixed significance level	Power of a statistical test	Type I and type II errors
Goodness-of-fit test	<i>P</i> -value	Wilcoxon's signed-rank test
Homogeneity test	Ranks	<i>z</i> -test
Hypotheses		

Statistical Inference for Two Samples



© HKPNC / iStockphoto

LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Structure comparative experiments involving two samples as hypothesis tests
 2. Test hypotheses and construct confidence intervals on the difference in means of two normal distributions
 3. Test hypotheses and construct confidence intervals on the ratio of the variances or standard deviations of two normal distributions
 4. Test hypotheses and construct confidence intervals on the difference in two population proportions
 5. Use the P -value approach for making decisions in hypotheses tests
 6. Compute power, and type II error probability, and make sample size decisions for two-sample tests on means, variances, and proportions
 7. Explain and use the relationship between confidence intervals and hypothesis tests
-