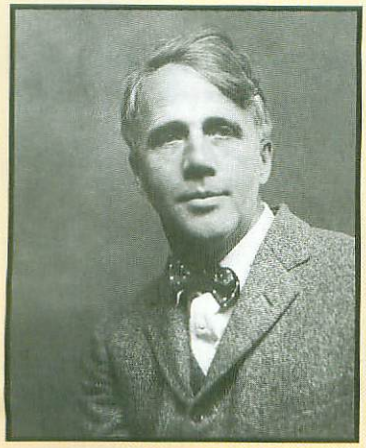
8

Estimation

We dance round in a ring and suppose,
But the Secret sits in the middle and knows.

—Robert Frost,
"The Secret Sits"*



Robert Lee Frost (1874–1963)

This celebrated American poet drew poetic symbols largely from common experiences observed in his rural New England.

In Chapter 1 we said that statistics is the study of how to collect, organize, analyze, and interpret numerical data. That part of statistics concerned with analysis, interpretation, and forming conclusions about the source of the data is called *statistical inference*. Problems of statistical inference require us to draw a *sample* of observations from a larger *population*. A sample usually contains incomplete information, so in a sense we must "dance round in a ring and suppose." Nevertheless, conclusions about the population can be obtained from sample data by the use of statistical estimates. This chapter introduces you to several widely used methods of estimation.

*Source: From *The Poetry of Robert Frost*, edited by Edward Connery Lathem. Copyright 1942 by Robert Frost, © 1970 by Lesley Frost Ballantine, © 1969 by Henry Holt and Company. Reprinted by permission of Henry Holt and Company, LLC.

PREVIEW QUESTIONS

- ◇ How do you estimate the expected value of a random variable? What assumptions are needed? How much confidence should be placed in such estimates? (SECTION 8.1)
- ◇ If you start out in the beginning design stage of a statistical project, how large a sample size should you plan to get? (SECTION 8.1)
- ◇ What famous statistician worked for Guinness brewing company in Ireland? What has this to do with constructing estimates from sample data? (SECTION 8.2)
- ◇ How do you estimate the proportion p for success in a binomial experiment? How does the normal approximation fit into this process? (SECTION 8.3)

For on-line student resources, visit **math.college.hmco.com/students** and follow the Statistics links to the Brase/Brase, *Understanding Basic Statistics,* 4th edition web site.

FOCUS PROBLEM

Trick or Treat!!!

About 28% of U.S. households turn out the lights and pretend not to be at home on Halloween (*Source: Are You Normal About Money?* by Bernice Kanner, Bloomberg Press).

Alice is a sociology student who is studying the affluent Cherry Creek neighborhood in Denver. As part of a larger survey, Alice interviewed a random sample of 35 households. One of the questions she asked was whether the resident turned out the lights and pretended not to be at home on Halloween. It was found that 11 of the 35 residents actually did this practice.

(a) Compute a 90% confidence interval or p, the proportion of all households in Cherry Creek that pretend not to be at home on Halloween.

(b) What assumptions are necessary to calculate the confidence interval of part (a)? Do you think these assumptions are met in this case? Explain.

(c) The national proportion is about 0.28. Is 0.28 in the confidence interval you computed? Based on your answer, does it seem that the Cherry Creek neighborhood is much different (either higher or lower proportion) from the population of all U.S. households? Explain.

(See Problem 8 of Section 8.3.)

8.1 Estimating μ When σ Is Known

FOCUS POINTS

- ✓ Explain the meaning of confidence level, error of estimate, and critical value.
- ✓ Find the critical value corresponding to a given confidence level.
- ✓ Compute confidence intervals for μ when σ is known. Interpret the results.
- ✓ Compute the sample size to be used for estimating a mean μ.

Because of time and money constraints, difficulty in finding population members, and so forth, we usually do not have access to *all* measurements of an *entire* population. Instead we rely on information from a sample.

In this section, we develop techniques for estimating the population mean μ using sample data. We assume the population standard deviation σ is known.

Let's begin by listing some basic assumptions used in the development of our formulas for estimating μ when σ is known.

Assumptions about the random variable x

1. We have a *simple random sample* of size n drawn from a population of x values.
2. The value of σ, the population standard deviation of x, *is known.*
3. If the x *distribution is normal,* then our methods work for *any sample size* n.
4. If x has an unknown distribution, then we require a *sample size* $n \geq 30$. However, if the x distribution is distinctly skewed and definitely not mound-shaped, a sample of size 50 or even 100 or higher may be necessary.

Point estimate

An estimate of a population parameter given by a single number is called a *point estimate* for that parameter. It will come as no great surprise that we use $\bar{x}$ (the sample mean) as the point estimate for μ (the population mean).

A **point estimate** of a population parameter is an estimate of the parameter using a single number.

$\bar{x}$ is the **point estimate** for μ.

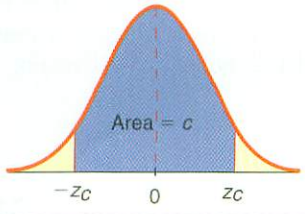
Margin of error

Even with a large random sample, the value of $\bar{x}$ usually is not *exactly* equal to the population mean μ. The *margin of error* is the magnitude of the difference between the sample point estimate and the true population parameter value.

When using $\bar{x}$ as a point estimate for μ, the **margin of error** is the magnitude of $\bar{x} - \mu$ or $|\bar{x} - \mu|$.

FIGURE 8-1

Confidence Level c and Corresponding Critical Value z_c Shown on the Standard Normal Curve



We cannot say exactly how close $\bar{x}$ is to μ when μ is unknown. Therefore, the exact margin of error is unknown when the population parameter is unknown. Of course, μ is usually not known or there would be no need to estimate it. In this section, we will use the language of probability to give us an idea of the size of the margin of error when we use $\bar{x}$ as a point estimate for μ.

First, we need to learn about *confidence levels.* The reliability of an estimate will be measured by the confidence level.

Suppose we want a confidence level of c (see Figure 8-1). Theoretically, you can choose c to be any value between 0 and 1, but usually c is equal to a number such

as 0.90, 0.95, or 0.99. In each case, the value z_c is the number such that the area under the standard normal curve falling between $-z_c$ and z_c is equal to c. The value z_c is called the *critical value* for a confidence level of c.

Finding the critical value

For a confidence level c, the **critical value** z_c is the number such that the area under the standard normal curve between $-z_c$ and z_c equals c.

The area under the normal curve from $-z_c$ to z_c is the probability that the standardized normal variable z lies in that interval. This means that

$$P(-z_c < z < z_c) = c$$

EXAMPLE 1
Find a critical value

Let us use Table 3 of the Appendix to find a number $z_{0.99}$ such that 99% of the area under the standard normal curve lies between $-z_{0.99}$ and $z_{0.99}$. That is, we will find $z_{0.99}$ such that

$$P(-z_{0.99} < z < z_{0.99}) = 0.99$$

SOLUTION: In Section 7.3, we saw how to find the z value when we were given an area between $-z$ and z. The first thing we did was to find the corresponding area to the left of $-z$. If A is the area between $-z$ and z, then $(1 - A)/2$ is the area to the left of z. In our case, the area between $-z$ and z is 0.99. The corresponding area in the left tail is $(1 - 0.99)/2 = 0.005$ (see Figure 8-2).

Next, we use Table 3 of the Appendix to find the z value corresponding to a left-tail area of 0.0050. Table 8-1 shows an excerpt from Table 3 of the Appendix.

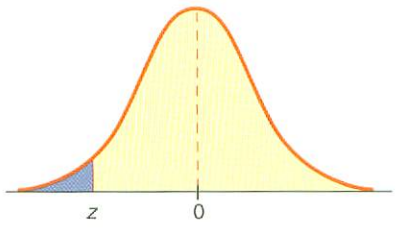
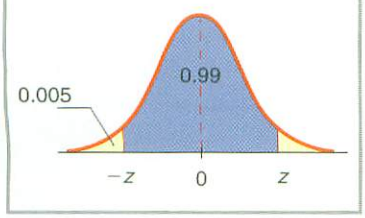


TABLE 8-1 Excerpt from Table 3 of the Appendix

z	.00	...	.07	.08	.09
−3.4	.0003		.0003	.0003	.0002
⋮					
−2.5	.0062		.0051	.0049	.0048
				↑	
				.0050	

FIGURE 8-2
Area Between $-z$ and z Is 0.99



From Table 8-1, we see that the desired area, 0.0050, is exactly halfway between the areas corresponding to $z = -2.58$ and $z = -2.57$. Because the two area values are so close together, we use the more conservative z value -2.58 rather than interpolate. In fact, $z_{0.99} \approx 2.576$. However, to two decimal places, we use $z_{0.99} = 2.58$ as the critical value for a confidence level of $c = 0.99$. We have

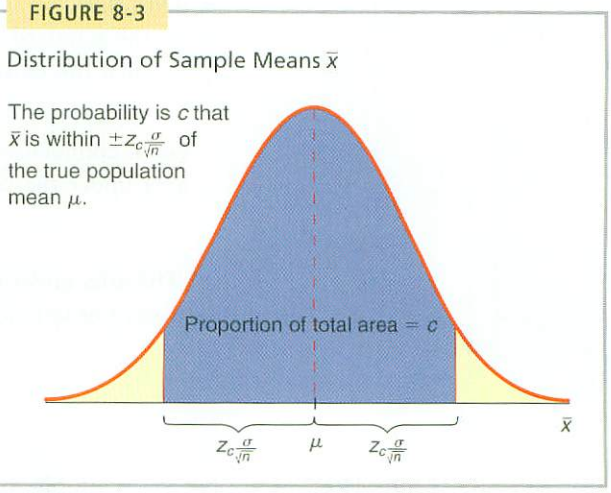
$$P(-2.58 < z < 2.58) = 0.99$$ ◆

The results of Example 1 will be used a great deal in our later work. For convenience, Table 8-2 gives some levels of confidence and corresponding critical values z_c. The same information is provided in Table 3(b) of the Appendix.

TABLE 8-2 Some Levels of Confidence and Their Corresponding Critical Values

Level of Confidence c	Critical Value z_c
0.70, or 70%	1.04
0.75, or 75%	1.15
0.80, or 80%	1.28
0.85, or 85%	1.44
0.90, or 90%	1.645
0.95, or 95%	1.96
0.98, or 98%	2.33
0.99, or 99%	2.58

FIGURE 8-3



An estimate is not very valuable unless we have some kind of measure of how "good" it is. The language of probability can give us an idea of the size of the margin of error caused by using the sample mean $\bar{x}$ as an estimate for the population mean.

Remember that $\bar{x}$ is a random variable. Each time we draw a sample of size n from a population, we can get a different value for $\bar{x}$. According to the central limit theorem, if the sample size is large, then $\bar{x}$ has a distribution that is approximately normal with mean $\mu_{\bar{x}} = \mu$, the population mean we are trying to estimate. The standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. If x has a normal distribution, these results are true *for any sample size.* (See Theorem 7.1.)

This information, together with our work on confidence levels, leads us to the probability statement

$$P\left(-z_c \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_c \frac{\sigma}{\sqrt{n}}\right) = c \tag{1}$$

◇ **COMMENT** To derive Equation (1), we start with the probability statement $P(-z_c < z < z_c) = c$. Since $n \geq 30$, we can use the central limit theorem and replace z by $(\bar{x} - \mu)/(\sigma/\sqrt{n})$. Finally, we multiply all parts of the inequality by $(\sigma/\sqrt{n})$ to obtain Equation (1). ◇

Equation (1) uses the language of probability to give us an idea of the size of the margin of error for the corresponding confidence level c. In words, Equation (1) states that the probability is c that our point estimate $\bar{x}$ is within a distance $\pm z_c(\sigma/\sqrt{n})$ of the population mean μ. This relationship is shown in Figure 8-3.

The *margin of error* (or absolute error) using $\bar{x}$ as a point estimate for μ is $|\bar{x} - \mu|$. In most practical problems, μ is unknown, so the margin of error is also unknown. However, Equation (1) allows us to compute an *error tolerance E,* which serves as a bound on the margin of error. Using a c% level of confidence, we can say that the point estimate $\bar{x}$ differs from the population mean μ by a *maximal margin of error*

Maximal margin of error, E

$$E = z_c \frac{\sigma}{\sqrt{n}} \tag{2}$$

Note: Formula (2) for E is based on the fact that the sampling distribution for $\bar{x}$ is exactly normal, with mean μ and standard deviation $\sigma/\sqrt{n}$. This occurs whenever the x distribution is normal with mean μ and standard deviation σ. If the x distribution is not normal, then according to the central limit theorem, large samples ($n \geq 30$) produce an $\bar{x}$ distribution that is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Using Equations (1) and (2), we conclude that

$$P(-E < \bar{x} - \mu < E) = c \tag{3}$$

Confidence interval for μ with σ known

Applying a little algebra to formula (3) produces

$$P(\bar{x} - E < \mu < \bar{x} + E) = c \tag{4}$$

Equation (4) states that there is a chance of c that the interval from $\bar{x} - E$ to $\bar{x} + E$ contains the population mean μ. We call this interval a *c confidence interval for μ.*

A **c confidence interval for μ** is an interval computed from sample data in such a way that c is the probability of generating an interval containing the actual value of μ.

We may get a different confidence interval for each different sample that is taken. Some intervals will contain the population mean μ and others will not. However, in the long run, the proportion of confidence intervals that contain μ is c.

PROCEDURE

How to find a confidence interval for μ when σ is known

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean $\bar{x}$. The value of σ is already known (perhaps from a previous study).

If you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, then use a sample size of $n \geq 30$.

Confidence interval for μ when σ is known

$$\bar{x} - E < \mu < \bar{x} + E \tag{5}$$

where $\bar{x}$ = sample mean of a simple random sample

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

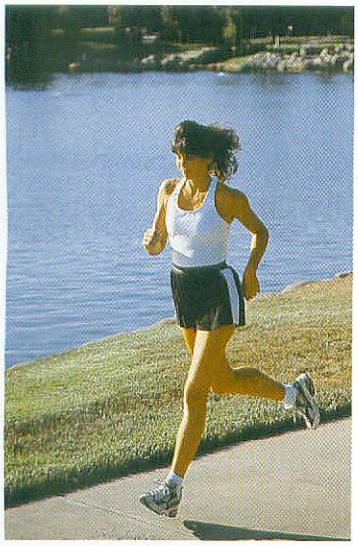
c = confidence level ($0 < c < 1$)

z_c = critical value for confidence level c based on the standard normal distribution (See Table 3(b) of the Appendix for frequently used values.)

EXAMPLE 2

Confidence interval for μ with σ known

Julia enjoys jogging. She has been jogging over a period of several years, during which time her physical condition has remained constantly good. Usually, she jogs 2 miles per day. The standard deviation of her times is $\sigma = 1.80$ min. During the past year Julia has recorded her times required to run 2 miles. She has a random sample of 90 of these times. For these 90 times the mean was $\bar{x} = 15.60$ minutes. Let μ be the mean jogging time for the entire distribution of Julia's 2-mile running times (taken over the past year). Find a 0.95 confidence interval for μ.



SOLUTION: The interval from $\bar{x} - E$ to $\bar{x} + E$ will be a 95% confidence interval for μ. In this case, $c = 0.95$, so $z_c = 1.96$ (see Table 8-2). The sample size $n = 90$ is large enough for the $\bar{x}$ distribution to be approximately normal, with mean μ and standard deviation $\sigma/\sqrt{n}$. Therefore,

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

$$E = 1.96\left(\frac{1.80}{\sqrt{90}}\right)$$

$$E \approx 0.37$$

Using Equation (5), the given value of $\bar{x}$, and our computed value for E, we get the 95% confidence interval for μ.

$$\bar{x} - E < \mu < \bar{x} + E$$

$$15.60 - 0.37 < \mu < 15.60 + 0.37$$

$$15.23 < \mu < 15.97$$

We conclude with 95% confidence that the interval from 15.23 min. to 15.97 min. is one that contains the population mean μ of jogging times for Julia. ◇

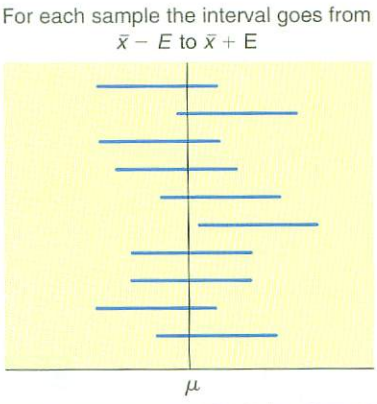
A few comments are in order about the general meaning of the term *confidence interval*. It is important to realize that the endpoints $\bar{x} \pm E$ are really statistical *variables*. Equation (4) states that we have a chance c of obtaining a sample such that the interval, once it is computed, will contain the parameter μ. Of course, after the confidence interval is numerically fixed, it either does or does not contain μ. So the probability is 1 or 0 that the interval, when it is fixed, will contain μ. A nontrivial probability statement can be made only about variables, not constants. Therefore, Equation (4) really states that if we repeat the experiment many times and get lots of confidence intervals (for the same sample size), then the proportion of all intervals that will turn out to contain the mean μ is c.

In Figure 8-4, the horizontal lines represent 0.90 confidence intervals for various samples of the same size from a distribution. Some of these intervals contain μ, and others do not. Since the intervals are 0.90 confidence intervals, about 90% of all such intervals should contain μ. For each sample, the interval goes from $\bar{x} - E$ to $\bar{x} + E$.

◇ **COMMENT** Please see Using Technology at the end of this chapter for a computer demonstration of this discussion about confidence intervals. ◇

FIGURE 8-4

0.90 Confidence Intervals for Samples of the Same Size



GUIDED EXERCISE 1

Confidence interval for μ with σ known

Walter usually meets Julia at the track. He prefers to jog 3 miles. From long experience, he knows that $\sigma = 2.40$ minutes for his jogging times. For a random sample of 90 jogging sessions, his mean time was $\bar{x} = 22.50$ minutes. Let μ be the mean jogging time for the entire distribution of Walter's 3-mile jogging times over the past several years. Find a 0.99 confidence interval for μ.

(a) What is the value of $z_{0.99}$? (See Table 8-2.) ➡ $z_{0.99} = 2.58$

(b) Is the $\bar{x}$ distribution approximately normal? ➡ Yes; we know this from the central limit theorem.

(c) What is the value of E? ➡ $E = z_c \dfrac{\sigma}{\sqrt{n}} = 2.58\left(\dfrac{2.40}{\sqrt{90}}\right) \approx 0.65$

(d) What are the endpoints for a 0.99 confidence interval for μ? ➡ The endpoints are given by

$\bar{x} - E \approx 22.50 - 0.65 = 21.85$

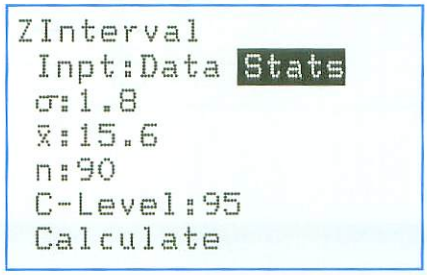
$\bar{x} + E \approx 22.50 + 0.65 = 23.15$

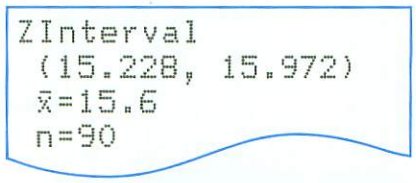
(e) How can we interpret the confidence interval? ➡ We are 99% certain that the interval from 21.85 to 23.15 is an interval that contains the population mean time μ.

When we use samples to estimate the mean of a population, we generate a small error. However, samples are useful even when it is possible to survey the entire population because the use of a sample may yield savings of time or effort in collecting data.

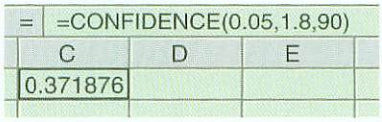
TECH NOTE The TI-84Plus and TI-83Plus calculators, Excel, and Minitab all support confidence intervals for μ when σ is known. The level of support varies according to the technology. When a confidence interval is given, the standard mathematical notation (lower value, upper value) is used. For instance, the notation (15.23, 15.97) means the interval from 15.23 to 15.97.

TI-84Plus/TI-83Plus This calculator gives the most extensive support. The user can opt to enter raw data or just summary statistics. In each case, the value of σ must be specified. Press the **STAT** key and select **TESTS**, then use **7:ZInterval.** The TI-84Plus/ TI-83Plus output shows the results for Example 2.





Excel Excel gives only the value of the maximal error of estimate E. Use the menu choice **Paste Function** (f_x) ➤ **Statistics** ➤ **Confidence(alpha, σ, n).** In the dialogue box, the value of alpha is 1 − confidence level. The Excel output shows the value of E for Example 2.



An alternate approach incorporating raw data (using the Student's t distribution presented in the next section) uses the menu choices **Tools** ➤ **Data Analysis** ➤ **Describe Statistics.** Again, the value of E for the interval is given.

Minitab Raw data are required. Use the menu choices **Stat** ➤ **Basic Statistics** ➤ **1-SampleZ.**

Sample Size for Estimating the Mean μ

In the design stages of statistical research projects, it is a good idea to decide in advance on the confidence level you wish to use and to select the *maximum* margin of error E you want for your project. How you choose to make these decisions depends on the requirements of the project and the practical nature of the problem.

Whatever specifications you make, the next step is to determine the sample size. Solving the formula that gives the maximal margin of error E for n enables us to determine the minimum sample size.

PROCEDURE

How to find the sample size n for estimating μ when σ is known

Assuming the distribution of sample means $\bar{x}$ is approximately normal, then

$$n = \left(\frac{z_c \sigma}{E}\right)^2 \tag{6}$$

where E = specified maximal error of estimate

σ = population standard deviation

z_c = critical value from the normal distribution for the desired confidence level c. Commonly used values of z_c can be found in Table 3(b) of the Appendix.

If n is not a whole number, increase n to the next higher whole number. Note that n is the minimal sample size for a specified confidence level and maximal error of estimate E.

◇ **COMMENT:** If you have a *preliminary study* involving a sample size of 30 or larger, then for most practical purposes it is safe to approximate σ with the sample standard deviation s in the formula for sample size. ◇

EXAMPLE 3

Sample size for estimating μ

A wildlife study is designed to find the mean weight of salmon caught by an Alaskan fishing company. A recent study of a random sample of 50 salmon showed $s \approx 2.15$ lb. How large a sample should be taken to be 99% confident that the sample mean $\bar{x}$ is within 0.20 lb of the true mean weight μ?

Salmon moving upstream

SOLUTION: In this problem, $z_{0.99} = 2.58$ (see Table 8-2) and $E = 0.20$. The preliminary study of 50 fish is large enough to permit a good approximation of σ by $s = 2.15$. Therefore, Equation (6) becomes

$$n = \left(\frac{z_c \sigma}{E}\right)^2 \approx \left(\frac{(2.58)(2.15)}{0.20}\right)^2 = 769.2$$

Note: In determining sample size, any fractional value of n is always rounded to the *next higher whole number.* We conclude that a sample size of 770 will be large enough to satisfy the specifications. Of course, a sample size larger than 770 also works. ◇

VIEWPOINT

Music and Techno Theft

Performing rights organizations ASCAP (American Society of Composers, Authors, and Publishers) and BMI (Broadcast Music, Inc.) collect royalties for songwriters and music publishers. Radio, television, cable, nightclubs, restaurants, elevators, and even beauty parlors play music that is copyrighted by a composer or publisher. The royalty payment for this music turns out to be more than a billion dollars a year (Source: *The Wall Street Journal*). How do ASCAP and BMI know who is playing what music? The answer is, they don't know! Instead of tracking *exactly* what gets played, they use random sampling and *confidence intervals*. For example, each radio station (there are more than 10,000 in the U.S.) has randomly chosen days of programming analyzed every year. The results are used to assess royalty fees. In fact, Deloitte & Touche (a financial services company) administers the sampling process.

Although the system is not perfect, it helps bring order into an otherwise chaotic accounting system. Such methods of "copyright policing" help prevent techno theft, ensuring that many songwriters and recording artists get a reasonable return for their creative work.

SECTION 8.1 PROBLEMS

Answers may vary slightly due to rounding.

1. *Zoology: Hummingbirds* Allen's hummingbird (*Selasphorus sasin*) has been studied by zoologist Bill Alther (Reference: *Hummingbirds*, K. Long and W. Alther). A small group of 15 Allen's hummingbirds has been under study in Arizona. The average weight for these birds is $\bar{x} = 3.15$ gm. Based on previous studies, we can assume that the weights of Allen's hummingbirds have a normal distribution with $\sigma = 0.33$ gm.
   (a) Find an 80% confidence interval for the average weights of Allen's hummingbirds in the study region. What is the margin of error?
   (b) What conditions are necessary for your calculations?
   (c) Give a brief interpretation of your results in the context of this problem.
   (d) *Sample Size* Find the sample size necessary for an 80% confidence level with a maximal error of estimate $E = 0.08$ for the mean weights of the hummingbirds.

2. *Diagnostic Tests: Uric Acid* Overproduction of uric acid in the body can be an indication of cell breakdown. This may be an advance indication of illness such as gout, leukemia, or lymphoma (Reference: *Manual of Laboratory and Diagnostic Tests*, F. Fischbach). Over a period of months, an adult male patient has taken eight blood tests for uric acid. The mean concentration was $\bar{x} = 5.35$ mg/dl. The distribution of uric acid in healthy adult males can be assumed to be normal with $\sigma = 1.85$ mg/dl.
   (a) Find a 95% confidence interval for the population mean concentration of uric acid in this patient's blood. What is the margin of error?
   (b) What conditions are necessary for your calculations?
   (c) Give a brief interpretation of your results in the context of this problem.
   (d) *Sample Size* Find the sample size necessary for a 95% confidence level with maximal error of estimate $E = 1.10$ for the mean concentration of uric acid in this patient's blood.

3. *Diagnostic Tests: Plasma Volume* Total plasma volume is important in determining the required plasma component in blood replacement therapy for a person undergoing surgery. Plasma volume is influenced by the overall health and physical activity of an individual. (Reference: See Problem 2.) Suppose that a random sample of 45 male firefighters are tested and that they have a plasma volume sample mean of $\bar{x} = 37.5$ ml/kg (milliliters plasma per kilogram body weight). Assume that $\sigma = 7.50$ ml/kg for the distribution of blood plasma.
   (a) Find a 99% confidence interval for the population mean blood plasma volume in male firefighters. What is the margin of error?
   (b) What conditions are necessary for your calculations?
   (c) Give a brief interpretation of your results in the context of this problem.
   (d) *Sample Size* Find the sample size necessary for a 99% confidence level with maximal error of estimate $E = 2.50$ for the mean plasma volume in male firefighters.

4. *Agriculture: Watermelon* What price do farmers get for their watermelon crops? In the third week of July, a random sample of 40 farming regions gave a sample mean of $\bar{x} = \$6.88$ per 100 pounds of watermelon. Assume that σ is known to be \$1.92 per 100 pounds (Reference: *Agricultural Statistics,* U.S. Department of Agriculture).
   (a) Find a 90% confidence interval for the population mean price (per 100 pounds) that farmers in this region get for their watermelon crop. What is the margin of error?
   (b) *Sample Size* Find the sample size necessary for a 90% confidence level with maximal error of estimate $E = 0.3$ for the mean price per 100 pounds of watermelon.
   (c) A farm brings 15 tons of watermelon to market. Find a 90% confidence interval for the population mean cash value of this crop. What is the margin of error? *Hint:* 1 ton is 2000 pounds.

5. *FBI Report: Larceny* Thirty small communities in Connecticut (population near 10,000 each) gave an average of $\bar{x} = 138.5$ reported cases of larceny per year. Assume that σ is known to be 42.6 cases per year (Reference: *Crime in the United States,* Federal Bureau of Investigation).
   (a) Find a 90% confidence interval for the population mean annual number of reported larceny cases in such communities. What is the margin of error?
   (b) Find a 95% confidence interval for the population mean annual number of reported larceny cases in such communities. What is the margin of error?
   (c) Find a 99% confidence interval for the population mean annual number of reported larceny cases in such communities. What is the margin of error?
   (d) Compare the margins of error for parts (a) through (c). As the confidence level increases, does the margin of error increase?
   (e) Compare the lengths of the confidence intervals for parts (a) through (c). As the confidence level increases, does the confidence interval increase in length?

6. *Salaries: Student Services* Consider college officials in admissions, registration, counseling, financial aid, campus ministry, food services, and so on. How much money do these people make each year? Suppose you read in your local newspaper that 45 officials in student services earned an average of $\bar{x} = \$50{,}340$ each year (Reference: *Chronicle of Higher Education*).
   (a) Assume that $\sigma = \$16{,}920$ for salaries of college officials in student services. Find a 90% confidence interval for the population mean salary of such personnel. What is the margin of error?
   (b) Assume that $\sigma = \$10{,}780$ for salaries of college officials in student services. Find a 90% confidence interval for the population mean salary of such personnel. What is the margin of error?

(c) Assume that $\sigma =$ \$4830 for salaries of college officials in student services. Find a 90% confidence interval for the population mean salary of such personnel. What is the margin of error?

(d) Compare the margins of error for parts (a) through (c). As the standard deviation decreases, does the margin of error decrease?

(e) Compare the lengths of the confidence intervals for parts (a) through (c). As the standard deviation decreases, does the length of a 90% confidence interval decrease?

7. *Salaries: College Administrators* How much do college administrators (not teachers or service personnel) make each year? Suppose you read the local newspaper and find that the average annual salary of administrators in the local college is $\bar{x} =$ \$58,940. Assume that σ is known to be \$18,490 for college administrator salaries (Reference: *The Chronicle of Higher Education*).

(a) Suppose that $\bar{x} =$ \$58,940 is based on a random sample of $n = 36$ administrators. Find a 90% confidence interval for the population mean annual salary of local college administrators. What is the margin of error?

(b) Suppose that $\bar{x} =$ \$58,940 is based on a random sample of $n = 64$ administrators. Find a 90% confidence interval for the population mean annual salary of local college administrators. What is the margin of error?

(c) Suppose that $\bar{x} =$ \$58,940 is based on a random sample of $n = 121$ administrators. Find a 90% confidence interval for the population mean annual salary of local college administrators. What is the margin of error?

(d) Compare the margins of error for parts (a) through (c). As the sample size increases, does the margin of error decrease?

(e) Compare the lengths of the confidence intervals for parts (a) through (c). As the sample size increases, does the length of a 90% confidence interval decrease?

8. *Ecology: Sand Dunes* At wind speeds above 1000 cm/sec, significant sand-moving events begin to occur. Wind speeds below 1000 cm/sec deposit sand and wind speeds above 1000 cm/sec move sand to new locations. The cyclic nature of wind and moving sand determines the shape and location of large dunes (Reference: *Hydraulic, Geologic, and Biologic Research at Great Sand Dunes National Monument and Vicinity, Colorado,* Proceedings of the National Park Service Research Symposium). At a test site, the prevailing direction of the wind did not change noticeably. However, the velocity did change. Sixty wind speed readings gave an average velocity of $\bar{x} = 1075$ cm/sec. Based on long-term experience, σ can be assumed to be 265 cm/sec.

(a) Find a 95% confidence interval for the population mean wind speed at this site.

(b) Does the confidence interval indicate that the population mean wind speed is such that the sand is always moving at this site? Explain.

9. *Profits: Banks* Jobs and productivity! How do banks rate? One way to answer this question is to examine annual profits per employee. *Forbes Top Companies,* edited by J. T. Davis (John Wiley & Sons), gave the following data about annual profits per employee (in units of one thousand dollars per employee) for representative companies in financial services. Companies such as Wells Fargo, First Bank System, and Key Banks were included. Assume $\sigma \approx 10.2$ thousand dollars.

42.9	43.8	48.2	60.6	54.9	55.1	52.9	54.9	42.5	33.0	33.6
36.9	27.0	47.1	33.8	28.1	28.5	29.1	36.5	36.1	26.9	27.8
28.8	29.3	31.5	31.7	31.1	38.0	32.0	31.7	32.9	23.1	54.9
43.8	36.9	31.9	25.5	23.2	29.8	22.3	26.5	26.7		

(a) Use a calculator or appropriate computer software to verify that, for the preceding data, $\bar{x} \approx 36.0$.
(b) Let us say that the preceding data are representative of the entire sector of (successful) financial services corporations. Find a 75% confidence interval for μ, the average annual profit per employee for all successful banks.
(c) Let us say that you are the manager of a local bank with a large number of employees. Suppose the annual profits per employee are less than 30 thousand dollars per employee. Do you think that this might be somewhat low compared with other successful financial institutions? Explain by referring to the confidence interval you computed in part (b).
(d) Suppose the annual profits are more than 40 thousand dollars per employee. As manager of the bank, would you feel somewhat better? Explain by referring to the confidence interval you computed in part (b).
(e) Repeat parts (b), (c), and (d) for a 90% confidence level.

10. *Profits: Retail* Jobs and productivity! How do retail stores rate? One way to answer this question is to examine annual profits per employee. The following data give annual profits per employee (in units of one thousand dollars per employee) for companies in retail sales. (See reference in Problem 9.) Companies such as Gap, Nordstrom, Circuit City, Dillards, JCPenney, Sears, Wal-Mart, Office Depot, and Toys 'Я' Us are included. Assume $\sigma \approx 3.8$ thousand dollars.

4.4	6.5	4.2	8.9	8.7	8.1	6.1	6.0	2.6	2.9	8.1	−1.9
11.9	8.2	6.4	4.7	5.5	4.8	3.0	4.3	−6.0	1.5	2.9	4.8
−1.7	9.4	5.5	5.8	4.7	6.2	15.0	4.1	3.7	5.1	4.2	

(a) Use a calculator or appropriate computer software to verify that, for the preceding data, $\bar{x} \approx 5.1$.
(b) Let us say that the preceding data are representative of the entire sector of retail sales companies. Find an 80% confidence interval for μ, the average annual profit per employee for retail sales.
(c) Let us say that you are the manager of a retail store with a large number of employees. Suppose the annual profits per employee are less than 3 thousand dollars per employee. Do you think that this might be low compared with other retail stores? Explain by referring to the confidence interval you computed in part (b).
(d) Suppose the annual profits are more than 6.5 thousand dollars per employee. As store manager, would you feel somewhat better? Explain by referring to the confidence interval you computed in part (b).
(e) Repeat parts (b), (c), and (d) for a 95% confidence interval.

8.2 Estimating μ When σ Is Unknown

FOCUS POINTS

- ✓ Learn about degrees of freedom and Student's t distributions.
- ✓ Find critical values using degrees of freedom and confidence level.
- ✓ Compute confidence intervals for μ when σ is unknown. What does this information tell you?

In order to use the normal distribution to find confidence intervals for a population mean μ, we need to know the value of σ, the population standard deviation. However, much of the time, when μ is unknown, σ is unknown as well. In such cases, we use the sample standard deviation s to approximate σ. When we use s to approximate σ, the sampling distribution for $\bar{x}$ follows a new distribution called a *Student's t distribution.*

Student's t Distributions

Student's t distributions were discovered in 1908 by W. S. Gosset. He was employed as a statistician by Guinness brewing company, a company that discouraged publication of research by its employees. As a result, Gosset published his research under the pseudonym *Student*. Gosset was the first to recognize the importance of developing statistical methods for obtaining reliable information from samples of populations with unknown σ. Gosset used the variable t when he introduced the distribution in 1908. To this day and in his honor it is still called a Student's t distribution. It might be more fitting to call this distribution *Gosset's t distribution;* however, in the literature of mathematical statistics, it is known as a *Student's t distribution.*

The variable t is defined as follows. A Student's t distribution depends on sample size n.

Assume that x has a normal distribution with mean μ. For samples of size n with sample mean $\bar{x}$ and sample standard deviation s, the **t variable**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{7}$$

has a **Student's t distribution** with **degrees of freedom** $d.f. = n - 1$.

If many random samples of size n are drawn, then we get many t values from Equation (7). These t values can be organized into a frequency table, and a histogram can be drawn, thereby giving us an idea of the shape of the t distribution (for a given n).

Fortunately, all this work is not necessary because mathematical theorems can be used to obtain a formula for the t distribution. However, it is important to observe that these theorems state that the shape of the t distribution depends only on n, provided the basic variable x has a normal distribution. So *when we use a t distribution, we will assume that the x distribution is normal.*

Degrees of freedom

Table 4 of the Appendix gives values of the variable t corresponding to what we call the number of *degrees of freedom,* abbreviated $d.f.$ For the methods used in this section, the number of degrees of freedom is given by the formula

$$d.f. = n - 1 \tag{8}$$

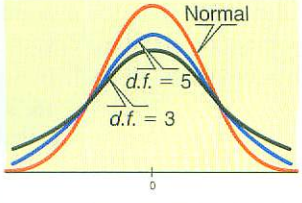
where $d.f.$ stands for the degrees of freedom and n is the sample size being used. Each choice for $d.f.$ gives a different t distribution.

The graph of a t distribution is always symmetrical about its mean, which (as for the z distribution) is 0. The main observable difference between a t distribution and the standard normal z distribution is that a t distribution has somewhat thicker tails.

Figure 8-5 shows a standard normal z distribution and a Student's t distribution with $d.f. = 3$ and $d.f. = 5$.

FIGURE 8-5

A Standard Normal Distribution and a Student's t Distribution with $d.f. = 3$ and $d.f. = 5$

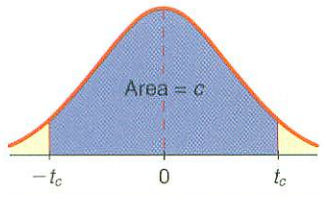


Properties of a Student's *t* distribution

1. The distribution is *symmetric* about the mean 0.
2. The distribution depends on the *degrees of freedom, d.f.* ($d.f. = n - 1$ for μ confidence intervals).
3. The distribution is *bell-shaped,* but has thicker tails than the standard normal distribution.
4. As the degrees of freedom increase, the *t* distribution *approaches* the standard normal distribution.

FIGURE 8-6

Area Under the *t* Curve Between $-t_c$ and t_c



Using Table 4 to Find Critical Values for Confidence Intervals

Table 4 of the Appendix gives various *t* values for different degrees of freedom *d.f.* We will use this table to find *critical values* t_c for a *c* confidence level. In other words, we want to find t_c such that an area equal to *c* under the *t* distribution for a given number of degrees of freedom falls between $-t_c$ and t_c. In the language of probability, we want to find t_c such that

$$P(-t_c < t < t_c) = c$$

This probability corresponds to the shaded area in Figure 8-6.

Table 4 of the Appendix has been arranged so that *c* is one of the column headings, and the degrees of freedom *d.f.* are the row headings. To find t_c for any specific *c*, we find the column headed by that *c* value and read down until we reach the row headed by the appropriate number of degrees of freedom *d.f.* (You will notice two other column headings: one-tail area and two-tail area. We will use these later, but for the time being, ignore them.)

Convention for using a Student's *t* distribution table

If the degrees of freedom *d.f.* you need are not in the table, use the closest *d.f.* in the table that is *smaller.* This procedure results in a critical value t_c that is more conservative in the sense that it is larger. The resulting confidence interval will be longer and have a probability that is slightly higher than *c*.

EXAMPLE 4
Student's t distribution

Use Table 8-3 (an excerpt from Table 4 of the Appendix) to find the critical value t_c for a 0.99 confidence level for a *t* distribution with sample size $n = 5$.

SOLUTION:

(a) First, we find the column with *c* heading 0.990.

(b) Next, we compute the number of degrees of freedom:
$d.f. = n - 1 = 5 - 1 = 4$.

(c) We read down the column under the heading $c = 0.99$ until we reach the row headed by 4 (under *d.f.*). The entry is 4.604. Therefore, $t_{0.99} = 4.604$. ◆

TABLE 8-3 Student's *t* Distribution Critical Values (Excerpt from Table 4, Appendix)

one-tail area	—	—	—	—
two-tail area	—	—	—	—
d.f. \ *c*	... 0.900	0.950	0.980	0.990 ...
⋮				
3	... 2.353	3.182	4.541	5.841 ...
4	... 2.132	2.776	3.747	4.604 ...
⋮				
7	... 1.895	2.365	2.998	3.449 ...
8	... 1.860	2.306	2.896	3.355 ...

GUIDED EXERCISE 2

Student's t distribution table

Use Table 4 of the Appendix (or Table 8-3 showing an excerpt from the table) to find t_c for a 0.90 confidence level for a t distribution with sample size $n = 9$.

(a) We find the column headed by $c =$ _____. ⇨ $c = 0.900$

(b) The degrees of freedom are given by $d.f. = n - 1 =$ _____. ⇨ $d.f. = n - 1 = 9 - 1 = 8$

(c) Read down the column found in part (a) until you reach the entry in the row headed by $d.f. = 8$. The value of $t_{0.90}$ is _____ for a sample of size 9. ⇨ $t_{0.90} = 1.860$ for a sample of size $n = 9$.

(d) Find t_c for a 0.95 confidence level for a t distribution with sample size $n = 9$. ⇨ $t_{0.95} = 2.306$ for a sample of size $n = 9$.

Maximal margin of error, *E*

In Section 8.1, we found bounds $\pm E$ on the margin of error for a c confidence level. Using the same basic approach, we arrive at the conclusion that

$$E = t_c \frac{s}{\sqrt{n}} \tag{9}$$

is the maximal margin of error for a c confidence level when σ is unknown (i.e., $|\bar{x} - \mu| < E$ with probability c). The analogue of Equation (1) in Section 8.1 is

$$P\left(-t_c \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_c \frac{s}{\sqrt{n}}\right) = c \tag{10}$$

◇ **COMMENT** Comparing Equation (10) with Equation (1) in Section 8.1, it becomes evident that we are using the same basic method on the t distribution that we did on the z distribution. ◇

Likewise, for samples from normal populations with unknown σ, Equation (4) of Section 8.1 becomes

$$P(\bar{x} - E < \mu < \bar{x} + E) = c \tag{11}$$

where $E = t_c(s/\sqrt{n})$. Let us organize what we have been doing in a convenient summary.

Confidence interval for μ with σ unknown

PROCEDURE

How to find a confidence interval for μ when σ is unknown

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n, $n > 1$) of x values from which you compute the sample mean $\bar{x}$ and the sample standard deviation s.

If you can assume that x has a normal distribution or simply a mound-shaped symmetric distribution, then any sample size n will work. If you cannot assume this, then use a sample size of $n \geq 30$.

Confidence interval for μ when σ is unknown

$$\bar{x} - E < \mu < \bar{x} + E$$

where $\bar{x}$ = sample mean of a simple random sample

$$E = t_c \frac{s}{\sqrt{n}}$$

c = confidence level $(0 < c < 1)$

t_c = critical value for confidence level c and degrees of freedom $d.f. = n - 1$
(See Table 4 of the Appendix.)

◇ **COMMENT** In our applications of Student's t distributions, we have made the basic assumption that x has a normal distribution. However, the same methods apply even if x is only approximately normal. In fact, the main requirement for using a Student's t distribution is that the distribution of x values be reasonably symmetrical and mound-shaped. If this is the case, then the methods we employ with the t distribution can be considered valid for most practical applications. ◇

EXAMPLE 5

Confidence interval for μ, σ unknown

Suppose an archaeologist discovers only seven fossil skeletons from a previously unknown species of miniature horse. Reconstructions of the skeletons of these seven miniature horses show the shoulder heights (in centimeters) to be

45.3 47.1 44.2 46.8 46.5 45.5 47.6

For these sample data, the mean is $\bar{x} \approx 46.14$ and the sample standard deviation is $s \approx 1.19$. Let μ be the mean shoulder height (in centimeters) for this entire species of miniature horse, and assume that the population of shoulder heights is approximately normal.

Find a 99% confidence interval for μ, the mean shoulder height of the entire population of such horses.

SOLUTION: In this case, $n = 7$, so $d.f. = n - 1 = 7 - 1 = 6$. For $c = 0.990$, Table 4 of the Appendix gives $t_{0.99} = 3.707$ (for $d.f. = 6$). The sample standard deviation is $s = 1.19$.

$$E = t_c \frac{s}{\sqrt{n}} = (3.707)\frac{1.19}{\sqrt{7}} \approx 1.67$$

The 99% confidence interval is

$$\bar{x} - E < \mu < \bar{x} + E$$
$$46.14 - 1.67 < \mu < 46.14 + 1.67$$
$$44.5 < \mu < 47.8$$

The archaeologist can be 99% confident that the interval from 44.5 cm to 47.8 cm is an interval that contains the population mean μ for shoulder height of this species of miniature horse. ◇

GUIDED EXERCISE 3

Confidence interval for μ, σ unknown

A company has a new process for manufacturing large artificial sapphires. In a trial run, 37 sapphires are produced. The mean weight for these 37 gems is $\bar{x} = 6.75$ carats, and the sample standard deviation is $s = 0.33$ carat. Let μ be the mean weight for the distribution of all sapphires produced by the new process.

(a) What is $d.f.$ for this setting?

➡ $d.f. = n - 1$, where n is the sample size. Since $n = 37$, $d.f. = 37 - 1 = 36$.

(b) Use Table 4 of the Appendix to find $t_{0.95}$. Note that $d.f. = 36$ is not in the table. Use the $d.f.$ closest to 36 that is smaller than 36.

➡ $d.f. = 35$ is the closest $d.f.$ in the table that is smaller than 36. Using $d.f. = 35$ and $c = 0.95$, we find $t_{0.95} = 2.030$.

(c) Find E.

➡ $$E = t_{0.95}\frac{s}{\sqrt{n}}$$
$$\approx 2.030\frac{0.33}{\sqrt{37}} \approx 0.11 \text{ carat}$$

(d) Find a 95% confidence interval for μ.

➡ $$\bar{x} - E < \mu < \bar{x} + E$$
$$6.75 - 0.11 < \mu < 6.75 + 0.11$$
$$6.64 \text{ carats} < \mu < 6.86 \text{ carats}$$

(e) Interpret the confidence interval in the context of the problem.

➡ The company can be 95% confident that the interval from 6.64 to 6.86 is an interval that contains the population mean weight of sapphires produced by the new process.

We have several formulas for confidence intervals for the population mean μ. How do we choose an appropriate one? We need to look at the sample size, the distribution of the original population, and whether or not the population standard deviation σ is known.

Summary: Confidence intervals for the mean

Assume that you have a random sample of size $n > 1$ from an x distribution and that you have computed $\bar{x}$ and s. A confidence interval for μ is

$$\bar{x} - E < \mu < \bar{x} + E$$

where E is the margin of error. How do you find E? It depends on how much you know about the x distribution.

Situation I (most common)

You don't know the population standard deviation σ. In this situation, you use the t distribution with margin of error

$$E = t_c \frac{s}{\sqrt{n}}$$

and degrees of freedom

$$d.f. = n - 1$$

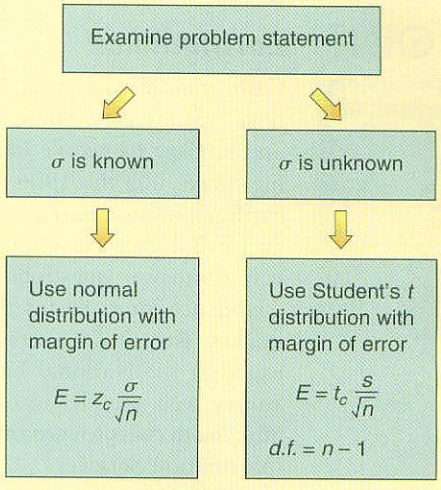
Although a t distribution can be used in many situations, you need to observe some guidelines. If n is less than 30, x should have a distribution that is mound-shaped and approximately symmetric. It's even better if the x distribution is normal. If n is 30 or more, the central limit theorem (Section 7.5) implies that these restrictions can be relaxed.

Situation II (almost never happens!)

You actually know the population value of σ. In addition, you know that x has a normal distribution. If you don't know that the x distribution is normal, then your sample size n must be 30 or larger. In this situation, you use the standard normal z distribution with margin of error

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

Which distribution should you use for $\bar{x}$?



Bootstrap

◇ **COMMENT** To find confidence intervals for μ based on small samples, you need to know that the population distribution is approximately normal. What if this is not the case? A procedure called *bootstrap* utilizes computer power to generate an approximation for the $\bar{x}$ sampling distribution. Essentially, the bootstrap method treats the sample as if it were the population. Then, using repetition, it takes many samples (often thousands) from the original sample. This process is called *resampling*. The sample mean $\bar{x}$ is computed for each resample and a distribution of sample means is created. For example, a 95% confidence interval reflects the range for the middle 95% of the bootstrap $\bar{x}$ distribution. (Reference: *An Introduction to the Bootstrap*, B. Efron and R. Tibshirani). ◇

TECH NOTE The TI-84Plus and TI-83Plus calculators, Excel, and Minitab support confidence intervals using the Student's t distribution.

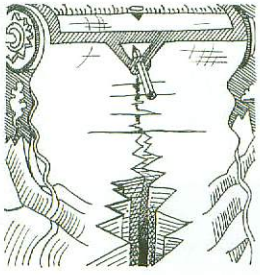
TI-84Plus/TI-83Plus Press the **STAT** key, select **TESTS**, and choose the option **8:TInterval.** You may use either raw data in a list or summary statistics.

Excel Excel gives only the value of the maximal margin of error E. You can easily construct the confidence interval by computing $\bar{x} - E$ and $\bar{x} + E$. Use the menu choices **Tools ➤ Data Analysis ➤ Describe Statistics.** In the dialogue box, check summary statistics and check confidence level for mean. Then set the desired confidence level. Under these menu choices, Excel uses the Student's t distribution.

Minitab Use the menu choices **Stat ➤ Basic Statistics ➤ 1-Sample t.** In the dialogue box, indicate the column that contains the raw data. The Minitab output shows the confidence interval for Example 5.

```
T Confidence Intervals
Variable     N       Mean      StDev     SE Mean          99.0 % CI
C1           7     46.143      1.190       0.450    (44.475, 47.810)
```

VIEWPOINT



Earthquakes!

California, Washington, Nevada, and even Yellowstone National Park all have earthquakes. Some earthquakes are severe! All earthquakes bring fear and anxiety to people living near the quake. Is San Francisco due for a really big quake like the 1906 major earthquake? How big are the sizes of recent earthquakes compared with really big earthquakes? What is the duration of an earthquake? How long is the time span between major earthquakes? One way to answer questions such as these is to use existing data to estimate confidence intervals for the average size, duration, and time interval between quakes. Recent data sets for computing such confidence intervals can be found at the National Earthquake Information Service of the U.S. Geological Survey web site. To access the site, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to National Earthquake Information Service.

SECTION 8.2 PROBLEMS

1. Use Table 4 of the Appendix to find t_c for a 0.95 confidence level when the sample size is 18.

2. Use Table 4 of the Appendix to find t_c for a 0.99 confidence level when the sample size is 4.

3. Use Table 4 of the Appendix to find t_c for a 0.90 confidence level when the sample size is 22.

4. Use Table 4 of the Appendix to find t_c for a 0.95 confidence level when the sample size is 12.

In Problems 5–11, assume that the population of x values has an approximately normal distribution. Answers may vary slightly due to rounding.

5. *Archaeology: Tree Rings* At Burnt Mesa Pueblo, the method of tree ring dating gave the following dates A.D. for an archaeological excavation site (*Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo*, edited by Kohler, Washington State University):

   1189 1271 1267 1272 1268 1316 1275 1317 1275

   (a) Use a calculator with mean and standard deviation keys to verify that the sample mean date is $\bar{x} \approx 1272$ with sample standard deviation $s \approx 37$ years.
   (b) Find a 90% confidence interval for the mean of all tree ring dates from this archaeological site.

6. *Wildlife: Wolf Pups* The number of pups in wolf dens of the southwestern United States is recorded below for 16 wolf dens (*The Wolf in the Southwest: The Making of an Endangered Species*, edited by D. E. Brown, University of Arizona Press).

   5 8 7 5 3 4 3 9
   5 8 5 6 5 6 4 7

   (a) Use a calculator with mean and standard deviation keys to verify that the sample mean is $\bar{x} \approx 5.63$ pups with sample standard deviation $s \approx 1.78$ pups.
   (b) Compute an 85% confidence interval for the population mean number of wolf pups per den in the southwestern United States.

7. *Wildlife: Mountain Lions* How much do wild mountain lions weigh? *The 77th Annual Report of the New Mexico Department of Game and Fish*, edited by Bill Montoya, gave the following information. Adult wild mountain lions (18 months or older) captured and released for the first time in the San Andres Mountains gave the following weights (lb):

   68 104 128 122 60 64

   (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} = 91.0$ lb and $s \approx 30.7$ lb.
   (b) Find a 75% confidence interval for the population average weight μ of all adult mountain lions in the specified region.

8. *Franchise: Candy Store* Do you want to own your own candy store? Wow! With some interest in running your own business and a decent credit rating, you can probably get a bank loan on startup costs for franchises such as Candy Express, The Fudge Company, Karmel Corn, and Rocky Mountain Chocolate Factory. Startup

costs (units in $1000) for a random sample of candy stores are given below (Source: *Entrepreneur Magazine,* Vol. 23, No. 10).

95 173 129 95 75 94 116 100 85

Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 106.9$ thousand dollars and $s \approx 29.4$ thousand dollars. Find a 90% confidence interval for the population average startup costs μ for candy store franchises.

9. *Diagnostic Tests: Total Calcium* Over the past several months, an adult patient has been treated for tetany (severe muscle spasms). This condition is associated with an average total calcium level below 6 mg/dl (Reference: *Manual of Laboratory and Diagnostic Tests,* F. Fischbach). Recently, the patient's total calcium tests gave the following readings (in mg/dl).

9.3 8.8 10.1 8.9 9.4 9.8 10.0
9.9 11.2 12.1

(a) Use a calculator to verify that $\bar{x} = 9.95$ and $s \approx 1.02$.
(b) Find a 99.9% confidence interval for the population mean of total calcium in this patient's blood.
(c) Based on your results in part (b), do you think this patient still has a calcium deficiency? Explain.

10. *Hospitals: Charity Care* What percentage of hospitals provide at least some charity care? The following problem is based on information taken from *State Health Care Data: Utilization, Spending, and Characteristics* (American Medical Association). Based on a random sample of hospital reports from eastern states, the following information was obtained (units in percentage of hospitals providing at least some charity care):

57.1 56.2 53.0 66.1 59.0 64.7 70.1 64.7 53.5 78.2

Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 62.3\%$ and $s \approx 8.0\%$. Find a 90% confidence interval for the population average μ of the percentage of hospitals providing at least some charity care.

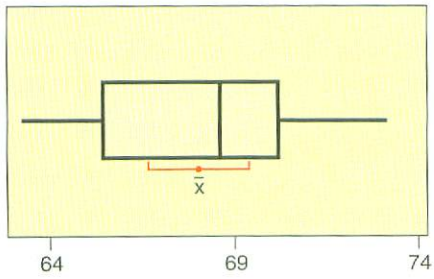
11. *Box Plots and Confidence Intervals: Heights* The distribution of heights of 18-year-old men in the United States is approximately normal with mean 68 inches and standard deviation 3 inches (U.S. Census Bureau). In Minitab we can simulate the drawing of random samples of size 20 from this population (➤ **Calc** ➤ **Random Data** ➤ **Normal** with 20 rows from a distribution with mean 68 and standard deviation 3). Then we can have Minitab compute a 95% confidence interval and draw a boxplot of the data (➤ **Stat** ➤ **Basic Statistics** ➤ **1—Sample t,** with boxplot selected in the graphs). The boxplots and confidence intervals for four different samples are shown in the accompanying figures. The four confidence intervals are

VARIABLE	N	MEAN	STDEV	SEMEAN	95.0 % CI
Sample 1	20	68.050	2.901	0.649	(66.692 , 69.407)
Sample 2	20	67.958	3.137	0.702	(66.490 , 69.426)
Sample 3	20	67.976	2.639	0.590	(66.741 , 69.211)
Sample 4	20	66.908	2.440	0.546	(65.766 , 68.050)

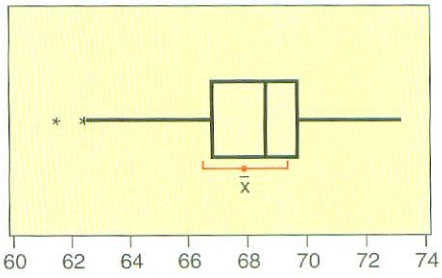
(a) Examine the figure [parts (a) to (d)]. How do the boxplots for the four samples differ? Why should you expect the boxplots to differ?

95% Confidence Intervals for Mean Height of 18-Year-Old Men (Sample size 20)

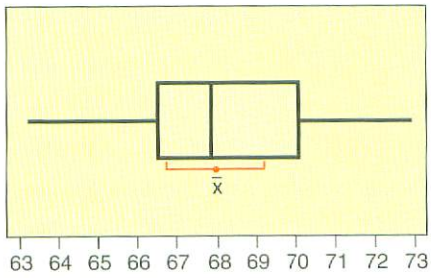
(a) Boxplot of Sample 1
(with 95% *t*-confidence interval for the mean)



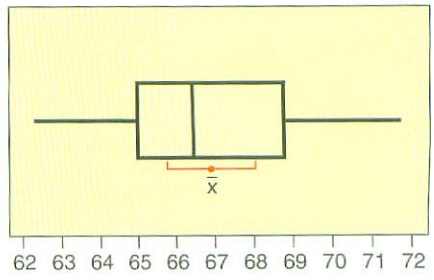
(b) Boxplot of Sample 2
(with 95% *t*-confidence interval for the mean)



(c) Boxplot of Sample 3
(with 95% *t*-confidence interval for the mean)



(d) Boxplot of Sample 4
(with 95% *t*-confidence interval for the mean)



(b) Examine the 95% confidence intervals for the four samples shown in the printout. Do the intervals differ in length? Do the intervals all contain the expected population mean of 68 inches? If we draw more samples, do you expect all of the resulting 95% confidence intervals to contain $\mu = 68$? Why or why not?

12. *Crime Rate: Denver* The following data represent crime rate per 1000 population for a random sample of 46 Denver neighborhoods (Reference: *The Piton Foundation*, Denver, Colorado).

63.2	36.3	26.2	53.2	65.3	32.0	65.0
66.3	68.9	35.2	25.1	32.5	54.0	42.4
77.5	123.2	66.3	92.7	56.9	77.1	27.5
69.2	73.8	71.5	58.5	67.2	78.6	33.2
74.9	45.1	132.1	104.7	63.2	59.6	75.7
39.2	69.9	87.5	56.0	154.2	85.5	77.5
84.7	24.2	37.5	41.1			

(a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 64.2$ and $s \approx 27.9$ crimes per 1000 population.

(b) Let us say that the preceding data are representative of the population crime rate in Denver neighborhoods. Compute an 80% confidence interval for μ, the population mean crime rate for all Denver neighborhoods.

(c) Suppose you are advising the police department about police patrol assignments. One neighborhood has a crime rate of 57 crimes per 1000 population. Do you think that this rate is below the average population crime rate and that fewer

patrols could safely be assigned to this neighborhood? Use the confidence interval to justify your answer.

(d) Another neighborhood has a crime rate of 75 crimes per 1000 population. Does this crime rate seem to be higher than the population average? Would you recommend assigning more patrols to this neighborhood? Use the confidence interval to justify your answer.

(e) Repeat parts (b), (c), and (d) for a 95% confidence interval.

(f) In previous problems, we assumed that the x distribution was normal or approximately normal. Do we need to make such an assumption in this problem? Why or why not? *Hint:* See the central limit theorem in Section 7.5.

13. *Finance: P/E Ratio* The price of a share of stock divided by the company's estimated future earnings per share is called the P/E ratio. High P/E ratios usually indicate "growth" stocks or maybe stocks that are simply overpriced. Low P/E ratios indicate "value" stocks or bargain stocks. A random sample of 51 of the largest companies in the United States gave the following P/E ratios (Reference: *Forbes*).

11	35	19	13	15	21	40	18	60	72	9	20
29	53	16	26	21	14	21	27	10	12	47	14
33	14	18	17	20	19	13	25	23	27	5	16
8	49	44	20	27	8	19	12	31	67	51	26
19	18	32									

(a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 25.2$ and $s \approx 15.5$.

(b) Find a 90% confidence interval for the P/E population mean μ of all large U.S. companies.

(c) Find a 99% confidence interval for the P/E population mean μ of all large U.S. companies.

(d) Bank One (now merged with J. P. Morgan) had a P/E of 12, AT&T Wireless had a P/E of 72, and Disney had a P/E of 24. Examine the confidence intervals in parts (b) and (c). How would you describe these stocks at this time?

(e) In previous problems, we assumed that the x distribution was normal or approximately normal. Do we need to make such an assumption in this problem? Why or why not? *Hint:* See the central limit theorem in Section 7.5.

14. *Baseball: Home Run Percentage* The home run percentage is the number of home runs per 100 times at bat. A random sample of 43 professional baseball players gave the following data for home run percentages (Reference: *The Baseball Encyclopedia*, Macmillan).

1.6	2.4	1.2	6.6	2.3	0.0	1.8	2.5	6.5	1.8
2.7	2.0	1.9	1.3	2.7	1.7	1.3	2.1	2.8	1.4
3.8	2.1	3.4	1.3	1.5	2.9	2.6	0.0	4.1	2.9
1.9	2.4	0.0	1.8	3.1	3.8	3.2	1.6	4.2	0.0
1.2	1.8	2.4							

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x} \approx 2.29$ and $s \approx 1.40$.

(b) Compute a 90% confidence interval for the population mean μ of home run percentages for all professional baseball players. *Hint:* If you use Table 4 of the Appendix, be sure to use the closest *d.f.* that is *smaller.*

(c) Compute a 99% confidence interval for the population mean μ of home run percentages for all professional baseball players.

(d) The home run percentages for three professional players are

Tim Huelett, 2.5 Herb Hunter, 2.0 Jackie Jensen, 3.8

Examine your confidence intervals and describe how home run percentages for these players compare to the population average.

(e) In previous problems, we assumed that the x distribution was normal or approximately normal. Do we need to make such an assumption in this problem? Why or why not? *Hint:* See the central limit theorem in Section 7.5.

15. *Expand Your Knowledge: Alternate Method for Confidence Intervals* When σ is unknown and the sample size is $n \geq 30$, there are two methods for computing confidence intervals for μ.

Method 1: Use the Student's t distribution with $d.f. = n - 1$.

This is the method used in the text. It is widely employed in statistical studies. Also, most statistical software packages use this method.

Method 2: When $n \geq 30$, use the sample standard deviation s as an estimate for σ, and then use the standard normal distribution.

This method is based on the fact that for large samples, s is a fairly good approximation for σ. Also, for large n, the critical values for the Student's t distribution approach those of the standard normal distribution.

Consider a random sample of size $n = 31$, with sample mean $\bar{x} = 45.2$ and sample standard deviation $s = 5.3$.

(a) Compute 90%, 95%, and 99% confidence intervals for μ using Method 1 with a Student's t distribution. Round endpoints to two digits after the decimal.

(b) Compute 90%, 95%, and 99% confidence intervals for μ using Method 2 with the standard normal distribution. Use s as an estimate for σ. Round endpoints to two digits after the decimal.

(c) Compare intervals for the two methods. Would you say that confidence intervals using a Student's t distribution are more conservative in the sense that they tend to be longer than intervals based on the standard normal distribution?

(d) Repeat parts (a) through (c) for a sample of size $n = 81$. With increased sample size, do the two methods give respective confidence intervals that are more similar?

8.3 Estimating p in the Binomial Distribution

FOCUS POINTS

- Compute the maximal margin of error for proportions using a given level of confidence.
- Compute confidence intervals for p and interpret the results.
- Interpret poll results.
- Compute the sample size to be used for estimating a proportion p when we have an estimate for p.
- Compute the sample size to be used for estimating a proportion p when we have no estimate for p.

The binomial distribution is completely determined by the number of trials n and the probability p of success in a single trial. For most experiments, the number of trials is chosen in advance. Then the distribution is completely determined by p. In this section, we will consider the problem of estimating p under the assumption that n has already been selected.

We are employing what are called *large-sample methods*. We will assume that the normal curve is a good approximation to the binomial distribution, and when necessary, we will use sample estimates for the standard deviation. Empirical studies have shown that these methods are quite good provided that *both*

$$np > 5 \quad \text{and} \quad nq > 5 \qquad \text{where } q = 1 - p \text{ is the probability of failure}$$

Let r be the number of successes out of n trials in a binomial experiment. We will take the sample proportion of successes $\hat{p}$ (read "p hat") $= r/n$ as our *point estimate for* p, the population proportion of successes.

Point estimates of p and q

The **point estimates for p and q** are

$$\hat{p} = \frac{r}{n}$$

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

where n = number of trials and r = number of successes.

For example, suppose that 800 students are selected at random from a student body of 20,000 and that they are each given shots to prevent a certain type of flu. These 800 students are then exposed to the flu, and 600 of them do not get the flu. What is the probability p that the shot will be successful for any single student selected at random from the entire population of 20,000 students? We estimate p for the entire student body by computing r/n from the sample of 800 students. The value $\hat{p} = r/n$ is 600/800, or 0.75. The value $\hat{p} = 0.75$ is then the *point estimate* for p.

Margin of error

The difference between the actual value of p and the estimate $\hat{p}$ is the size of our error caused by using $\hat{p}$ as a point estimate for p. The magnitude of $\hat{p} - p$ is called the *margin of error* for using $\hat{p} = r/n$ as a point estimate for p. In absolute value notation, the margin of error is $|\hat{p} - p|$.

To compute the bounds for the margin of error, we need some information about the distribution of $\hat{p} = r/n$ values for different samples of the same size n. It turns out that, for large samples, the distribution of $\hat{p}$ values is well approximated by a *normal curve* with

$$\text{mean } \mu = p \quad \text{and} \quad \text{standard error } \sigma = \sqrt{pq/n}$$

Since the distribution of $\hat{p} = r/n$ is approximately normal, we use features of the standard normal distribution to find the bounds for the difference $\hat{p} - p$. Recall that z_c is the number such that an area equal to c under the standard normal curve falls between $-z_c$ and z_c. Then, in terms of the language of probability,

$$P\left(-z_c\sqrt{\frac{pq}{n}} < \hat{p} - p < z_c\sqrt{\frac{pq}{n}}\right) = c \qquad (12)^*$$

Equation (12) states that the chance is c that the numerical difference between $\hat{p}$ and p is between $-z_c\sqrt{pq/n}$ and $z_c\sqrt{pq/n}$. With the c confidence level, our estimate $\hat{p}$ differs from p by no more than

Maximal margin of error, E

$$E = z_c\sqrt{pq/n}$$

As in Section 8.1, we call E the *maximal margin of error.*

*Recall from Section 7.6 that when n is large, the binomial distribution of the number of successes r is approximately normal, with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Therefore, $z = (r - np)/\sqrt{npq}$. Dividing the numerator and denominator by n shows that $\hat{p} = r/n$ has a normal distribution with mean $\mu = p$ and standard deviation $\sigma = \sqrt{pq/n}$. Beginning with the equation $P(-z_c < z < z_c) = c$, replacing z by $(\hat{p} - p)/\sqrt{pq/n}$, and multiplying all parts of the inequality by $\sqrt{pq/n}$, we obtain Equation (12).

Confidence interval for p

To find a c confidence interval for p, we will use E in place of the expression $z_c\sqrt{pq/n}$ in Equation (12). Then we get

$$P(-E < \hat{p} - p < E) = c \tag{13}$$

Some algebraic manipulation produces the mathematically equivalent statement

$$P(\hat{p} - E < p < \hat{p} + E) = c \tag{14}$$

Equation (14) states that the probability is c that p lies in the interval from $\hat{p} - E$ to $\hat{p} + E$. Therefore, the interval from $\hat{p} - E$ to $\hat{p} + E$ is the c confidence interval for p that we wanted to find.

There is one technical difficulty in computing the c confidence interval for p. The expression $E = z_c\sqrt{pq/n}$ requires that we know the values of p and q. In most situations, we will not know the actual values of p or q, so we will use our point estimates

$$p \approx \hat{p} \qquad \text{and} \qquad q = 1 - p \approx 1 - \hat{p}$$

to estimate E. These estimates are safe for most practical purposes, since we are dealing with large-sample theory ($np > 5$ and $nq > 5$).

For convenient reference, we'll summarize the information about c confidence intervals for p, the probability of success in a binomial distribution.

PROCEDURE

How to find a confidence interval for a proportion p

Consider a binomial experiment with n trials for which p represents the population probability of success and $q = 1 - p$ represents the population probability of failure. Let r be a random variable that represents the number of successes out of the n binomial trials.

The point estimates for p and q are

$$\hat{p} = \frac{r}{n} \qquad \text{and} \qquad \hat{q} = 1 - \hat{p}$$

The number of trials should be sufficiently large so that $np > 5$ and $nq > 5$. Since we do not know p or q, we use the approximations $n\hat{p} > 5$ and $n\hat{q} > 5$.

Confidence interval for p

$$\hat{p} - E < p < \hat{p} + E$$

where $E \approx z_c\sqrt{\dfrac{\hat{p}\hat{q}}{n}} = z_c\sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}}$

c = confidence level $(0 < c < 1)$

z_c = critical value for confidence level c based on the standard normal distribution (See Table 3(b) of the Appendix for frequently used values.)

EXAMPLE 6

Confidence interval for p

Let's return to our flu shot experiment described at the beginning of this section. Suppose that 800 students were selected at random from a student body of 20,000 and given shots to prevent a certain type of flu. All 800 students were exposed to the flu, and 600 of them did not get the flu. Let p represent the probability that the shot will be successful in preventing the flu for any single student selected at random from the entire population of 20,000. Let q be the probability that the shot is not successful.

(a) What is the number of trials n? What is the value of r?

SOLUTION: Since each of the 800 students receiving the shot may be thought of as a trial, then $n = 800$, and $r = 600$ is the number of successful trials.

(b) What are the point estimates for p and q?

SOLUTION: We estimate p by the sample point estimate

$$\hat{p} = \frac{r}{n} = \frac{600}{800} = 0.75$$

We estimate q by

$$\hat{q} = 1 - \hat{p} = 1 - 0.75 = 0.25$$

(c) Would it seem that the number of trials is large enough to justify a normal approximation to the binomial?

SOLUTION: Since $n = 800$, $p \approx 0.75$, and $q \approx 0.25$, we have

$$np \approx (800)(0.75) = 600 > 5 \quad \text{and} \quad nq \approx (800)(0.25) = 200 > 5$$

A normal approximation is certainly justified.

(d) Find a 99% confidence interval for p.

SOLUTION:

$z_{0.99} = 2.58$ (see Table 8-2 or Table 3(b) of the Appendix)

$$E \approx z_{0.99}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 2.58\sqrt{\frac{(0.75)(0.25)}{800}} \approx 0.0395$$

The 99% confidence interval is then

$$\hat{p} - E < p < \hat{p} + E$$
$$0.75 - 0.0395 < p < 0.75 + 0.0395$$
$$0.71 < p < 0.79$$

◆

GUIDED EXERCISE 4

Confidence interval for p

A random sample of 188 books purchased at a local bookstore showed that 66 of the books were murder mysteries. Let p represent the proportion of books sold by this store that are murder mysteries.

(a) What is a point estimate for p?

$$\hat{p} = \frac{r}{n} = \frac{66}{188} = 0.35$$

(b) Find a 90% confidence interval for p.

$$E = z_c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 1.645\sqrt{\frac{(0.35)(1-0.35)}{188}} \approx 0.0572$$

The confidence interval is

$$\hat{p} - E < p < \hat{p} + E$$

$$0.35 - 0.0572 < p < 0.35 + 0.0572$$

$$0.29 < p < 0.41$$

(c) What is the meaning of the confidence interval you just computed?

If we had computed the interval for many different sets of 188 books, we would have found that about 90% of the intervals actually contained p, the population proportion of mysteries. Consequently, we can be 90% confident that our interval is one of the intervals that contains the unknown value p.

(d) To compute the confidence interval, we used a normal approximation. Does this seem justified?

$n = 188$; $p \approx 0.35$; $q \approx 0.65$

Since $np \approx 65.8 > 5$ and $nq \approx 122.2 > 5$, the approximation is justified.

It is interesting to note that our sample point estimate $\hat{p} = r/n$ and the confidence interval for the population proportion p do not depend on the size of the population. In our bookstore example, it made no difference how many books the store sold. On the other hand, the size of the sample does affect the accuracy of a statistical estimate.

TECH NOTE The TI-84Plus and TI-83Plus calculators and Minitab provide confidence intervals for proportions.

TI-84Plus/TI-83Plus Press the **STAT** key, select **TESTS**, and choose option **A:1-PropZInt.** The letter x represents the number of successes r. The TI-84Plus/TI-83Plus output shows the results for Guided Exercise 4.

```
1-PropZInt
(.29381,.40832)
p̂=.3510638298
n=188
```

Minitab Use the menu selections **Stat ➤ Basic Statistics ➤ 1 Proportion.** In the dialogue box, select summarized data and fill in the number of trials and the number of successes. Under Options, select a confidence interval. Minitab uses the binomial distribution directly unless normal is checked. The Minitab output shows the results for Guided Exercise 4. Information relating to Chapter 9 material is also shown.

```
Test and Confidence Interval for One Proportion (Using Binomial)
Test of p = 0.5 vs p not = 0.5

                                                              Exact
Sample     X      N     Sample p          90.0 % CI          P-Value
  1       66    188    0.351064    (0.293222, 0.412466)      0.000
```

```
Test and Confidence Interval for One Proportion (Using Normal)
Test of p = 0.5 vs p not = 0.5
Sample   X    N   Sample p        90.0 % CI          Z-Value   P-Value
  1     66  188  0.351064  (0.293805, 0.408323)      -4.08     0.000
```

Interpreting Results from a Poll

Newspapers frequently report the results of an opinion poll. In articles that give more information, a statement about the margin of error accompanies the poll results. In most polls, the margin of error is given for a *95% confidence interval.*

General interpretation of poll results

1. When a poll states the results of a survey, the proportion reported to respond in the designated manner is $\hat{p}$, the sample estimate of the population proportion.
2. The *margin of error* is the maximal error E of a 95% confidence interval for p.
3. A 95% confidence interval for the population proportion p is

   poll report $\hat{p}$ − margin of error $E < p <$ poll report $\hat{p}$ + margin of error E

◇ **COMMENT:** Leslie Kish, a statistician at the University of Michigan, was the first to apply the term *margin of error.* He was a pioneer in the study of population sampling techniques. His book *Survey Sampling* is still widely used all around the world. ◇

Some articles clarify the meaning of the margin of error further by saying that it is an error due to sampling. For instance, the following comments accompany results of a political poll reported in an issue of *The Wall Street Journal.*

How Poll Was Conducted

The Wall Street Journal/NBC News poll was based on nationwide telephone interviews of 1508 adults conducted last Friday through Tuesday by the polling organizations of Peter Hart and Robert Teeter.

The sample was drawn from 315 randomly selected geographic points in the continental U.S. Each region was represented in proportion to its population. Households were selected by a method that gave all telephone numbers . . . an equal chance of being included.

One adult, 18 years or older, was selected from each household by a procedure to provide the correct number of male and female respondents.

Chances are 19 of 20 that if all adults with telephones in the U.S. had been surveyed, the findings would differ from these poll results by no more than 2.6 percentage points in either direction.

GUIDED EXERCISE 5

Reading a poll

Read the last paragraph of the article, "How Poll Was Conducted."

(a) What confidence level corresponds to the phrase "chances are 19 of 20 that if . . ."

➡ $\frac{19}{20} = 0.95$

A 95% confidence interval is being discussed.

(b) The article indicates that everyone in the sample was asked the question, "Which party, the Democratic Party or the Republican Party, do you think would do a better job handling . . . education?" Possible responses were Democrats, neither, both, or Republicans. The poll reported that 32% of the respondents said "Democrats." Does 32% represent the sample statistic $\hat{p}$ or the population parameter p for the proportion of adults responding "Democrat"?

➡ 32% represents a sample statistic $\hat{p}$ because 32% represents the percentage of the adults in the *sample* who responded "Democrats."

(c) Continue reading the last paragraph of the article. It goes on to state, ". . . if all adults with telephones in the U.S. had been surveyed, the findings would differ from these poll results by no more than 2.6 percentage points in either direction." Use this information together with parts (a) and (b) to find a 95% confidence interval for the proportion p of the specified population who would respond "Democrat" to the question.

➡ The value 2.6 percentage points represents the margin of error. Since the margin of error is for a 95% confidence interval, the confidence interval is

$$32\% - 2.6\% < p < 32\% + 2.6\%$$
$$29.4\% < p < 34.6\%$$

The poll indicates that at the time of the poll, between 29.4% and 34.6% of the specified population thought Democrats would do a better job handling education.

Sample Size for Estimating p

Suppose you want to specify the maximal margin of error in advance for a confidence interval for p at a given confidence level c. What sample size do you need? The answer depends on whether or not you have a preliminary estimate for the population probability of success p in a binomial distribution.

PROCEDURE

How to find the sample size n for estimating a proportion p

$$n = p(1 - p)\left(\frac{z_c}{E}\right)^2 \text{ if you have a preliminary estimate for } p \qquad (15)$$

$$n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 \text{ if you do } \textit{not} \text{ have a preliminary estimate for } p \qquad (16)$$

where E = specified maximal error of estimate

z_c = critical value from the normal distribution for the desired confidence level c. Commonly used values of z_c can be found in Table 3(b) of the Appendix.

If n is not a whole number, increase n to the next higher whole number. Also, if necessary, increase the sample size n to ensure that both $np > 5$ and $nq > 5$. Note that n is the minimal sample size for a specified confidence level and maximal error of estimate.

◇ **COMMENT:** To obtain Equation (15), simply solve the formula that gives the maximal error of estimate E of p for the sample size n. When you don't have an estimate for p, a little algebra can be used to show that the maximum value of $p(1 - p)$ is 1/4. ◇

EXAMPLE 7

Sample size for estimating p

A company is in the business of selling wholesale popcorn to grocery stores. The company buys directly from farmers. A buyer for the company is examining a large amount of corn from a certain farmer. Before the purchase is made, the buyer wants to estimate p, the probability that a kernel will pop.

Suppose that a random sample of n kernels is taken and r of these kernels pop. The buyer wants to be 95% sure that the point estimate $\hat{p} = r/n$ for p will be in error either way by less than 0.01.

(a) If no preliminary study is made to estimate p, how large a sample should the buyer use?

SOLUTION: In this case, we use Equation (16) with $z_{0.95} = 1.96$ (see Table 8-2) and $E = 0.01$.

$$n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 0.25(38{,}416) = 9604$$

The buyer would need a sample of $n = 9604$ kernels.

(b) A preliminary study showed that p was approximately 0.86. If the buyer uses the results of the preliminary study, how large a sample should be used?

SOLUTION: In this case, we use Equation (15) with $p \approx 0.86$. Again, from Table 8-2, $z_{0.95} = 1.96$, and from the problem, $E = 0.01$.

$$n = p(1 - p)\left(\frac{z_c}{E}\right)^2 = (0.86)(0.14)\left(\frac{1.96}{0.01}\right)^2 = 4625.29$$

The sample size should be at least $n = 4626$ kernels. This sample is less than half the sample size necessary without the preliminary study. ◆

VIEWPOINT

"Band-Aid" Surgery

Faster recovery time and less pain! Sounds great. An alternate surgical technique called *laparoscopic* ("Band-Aid") *surgery* involves small incisions in which tiny video cameras and long surgical instruments are maneuvered. Instead of a 10-inch incision, surgeons might use four little stabs of about $\frac{1}{2}$-inch in length. However, not every such surgery is successful. An article in the Health Section of *The Wall Street Journal* recommends using a surgeon who has done at least 50 such surgeries. Then the prospective patient should ask about the *rate of conversion*, that is, the proportion p of times the surgeon has been forced by complications to switch in midoperation to conventional surgery. A confidence interval for the proportion p would be useful patient information!

SECTION 8.3 PROBLEMS

For all these problems, carry at least four digits after the decimal in your calculations. Answers may vary slightly due to rounding.

1. *Myers-Briggs: Actors* Isabel Myers was a pioneer in the study of personality types. The following information is taken from *A Guide to the Development and Use of the Myers-Briggs Type Indicator*, by Myers and McCaulley (Consulting Psychologists Press). In a random sample of 62 professional actors, it was found that 39 were extroverts.
   (a) Let p represent the proportion of all actors who are extroverts. Find a point estimate for p.
   (b) Find a 95% confidence interval for p. Give a brief interpretation of the meaning of the confidence interval you have found.
   (c) Do you think that the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Explain why this would be an important consideration.

2. *Myers-Briggs: Judges* In a random sample of 519 judges, it was found that 285 were introverts (see reference of Problem 1).
   (a) Let p represent the proportion of all judges who are introverts. Find a point estimate for p.
   (b) Find a 99% confidence interval for p. Give a brief interpretation of the meaning of the confidence interval you have found.
   (c) Do you think that the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Explain why this would be an important consideration.

3. *Navajo Lifestyle: Traditional Hogans* A random sample of 5222 permanent dwellings on the entire Navajo Indian Reservation showed that 1619 were traditional Navajo hogans (*Navajo Architecture: Forms, History, Distributions,* by Jett and Spencer, University of Arizona Press).
   (a) Let p be the proportion of all permanent dwellings on the entire Navajo Reservation that are traditional hogans. Find a point estimate for p.
   (b) Find a 99% confidence interval for p. Give a brief interpretation of the confidence interval.
   (c) Do you think that $np > 5$ and $nq > 5$ are satisfied for this problem? Explain why this would be an important consideration.

4. *Archaeology: Pottery* Santa Fe black-on-white is a type of pottery commonly found at archaeological excavations in Bandelier National Monument. At one excavation site, a sample of 592 potsherds was found, of which 360 were identified as Santa Fe black-on-white (*Bandelier Archaeological Excavation Project: Summer 1990 Excavations at Burnt Mesa Pueblo and Casa del Rito,* edited by Kohler and Root, Washington State University).
   (a) Let p represent the population proportion of Santa Fe black-on-white potsherds at the excavation site. Find a point estimate for p.
   (b) Find a 95% confidence interval for p. Give a brief statement of the meaning of the confidence interval.
   (c) Do you think that the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Why would this be important?

5. *Health Care: Colorado Physicians* A random sample of 5792 physicians in Colorado showed that 3139 provided at least some charity care (i.e., treated poor people at no cost). These data are based on information from *State Health Care Data: Utilization, Spending, and Characteristics* (American Medical Association).
   (a) Let p represent the proportion of all Colorado physicians who provide some charity care. Find a point estimate for p.
   (b) Find a 99% confidence interval for p. Give a brief explanation of the meaning of your answer in the context of this problem.
   (c) Is the normal approximation to the binomial justified in this problem? Explain.

6. *Law Enforcement: Escaped Convicts* Case studies showed that out of 10,351 convicts who escaped from U.S. prisons, only 7867 were recaptured (*The Book of Odds,* by Shook and Shook, Signet).
   (a) Let p represent the proportion of all escaped convicts who will eventually be recaptured. Find a point estimate for p.
   (b) Find a 99% confidence interval for p. Give a brief statement of the meaning of the confidence interval.
   (c) Is use of the normal approximation to the binomial justified in this problem? Explain.

7. *Fishing: Barbless Hooks* In a combined study of northern pike, cutthroat trout, rainbow trout, and lake trout, it was found that 26 out of 855 fish died when caught and released using barbless hooks on flies or lures. All hooks were removed from the fish (Source: *A National Symposium on Catch and Release Fishing,* Humboldt State University Press).
   (a) Let p represent the proportion of all pike and trout that die (i.e., p is the mortality rate) when caught and released using barbless hooks. Find a point estimate for p.
   (b) Find a 99% confidence interval for p, and give a brief explanation of the meaning of the interval.
   (c) Is the normal approximation to the binomial justified in this problem? Explain.

8. *Focus Problem: Trick or Treat* In a survey of a random sample of 35 households in the Cherry Creek neighborhood of Denver, it was found that 11 households turned out the lights and pretended not to be home on Halloween.
   (a) Compute a 90% confidence interval for p, the proportion of all households in Cherry Creek that pretend not to be home on Halloween.
   (b) What assumptions are necessary to calculate the confidence interval of part (a)?
   (c) The national proportion of all households in the United States that turn out the lights and pretend not to be home on Halloween is 0.28. Is 0.28 in the confidence interval you computed? Based on your answer, does it seem that the Cherry Creek neighborhood is much different (either higher or lower proportion) from the population of all U.S. households? Explain.

9. *Marketing: Customer Loyalty* In a marketing survey, a random sample of 730 women shoppers revealed that 628 remained loyal to their favorite supermarket during the past year—i.e., did not switch stores (Source: *Trends in the United States: Consumer Attitudes and the Supermarket*, The Research Department, Food Marketing Institute).
   (a) Let p represent the proportion of all women shoppers who remain loyal to their favorite supermarket. Find a point estimate for p.
   (b) Find a 95% confidence interval for p. Give a brief explanation of the meaning of the interval.
   (c) As a news writer, how would you report the survey results regarding the percentage of supermarket shoppers who remained loyal to their favorite supermarket during the past year? What is the margin of error based on a 95% confidence interval?

10. *Marketing: Bargain Hunters* In a marketing survey, a random sample of 1001 supermarket shoppers revealed that 273 always stock up on an item when they find that item at a real bargain price. See reference in Problem 9.
   (a) Let p represent the proportion of all supermarket shoppers who always stock up on an item when they find a real bargain. Find a point estimate for p.
   (b) Find a 95% confidence interval for p. Give a brief explanation of the meaning of the interval.
   (c) As a news writer, how would you report the survey results on the percentage of supermarket shoppers who stock up on items when they find the item is a real bargain? What is the margin of error based on a 95% confidence interval?

11. *Lifestyle: Smoking* In a survey of 1000 large corporations, 250 said that, given a choice between a job candidate who smokes and an equally qualified nonsmoker, the nonsmoker would get the job (*USA Today*).
   (a) Let p represent the proportion of all corporations preferring a nonsmoking candidate. Find a point estimate for p.
   (b) Find a 0.95 confidence interval for p.
   (c) As a news writer, how would you report the survey results regarding the proportion of corporations that would hire the equally qualified nonsmoker? What is the margin of error based on a 95% confidence interval?

12. *Opinion Poll: Crime and Violence* A *New York Times*/CBS poll asked the question, "What do you think is the most important problem facing this country today?" Nineteen percent of the respondents answered "crime and violence." The margin of sampling error was plus or minus 3 percentage points. Following the convention that the margin of error is based on a 95% confidence interval, find a 95% confidence interval for the percentage of the population that would respond "crime and violence" to the question asked by the pollsters.

13. *Medical: Blood Type* A random sample of medical files is used to estimate the proportion p of all people who have blood type B.
    (a) If you have no preliminary estimate for p, how many medical files should you include in a random sample in order to be 85% sure that the point estimate $\hat{p}$ will be within a distance of 0.05 from p?
    (b) Answer part (a) if you use the preliminary estimate that about 8 out of 90 people have blood type B (Reference: *Manual of Laboratory and Diagnostic Tests*, F. Fischbach).

14. *Business: Phone Contact* How hard is it to reach a businessperson by phone? Let p be the proportion of calls to businesspeople for which the caller reaches the person being called on the *first* try.
    (a) If you have no preliminary estimate for p, how many business phone calls should you include in a random sample to be 80% sure that the point estimate $\hat{p}$ will be within a distance of 0.03 from p?
    (b) The *Book of Odds*, by Shook and Shook (Signet), reports that businesspeople can be reached by a single phone call approximately 17% of the time. Using this (national) estimate for p, answer part (a).

15. *Campus Life: Coeds* What percentage of the campus student body is female? Let p be the proportion of women students on your campus.
    (a) If no preliminary study is made to estimate p, how large a sample is needed to be 99% sure that a point estimate $\hat{p}$ will be within a distance of 0.05 from p?
    (b) The *Statistical Abstract of the United States*, 112th Edition, indicates that approximately 54% of college students are females. Answer part (a) using this estimate for p.

16. *Small Business: Bankruptcy* The National Council of Small Businesses is interested in the proportion of small businesses that declared Chapter 11 bankruptcy last year. Since there are so many small businesses, the National Council intends to estimate the proportion from a random sample. Let p be the proportion of small businesses that declared Chapter 11 bankruptcy last year.
    (a) If no preliminary sample is taken to estimate p, how large a sample is necessary to be 95% sure that a point estimate $\hat{p}$ will be within a distance of 0.10 from p?
    (b) In a preliminary random sample of 38 small businesses, it was found that six had declared Chapter 11 bankruptcy. How many *more* small businesses should be included in the sample to be 95% sure that a point estimate $\hat{p}$ will be within a distance of 0.10 from p?

SUMMARY

How do you get information about a population by looking at a sample? One way is to use point estimates and confidence intervals. In this chapter, you studied point estimates and confidence intervals for population parameters μ and p. The respective point estimates are $\bar{x}$ and $\hat{p}$. Confidence intervals are created by subtracting and adding the margin of error E for a specified confidence level.

The general structure for a c confidence interval is

$$\text{point estimate} - E < \text{parameter} < \text{point estimate} + E$$

Specific formulas for E and how to use them to construct confidence intervals at the level of

Confidence intervals for p require large-sample techniques so that the standard normal distribution can be used for critical values.

Confidence intervals have an associated probability c called the confidence level. The probability is c that the c confidence interval you compute is one of the many possible intervals containing the population parameter.

IMPORTANT WORDS & SYMBOLS

Section 8.1

Large samples, $n \geq 30$
Maximal margin of error E
Confidence level c
Critical values z_c
Point estimate for μ
Confidence interval for μ
c confidence interval

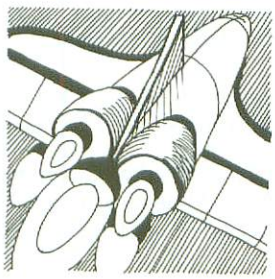
Section 8.2

Student's t variable
Degrees of freedom ($d.f.$)
Critical values t_c

Section 8.3

Point estimate for p, $\hat{p}$
Confidence interval for p
Margin of error for polls

VIEWPOINT



All Systems Go?

On January 28, 1986, the Space Shuttle *Challenger* caught fire and blew up only seconds after launch. A great deal of good engineering went into the design of the *Challenger*. However, when a system has several confidence levels operating at once, it can happen, in rare cases, that risks will increase rather than cancel out. Diane Vaughn is a professor of sociology at Boston College and author of the book *The Challenger Launch Decision* (University of Chicago Press). Her book contains an excellent discussion of risks, the normalization of deviants, and cost/safety tradeoffs. Vaughn's book is described as "a remarkable and important analysis of how social structures can induce consequential errors in a decision process" (Robert K. Merton, Columbia University).

CHAPTER REVIEW PROBLEMS

1. In your own words, carefully explain the meaning of the following terms: point estimate, critical value, maximal margin of error, confidence level, and confidence interval.

For Problems 2–10, categorize each problem according to the parameter being estimated, mean μ or proportion p. Then solve the problem.

2. *Auto Insurance: Claims* Anystate Auto Insurance Company took a random sample of 370 insurance claims paid out during a 1-year period. The average claim paid was \$1570. Assume $\sigma = \$250$. Find 0.90 and 0.99 confidence intervals for the mean claim payment.

3. *Psychology: Closure* Three experiments investigating the relation between need for cognitive closure and persuasion were reported in "Motivated Resistance and Openness to Persuasion in the Presence or Absence of Prior Information," by A. W. Kruglanski (*Journal of Personality and Social Psychology*, Vol. 65, No. 5, pp. 861–874). Part of the study involved administering a "need for closure scale" to a group of students enrolled in an introductory psychology course. The "need for closure scale" has scores ranging from 101 to 201. For the 73 students in the highest quartile of the distribution, the mean score was $\bar{x} = 178.70$. Assume the population standard deviation $\sigma = 7.81$. These students were all classified as high on their need for closure. Assume that the 73 students represent a random sample of all students who are classified as high on their need for closure. Find a 95% confidence interval for the population mean score μ on the "need for closure scale" for all students with a high need for closure.

4. *Psychology: Closure* How large a sample is needed in Problem 3 if we wish to be 99% confident that the sample mean score is within 2 points of the population mean score for students who are high on the need for closure?

5. *Archaeology: Excavations* The Wind Mountain archaeological site is located in southwestern New Mexico. Wind Mountain was home to an ancient culture of prehistoric Native Americans called Anasazi. A random sample of excavations at Wind Mountain gave the following depths (in centimeters) from present-day surface grade to the location of significant archaeological artifacts (Source: *Mimbres Mogollon Archaeology*, by A. Woosley and A. McIntyre, University of New Mexico Press).

85	45	120	80	75	55	65	60
65	95	90	70	75	65	68	

   (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 74.2$ cm and $s \approx 18.3$ cm.
   (b) Compute a 95% confidence interval for the mean depth μ at which archaeological artifacts from the Wind Mountain excavation site can be found.

6. *Archaeology: Pottery* Sherds of clay vessels were put together to reconstruct rim diameters of the original ceramic vessels at the Wind Mountain archaeological site (see source in Problem 5). A random sample of ceramic vessels gave the following rim diameters (in centimeters):

15.9	13.4	22.1	12.7	13.1	19.6	11.7	13.5	17.7	18.1

   (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 15.8$ cm and $s \approx 3.5$ cm.
   (b) Compute an 80% confidence interval for the population mean μ of rim diameters for such ceramic vessels found at the Wind Mountain archaeological site.

7. *Telephone Interviews: Survey* The National Study of the Changing Work Force conducted an extensive survey of 2958 wage and salaried workers on issues ranging from relationship with their bosses to household chores. The data were gathered through hour-long telephone interviews with a nationally representative sample (*The Wall Street Journal*). In response to the question, "What does success mean to you?" 1538 responded, "Personal satisfaction from doing a good job." Let p be the population proportion of all wage and salaried workers who would respond the same way to the stated question. Find a 90% confidence interval for p.

8. *Telephone Interviews: Survey* How large a sample is needed in Problem 7 if we wish to be 95% confident that the sample percentage of those equating success with personal satisfaction is within 1% of the population percentage? (*Hint:* Use $p \approx 0.52$ as a preliminary estimate.)

9. *Archaeology: Pottery* Three-circle, red-on-white is one distinctive pattern painted on ceramic vessels of the Anasazi period found at the Wind Mountain archaeological site (see source for Problem 5). At one excavation, a sample of 167 potsherds indicated that 68 were of the three-circle, red-on-white pattern.
   (a) Find a point estimate $\hat{p}$ for the proportion of all ceramic potsherds at this site that are of the three-circle, red-on-white pattern.
   (b) Compute a 95% confidence interval for the population proportion p of all ceramic potsherds with this distinctive pattern found at the site.

10. *Archaeology: Pottery* Consider the three-circle, red-on-white pattern discussed in Problem 9. How many ceramic potsherds must be found and identified if we are to be 95% confident that the sample proportion $\hat{p}$ of such potsherds is within 6% of the population proportion of three-circle, red-on-white patterns found at this excavation site? (*Hint:* Use the results of Problem 9 as a preliminary estimate.)

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Digging clams

1. Garrison Bay is a small bay in Washington state. A popular recreational activity in the bay is clam digging. For several years, the clam harvest has been monitored and the size distribution of clams recorded. Data for lengths and widths of little neck clams (*Protothaca staminea*) were recorded by a method of systematic sampling in a study done by S. Scherba and V. F. Gallucci ("The Application of Systematic Sampling to a Study of Infaunal Variation in a Soft Substrate Intertidal Environment," *Fishery Bulletin* 74:937–948). The data in Tables 8-4 and 8-5 give lengths and widths for 35 little neck clams.
   (a) Use a calculator to compute the sample mean and sample standard deviation for the lengths and widths. Compute the coefficient of variation for each.
   (b) Compute a 95% confidence interval for the population mean length of all Garrison Bay little neck clams.
   (c) How many more little neck clams would be needed in a sample if you wanted to be 95% sure that the sample mean length is within a maximal margin of error of 10 mm of the population mean length?
   (d) Compute a 95% confidence interval for the population mean width of all Garrison Bay little neck clams.

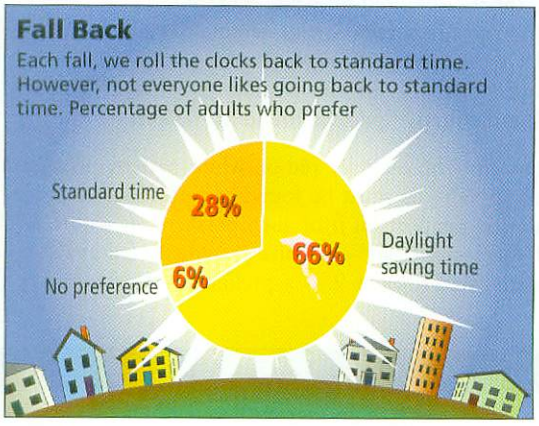
TABLE 8-4 Lengths of Little Neck Clams (mm)

530	517	505	512	487	481	485	479	452	468
459	449	472	471	455	394	475	335	508	486
474	465	420	402	410	393	389	330	305	169
91	537	519	509	511					

TABLE 8-5 Widths of Little Neck Clams (mm)

494	477	471	413	407	427	408	430	395	417
394	397	402	401	385	338	422	288	464	436
414	402	383	340	349	333	356	268	264	141
77	498	456	433	447					

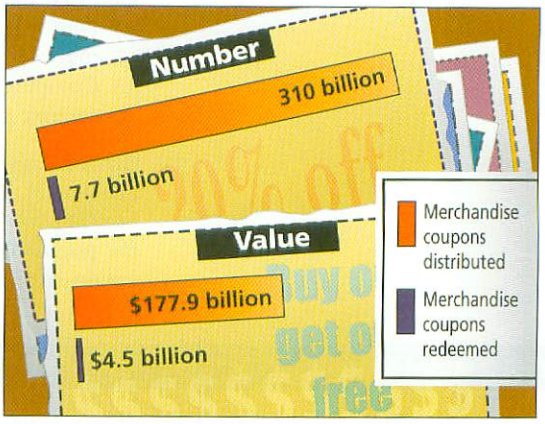
FIGURE 8-7



Source: Hilton Time Survey of 1024 adults

FIGURE 8-8

Coupons: Limited Use



Source: NCH Promotional Services

(e) How many more little neck clams would be needed in a sample if you wanted to be 95% sure that the sample mean width is within a maximal margin of error of 10 mm of the population mean width?

2. Examine Figure 8-7, "Fall Back."
   (a) Of the 1024 adults surveyed, 66% were reported to favor daylight saving time. How many people in the sample preferred daylight saving time? Using the statistic $\hat{p} = 0.66$ and sample size $n = 1024$, find a 95% confidence interval for the proportion of people p who favor daylight saving time. How could you report this information in terms of a margin of error?
   (b) Look at Figure 8-7 to find the sample statistic $\hat{p}$ for the proportion of people preferring standard time. Find a 95% confidence interval for the population proportion p of people who favor standard time. Report the same information in terms of a margin of error.

3. Examine Figure 8-8, "Coupons: Limited Use."
   (a) Use Figure 8-8 to estimate the percentage of merchandise coupons that were redeemed. Also estimate the percentage dollar value of the coupons that were redeemed. Are these numbers approximately equal?
   (b) Suppose that you are a marketing executive working for a national chain of toy stores. You wish to estimate the percentage of coupons that will be redeemed for the toy stores. How many coupons should you check to be 95% sure that the percentage of coupons redeemed is within 1% of the population proportion of all coupons redeemed for the toy store?
   (c) Use the results of part (a) as a preliminary estimate for p, the percentage of coupons that are redeemed, and redo part (b).
   (d) Suppose that you sent out 937 coupons and found that 27 were redeemed. Explain why you could be 95% confident that the proportion of such coupons redeemed in the future would be between 1.9% and 3.9%.
   (e) Suppose that the dollar value of a collection of coupons was $10,000. Use the data in Figure 8-8 to find the expected value and standard deviation of the dollar value of the redeemed coupons. What is the probability that between $225 and $275 (out of the $10,000) is redeemed?

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. In this chapter, we have studied confidence intervals. Carefully read the following statements about confidence intervals:
   (a) Once the endpoints of a confidence interval are numerically fixed, then the parameter in question (either μ or p) does or does not fall inside the "fixed" interval.
   (b) A given fixed interval either does or does not contain the parameter μ or p; therefore, the probability is 1 or 0 that the parameter is in the interval.

   Next, read the following statements. Then discuss all four statements in the context of what we actually mean by a confidence interval.
   (c) Nontrivial probability statements can be made only about variables, not constants.
   (d) The confidence level c represents the proportion of all (fixed) intervals that would contain the parameter if we repeated the process many, many times.

2. Throughout Chapter 8, we have used the normal distribution, the central limit theorem, or the Student's t distribution.
   (a) Give a brief outline describing how confidence intervals for means use the central limit theorem or Student's t distribution in their basic construction.
   (b) Give a brief outline describing how the normal approximation to the binomial distribution is used in the construction of confidence intervals for a proportion p.
   (c) Give a brief outline describing how the sample size for a predetermined error tolerance and level of confidence is determined from the normal distribution or the central limit theorem.

3. When the results of a survey or a poll are published, the sample size is usually given, as well as the margin of error. For example, suppose the *Honolulu Star Bulletin* reported that it surveyed 385 Honolulu residents and 78% said they favor mandatory jail sentences for people convicted of driving under the influence of drugs or alcohol (with margin of error of 3 percentage points in either direction). Usually the confidence level of the interval is not given, but it is standard practice to use the margin of error for a 95% confidence interval when no other confidence level is given.
   (a) The paper reported a point estimate of 78% with margin of error of $\pm 3\%$. Write this information in the form of a confidence interval for p, the population proportion of residents favoring mandatory jail sentences for people convicted of driving under the influence. What is the assumed confidence level?
   (b) The margin of error is simply the error due to using a sample instead of the entire population. It does not take into account the bias that might be introduced by the wording of the question, by the truthfulness of the respondents, or by other factors. Suppose the question was asked in this fashion: "Considering the devastating injuries suffered by innocent victims in auto accidents caused by drunken or drugged drivers, do you favor a mandatory jail sentence for those convicted of driving under the influence of drugs or alcohol?" Do you think the wording of the question would influence the respondents? Do you think the population proportion of those favoring mandatory jail sentences is accurately represented by a confidence interval based on responses to such a question? Explain your answer.

   Suppose the question had been: "Considering the existing overcrowding of our prisons, do you favor a mandatory jail sentence for *people convicted* of driving under the influence of drugs or alcohol?" Do you think the population proportion of those favoring mandatory jail sentences is accurately represented by a confidence interval based on responses to such a question? Explain your answer.

Using Technology

TI-84PLUS/TI-83PLUS • EXCEL • MINITAB • SPSS

APPLICATION

Confidence Interval Demonstration

When we generate different random samples of the same size from a population, we discover that $\bar{x}$ varies from sample to sample. Likewise, different samples produce different confidence intervals for μ. The endpoints $\bar{x} \pm E$ of a confidence interval are statistical variables. A 90% confidence interval tells us that if we obtain lots of confidence intervals (for the same sample size), then the proportion of all intervals that will turn out to contain μ is 90%.

(a) Use the technology of your choice to generate 10 large random samples from a population with a known mean μ.

(b) Construct a 90% confidence interval for the mean for each sample.

(c) Examine the confidence intervals and note the percentage of the intervals that contain the population mean μ. We have 10 confidence intervals. Will exactly 90% of 10 intervals always contain μ? Explain. What if we have 1000 intervals?

Technology Hints for Confidence Interval Demonstration

TI-84Plus/TI-83Plus

The TI-84Plus/TI-83Plus generates random samples from uniform, normal, and binomial distributions. Press the MATH key and select PRB. Choice **5:randInt(lower, upper, sample size *n*)** generates random samples of size n from the integers between the specified lower and upper values. Choice **6:randNorm(μ, σ, sample size n)** generates random samples of size n from a normal distribution with specified mean and standard deviation. Choice **7:randBin(number of trials, p, sample size)** generates samples of the specified size from the designated binomial distribution. Under STAT, select EDIT and highlight the list name such as L1. At the = sign, use the MATH key to access the desired population distribution. Finally, use the **Zinterval** under the TESTS option of the STAT key to generate 90% confidence intervals.

Excel

Use the menu choices **Tools ➤ Data Analysis ➤ Random Number Generator.** In the dialogue box, the number of variables refers to the number of samples. The number of random numbers refers to the number of data in each sample. Select the population distribution (uniform, normal, binomial). The command **Paste function** f_x **➤ Statistical ➤ Confidence(1 − confidence level, σ, sample size)** gives the maximal margin of error E. To find a 90% confidence interval for each sample, use **Confidence(0.10, σ, sample size)** to find the maximal margin of error E. Note that if you use the population standard deviation σ in the function, the value of E will be the same for all samples of the same size. Next, find the sample mean $\bar{x}$ for each sample (use **Paste function** f_x **➤ Statistical ➤ Average**). Finally, construct the endpoints $\bar{x} \pm E$ of the confidence interval for each sample.

Minitab

Minitab provides options for sampling from a variety of distributions. To generate random samples from a specific distribution, use the menu selection **Calc ➤ Random Data ➤** and then select the population distribution. In the dialogue box, the *number of rows of data* represents the *sample size*. The *number of samples* corresponds to the number of columns selected for data storage. For example, c1−c10 in data storage produces 10 different random samples of the specified size. Use the

menu selection **Stat ➤ Basic Statistics ➤ 1 sample z** to generate confidence intervals for the mean μ from each sample. In the variables box, list all the columns containing your samples. For instance, using c1−c10 in the variables list will produce confidence intervals for each of the 10 samples stored in columns c1 through c10.

The Minitab display shows 90% confidence intervals for 10 different random samples of size 50 taken from a normal distribution with $\mu = 30$ and $\sigma = 4$. Notice that, as expected, 9 out of 10 of the intervals contain $\mu = 30$.

Minitab Display

Z Confidence Intervals (Samples from a Normal Population with $\mu = 30$ and $\sigma = 4$)

The assumed sigma = 4.00

Variable	N	Mean	StDev	SE Mean	90.0 % CI
C1	50	30.265	4.300	0.566	(29.334, 31.195)
C2	50	31.040	3.957	0.566	(30.109, 31.971)
C3	50	29.940	4.195	0.566	(29.010, 30.871)
C4	50	30.753	3.842	0.566	(29.823, 31.684)
C5	50	30.047	4.174	0.566	(29.116, 30.977)
C6	50	29.254	4.423	0.566	(28.324, 30.185)
C7	50	29.062	4.532	0.566	(28.131, 29.992)
C8	50	29.344	4.487	0.566	(28.414, 30.275)
C9	50	30.062	4.199	0.566	(29.131, 30.992)
C10	50	29.989	3.451	0.566	(29.058, 30.919)

SPSS

SPSS uses a Student's t distribution to generate confidence intervals for the mean. Use the menu choices **Analyze ➤ Compare Means** and then **One-Sample T Test** for confidence intervals for a single mean. In the dialogue box, use 0 for the test value. Click **Options...** to provide the confidence level.

To generate 10 random samples of size $n = 30$ from a normal distribution with $\mu = 30$ and $\sigma = 4$, first enter consecutive integers from 1 to 30 in a column of the data editor. Then, under variable view, enter the variable names Sample1 through Sample10. Use the menu choices **Transform ➤ Compute.** In the dialogue box, use Sample1 for the target variable, then select the function **RV.Normal(mean, stddev).** Use 30 for the mean and 4 for the standard deviation. Continue until you have 10 samples. To sample from other distributions, use appropriate functions in the Compute dialogue box.

The SPSS display shows 90% confidence intervals for 10 different random samples of size 30 taken from a normal distribution with $\mu = 30$ and $\sigma = 4$. Notice that, as expected, 9 of the 10 intervals contain the population mean $\mu = 30$.

SPSS Display

90% t-confidence intervals for random samples of size $n = 30$ from a normal distribution with $\mu = 30$ and $\sigma = 4$.

	t	df	Sig(2-tail)	Mean	Lower	Upper
SAMPLE1	42.304	29	.000	29.7149	28.5214	30.9084
SAMPLE2	43.374	29	.000	30.1552	28.9739	31.3365
SAMPLE3	53.606	29	.000	31.2743	30.2830	32.2656
SAMPLE4	35.648	29	.000	30.1490	28.7120	31.5860
SAMPLE5	47.964	29	.000	31.0161	29.9173	32.1148
SAMPLE6	34.718	29	.000	30.3519	28.8665	31.8374
SAMPLE7	34.698	29	.000	30.7665	29.2599	32.2731
SAMPLE8	39.731	29	.000	30.2388	28.9456	31.5320
SAMPLE9	44.206	29	.000	29.7256	28.5831	30.8681
SAMPLE10	49.981	29	.000	29.7273	28.7167	30.7379