7

Rusty Jarrett/Getty Images for NASCAR/ Getty Images

7.1 Estimating μ When σ Is Known

7.2 Estimating μ When σ Is Unknown

7.3 Estimating p in the Binomial Distribution

7.4 Estimating $\mu_1 - \mu_2$ and $p_1 - p_2$

Library of Congress

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

—Robert Frost, "The Secret Sits"*

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 *random 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," to quote the words of the celebrated American poet Robert Lee Frost (1874–1963). 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.

* "The Secret Sits," 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, Inc. Reprinted by permission of Henry Holt and Company, Inc.

For online student resources, visit The Brase/Brase, *Understandable Statistics,* 10th edition web site at **http://www.cengage.com/statistics/brase**.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Estimation

Bobby Deal / RealDealPhoto, 2010/Used under license from Shutterstock.Com

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 7.1)

At the beginning design stage of a statistical project, how large a sample size should you plan to get? (Section 7.1)

What famous statistician worked for Guinness brewing company in Ireland? What has this got to do with constructing estimates from sample data? (Section 7.2)

How do you estimate the proportion p of successes in a binomial experiment? How does the normal approximation fit into this process? (Section 7.3)

Sometimes even small differences can be extremely important. How do you estimate differences? (Section 7.4)

FOCUS PROBLEM

The Trouble with Wood Ducks

The National Wildlife Federation published an article entitled "The Trouble with Wood Ducks" (*National Wildlife*, Vol. 31, No. 5). In this article, wood ducks are described as beautiful birds living in forested areas such as the Pacific Northwest and southeast United States. Because of overhunting and habitat destruction, these birds were in danger of extinction. A federal ban on hunting wood ducks in 1918 helped save the species from extinction. Wood ducks like to nest in tree cavities. However, many such trees were disappearing due to heavy timber cutting. For a period of time it seemed that nesting boxes were the solution to disappearing trees. At first, the wood duck population grew, but after a few seasons, the population declined sharply. Good biology research combined with good statistics provided an answer to this disturbing phenomenon.

Cornell University professors of ecology Paul Sherman and Brad Semel found that the nesting boxes were placed too close to each other. Female wood ducks prefer a secluded nest that is a considerable distance from the next wood duck nest. In fact, female wood duck behavior changed when the nests were too close to each other. Some females would lay their

Photo Researchers

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

eggs in another female's nest. The result was too many eggs in one nest. The biologists found that if there were too many eggs in a nest, the proportion of eggs that hatched was considerably reduced. In the long run, this meant a decline in the population of wood ducks.

In their study, Sherman and Semel used two placements of nesting boxes. Group I boxes were well separated from each other and well hidden by available brush. Group II boxes were highly visible and grouped closely together.

In group I boxes, there were a total of 474 eggs, of which a field count showed that about 270 hatched. In group II boxes, there were a total of 805 eggs, of which a field count showed that, again, about 270 hatched.

The material in Chapter 7 will enable us to answer many questions about the hatch ratios of eggs from nests in the two groups.

(a) Find a point estimate $\hat{p}_1$ for p_1, the proportion of eggs that hatch in group I nest box placements. Find a 95% confidence interval for p_1.

(b) Find a point estimate $\hat{p}_2$ for p_2, the proportion of eggs that hatch in group II nest box placements. Find a 95% confidence interval for p_2.

(c) Find a 95% confidence interval for $p_1 - p_2$. Does the interval indicate that the proportion of eggs hatched from group I nest box placements is higher than, lower than, or equal to the proportion of eggs hatched from group II nest boxes?

(d) What conclusions about placement of nest boxes can be drawn? In the article, additional concerns are raised about the higher cost of placing and maintaining group I nest boxes. Also at issue is the cost efficiency per successful wood duck hatch. Data in the article do not include information that would help us answer questions of cost efficiency. However, the data presented do help us answer questions about the proportions of successful hatches in the two nest box configurations. (See Problem 26 of Section 7.4.)

SECTION 7.1 Estimating μ When σ Is Known

FOCUS POINTS

- Explain the meanings 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*.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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

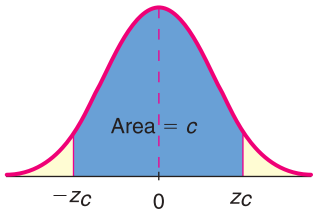
Even with a large random sample, the value of $\bar{x}$ usually is not *exactly* equal to the population mean μ. The *margin of error* is the magnitude of the difference between the sample point estimate and the true population parameter value.

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

Confidence level, c

FIGURE 7-1

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 7-1). Theoretically, we 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 5 of Appendix II to find a number $z_{0.99}$ such that 99% of the area under the standard curve lies between $-z_{0.99}$ and $z_{0.99}$. That is, we will find $z_{0.99}$ such that

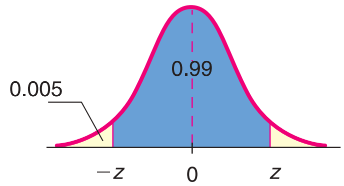
$$P(-z_{0.99} < z < z_{0.99}) = 0.99$$

SOLUTION: In Section 6.3, 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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

FIGURE 7-2

Area Between $-z$ and z Is 0.99



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 7-2).

Next, we use Table 5 of Appendix II to find the z value corresponding to a left-tail area of 0.0050. Table 7-1 shows an excerpt from Table 5 of Appendix II.

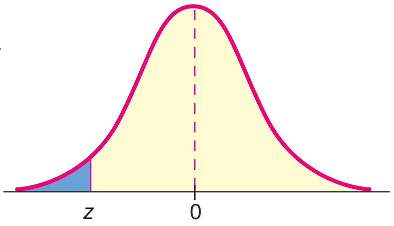


TABLE 7-1 **Excerpt from Table 5 of Appendix II**

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

From Table 7-1, 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

$$P(-2.58 < z < 2.58) \approx 0.99$$

The results of Example 1 will be used a great deal in our later work. For convenience, Table 7-2 gives some levels of confidence and corresponding critical values z_c. The same information is provided in Table 5(b) of Appendix II.

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 6.1.)

This information, together with our work on confidence levels, leads us (as shown in the optional derivation that follows) to the probability statement

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

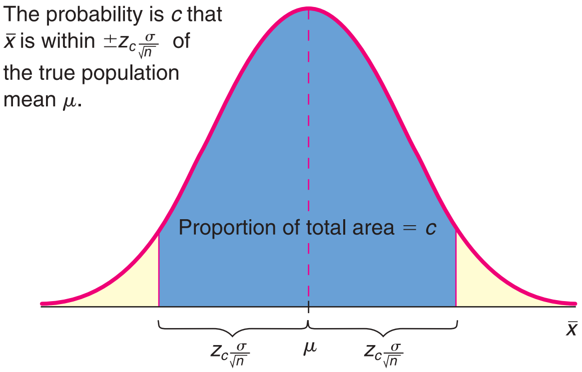
TABLE 7-2 **Some Levels of Confidence and Their Corresponding Critical Values**

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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

FIGURE 7-3

Distribution of Sample Means $\bar{x}$



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 7-3.

In the following optional discussion, we derive Equation (1). If you prefer, you may jump ahead to the discussion about the margin of error.

Optional derivation of Equation (1)

For a c confidence level, we know that

$$P(-z_c < z < z_c) = c \tag{2}$$

This statement gives us information about the size of z, but we want information about the size of $\bar{x} - \mu$. Is there a relationship between z and $\bar{x} - \mu$? The answer is yes since, by the central limit theorem, $\bar{x}$ has a distribution that is approximately normal, with mean μ and standard deviation $\sigma/\sqrt{n}$. We can convert $\bar{x}$ to a standard z score by using the formula

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{3}$$

Substituting this expression for z into Equation (2) gives

$$P\left(-z_c < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_c\right) = c \tag{4}$$

Multiplying all parts of the inequality in (4) by $\sigma/\sqrt{n}$ gives us

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

Equation (1) is precisely the equation we set out to derive.

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 that 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

$$E = z_c\frac{\sigma}{\sqrt{n}} \tag{5}$$

Note: Formula (5) 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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 (5), we conclude that

$$P(-E < \bar{x} - \mu < E) = c \tag{6}$$

Equation (6) states that the probability is c that the difference between $\bar{x}$ and μ is no more than the maximal error tolerance E. If we use a little algebra on the inequality

$$-E < \bar{x} - \mu < E \tag{7}$$

for μ, we can rewrite it in the following mathematically equivalent way:

Confidence interval for μ with σ known

$$\bar{x} - E < \mu < \bar{x} + E \tag{8}$$

Since formulas (7) and (8) are mathematically equivalent, their probabilities are the same. Therefore, from (6), (7), and (8), we obtain

$$P(\bar{x} - E < \mu < \bar{x} + E) = c \tag{9}$$

Equation (9) states that there is a chance 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. In other words, c is the proportion of confidence intervals, based on random samples of size n, that actually contain 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

Requirements

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

$$\bar{x} - E < \mu < \bar{x} + E \tag{10}$$

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

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

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

z_c = critical value for confidence level c based on the standard normal distribution (see Table 5(b) of Appendix II for frequently used values).

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 $\sigma = 1.80$ minutes. During the past year, Julia has recorded her times 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.

Myrleen Ferguson Cate/PhotoEdit

SOLUTION: *Check Requirements* We have a simple random sample of running times, and the sample size $n = 90$ is large enough for the $\bar{x}$ distribution to be approximately normal. We also know σ. It is appropriate to use the normal distribution to compute a confidence interval for μ.

To compute E for the 95% confidence interval $\bar{x} - E$ to $\bar{x} + E$, we use the fact that $z_c = 1.96$ (see Table 7-2), together with the values $n = 90$ and $\sigma = 1.80$. Therefore,

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

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

$$E \approx 0.37$$

Using Equation (10), the given value of $\bar{x}$, and our computed value for E, we get the 95% confidence interval for μ.

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

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

$$15.23 < \mu < 15.97$$

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

CRITICAL THINKING

Interpreting Confidence Intervals

A few comments are in order about the general meaning of the term *confidence interval*.

- Since $\bar{x}$ is a random variable, the endpoints $\bar{x} \pm E$ are also random variables. Equation (9) states that we have a chance c of obtaining a sample such that the interval, once it is computed, will contain the parameter μ.
- After the confidence interval is numerically fixed for a specific sample, 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.

- Equation (9), $P(\bar{x} - E < \mu < \bar{x} + E) = c$, really states that if we draw many random samples of size n and get lots of confidence intervals, then the proportion of all intervals that will turn out to contain the mean μ is c.

  For example, in Figure 7-4, on the next page, the horizontal lines represent 0.90 confidence intervals for various samples of the same size from an x distribution. Some of these intervals contain μ and others do not. Since the

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

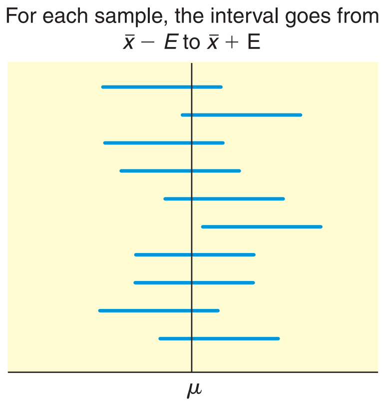
intervals are 0.90 confidence intervals, about 90% of all such intervals should contain μ. For each sample, the interval goes from $\bar{x} - E$ to $\bar{x} + E$.

- Once we have a *specific* confidence interval for μ, such as $3 < \mu < 5$, all we can say is that we are c% confident that we have one of the intervals that actually contains μ. Another appropriate statement is that at the c confidence level, our interval contains μ.

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

FIGURE 7-4

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, the mean time was $\bar{x} = 22.50$ minutes. Let μ be the mean jogging time for the entire distribution of Walter's 3-mile running times over the past several years. Find a 0.99 confidence interval for μ.

(a) *Check Requirements* Is the $\bar{x}$ distribution approximately normal? Do we know σ?

→ Yes; we know this from the central limit theorem. Yes, $\sigma = 2.40$ minutes.

(b) What is the value of $z_{0.99}$? (See Table 7-2.)

→ $z_{0.99} = 2.58$

(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) *Interpretation* Explain what the confidence interval tells us.

→ 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.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

LOOKING FORWARD

The basic structure for most confidence intervals for population parameters is

$$\text{sample statistic} - E < \text{population parameter} < \text{sample statistic} + E$$

where E is the maximal margin of error based on the sample statistic distribution and the level of confidence c. We will see this same format used for confidence intervals of the mean when σ is unknown (Section 7.2), for proportions (Section 7.3), for differences of means from independent samples (Section 7.4), for differences of proportions (Section 7.4), and for parameters of linear regression (Chapter 9). This structure for confidence intervals is so basic that some software packages, such as Excel simply give the value of E for a confidence interval and expect the user to finish the computation.

TECH NOTES

The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all support confidence intervals for μ from large samples. 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/TI-*n*spire (with TI-84Plus keypad) These calculators give 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, then select **TESTS**, and use **7:ZInterval**. The TI-84Plus/TI-83Plus/TI-*n*spire output shows the results for Example 2.

```
ZInterval
 Inpt:Data Stats
 σ:1.8
 x̄:15.6
 n:90
 C-Level:95
 Calculate
```

```
ZInterval
 (15.228, 15.972)
 x̄=15.6
 n=90
```

Excel 2007 Excel gives only the value of the maximal error of estimate E. On the **Home** screen click the **Insert Function** f_x. In the dialogue box, select **Statistical** for the category, and then select **Confidence.** In the resulting dialogue box, the value of **alpha** is 1 - confidence level. For example, alpha is 0.05 for a 95% confidence interval. The values of σ and n are also required. The Excel output shows the value of E for Example 2.

f_x =CONFIDENCE(0.05,1.8,90)

C	D	E
0.371876		

An alternate approach incorporating raw data (using the Student's t distribution presented in the next section) uses a selection from the Data Analysis package. Click the **Data** tab on the home ribbon. From the Analysis group, select **Data Analysis.** In the dialogue box, select **Descriptive Statistics.** Check the box by **Confidence Level for Mean,** and enter the confidence level. Again, the value of E for the interval is given.

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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 *maximal* 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 minimal sample size.

PROCEDURE

HOW TO FIND THE SAMPLE SIZE n FOR ESTIMATING μ WHEN σ IS KNOWN

Requirements

The distribution of sample means $\bar{x}$ is approximately normal.

Formula for sample size

$$n = \left(\frac{z_c \sigma}{E}\right)^2 \tag{11}$$

where E = specified maximal margin of error

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 5(b) of Appendix II.

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 preliminary study of a random sample of 50 salmon showed $s \approx 2.15$ pounds. How large a sample should be taken to be 99% confident that the sample mean $\bar{x}$ is within 0.20 pound of the true mean weight μ?

Photo Researchers

Salmon moving upstream

SOLUTION: In this problem, $z_{0.99} = 2.58$ (see Table 7-2) 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.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 salons 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: Wall Street Journal*). How do ASCAP and BMI know who is playing what music? The answer is, they don't! 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 United States) 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 7.1 PROBLEMS

In Problems 1–8, answer true or false. Explain your answer.

1. *Statistical Literacy* The value z_c is a value from the standard normal distribution such that $P(-z_c < x < z_c) = c$.
2. *Statistical Literacy* The point estimate for the population mean μ of an x distribution is $\bar{x}$, computed from a random sample of the x distribution.
3. *Statistical Literacy* Consider a random sample of size n from an x distribution. For such a sample, the margin of error for estimating μ is the magnitude of the difference between $\bar{x}$ and μ.
4. *Statistical Literacy* Every random sample of the same size from a given population will produce exactly the same confidence interval for μ.
5. *Statistical Literacy* A larger sample size produces a longer confidence interval for μ.
6. *Statistical Literacy* If the original x distribution has a relatively small standard deviation, the confidence interval for μ will be relatively short.
7. *Statistical Literacy* If the sample mean $\bar{x}$ of a random sample from an x distribution is relatively small, then the confidence interval for μ will be relatively short.
8. *Statistical Literacy* For the same random sample, when the confidence level c is reduced, the confidence interval for μ becomes shorter.
9. *Critical Thinking* Sam computed a 95% confidence interval for μ from a specific random sample. His confidence interval was $10.1 < \mu < 12.2$. He claims that the probability that μ is in this interval is 0.95. What is wrong with his claim?
10. *Critical Thinking* Sam computed a 90% confidence interval for μ from a specific random sample of size n. He claims that at the 90% confidence level, his confidence interval contains μ. Is his claim correct? Explain.

Answers may vary slightly due to rounding.

11. *Basic Computation: Confidence Interval* Suppose x has a normal distribution with $\sigma = 6$. A random sample of size 16 has sample mean 50.
    (a) *Check Requirements* Is it appropriate to use a normal distribution to compute a confidence interval for the population mean μ? Explain.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) Find a 90% confidence interval for μ.
(c) *Interpretation* Explain the meaning of the confidence interval you computed.

12. *Basic Computation: Confidence Interval* Suppose x has a mound-shaped distribution with $\sigma = 9$. A random sample of size 36 has sample mean 20.
(a) *Check Requirements* Is it appropriate to use a normal distribution to compute a confidence interval for the population mean μ? Explain.
(b) Find a 95% confidence interval for μ.
(c) *Interpretation* Explain the meaning of the confidence interval you computed.

13. *Basic Computation: Sample Size* Suppose x has a mound-shaped distribution with $\sigma = 3$.
(a) Find the minimal sample size required so that for a 95% confidence interval, the maximal margin of error is $E = 0.4$.
(b) *Check Requirements* Based on this sample size, can we assume that the $\bar{x}$ distribution is approximately normal? Explain.

14. *Basic Computation: Sample Size* Suppose x has a normal distribution with $\sigma = 1.2$.
(a) Find the minimal sample size required so that for a 90% confidence interval, the maximal margin of error is $E = 0.5$.
(b) *Check Requirements* Based on this sample size and the x distribution, can we assume that the $\bar{x}$ distribution is approximately normal? Explain.

15. *Zoology: Hummingbirds* Allen's hummingbird (*Selasphorus sasin*) has been studied by zoologist Bill Alther (Reference: *Hummingbirds* by 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 $\bar{x} = 3.15$ grams. Based on previous studies, we can assume that the weights of Allen's hummingbirds have a normal distribution, with $\sigma = 0.33$ gram.
(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) *Interpret* Compare your results in the context of this problem.
(d) *Sample Size* Find the sample size necessary for an 80% confidence level with a maximal margin of error $E = 0.08$ for the mean weights of the hummingbirds.

16. *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* by 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) *Interpret* Compare your results in the context of this problem.
(d) *Sample Size* Find the sample size necessary for a 95% confidence level with maximal margin of error $E = 1.10$ for the mean concentration of uric acid in this patient's blood.

17. *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 16.) Suppose that a random sample of 45 male firefighters are tested and that they have a plasma

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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) *Interpret* Compare your results in the context of this problem.

(d) *Sample Size* Find the sample size necessary for a 99% confidence level with maximal margin of error $E = 2.50$ for the mean plasma volume in male firefighters.

18. *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 margin of error $E = 0.3$ 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.

19. *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 levels increase, do the margins of error increase?

(e) *Critical Thinking:* Compare the lengths of the confidence intervals for parts (a) through (c). As the confidence levels increase, do the confidence intervals increase in length?

20. *Confidence Intetvals: Values of σ* A random sample of size 36 is drawn from an x distribution. The sample mean is 100.

(a) Suppose the x distribution has $\sigma = 30$. Compute a 90% confidence interval for μ. What is the value of the margin of error?

(b) Suppose the x distribution has $\sigma = 20$. Compute a 90% confidence interval for μ. What is the value of the margin of error?

(c) Suppose the x distribution has $\sigma = 10$. Compute a 90% confidence interval for μ. What is the value of 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) *Critical Thinking* 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?

21. *Confidence Intervals: Sample Size* A random sample is drawn from a population with $\sigma = 12$. The sample mean is 30.

(a) Compute a 95% confidence interval for μ based on a sample of size 49. What is the value of the margin of error?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) Compute a 95% confidence interval for μ based on a sample of size 100. What is the value of the margin of error?
(c) Compute a 95% confidence interval for μ based on a sample of size 225. What is the value of 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) *Critical Thinking* 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?

22. ***Ecology: Sand Dunes*** At wind speeds above 1000 centimeters per second (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) ***Interpretation*** Does the confidence interval indicate that the population mean wind speed is such that the sand is always moving at this site? Explain.

23. ***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) ***Interpretation*** 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 this might be somewhat low compared with other successful financial institutions? Explain by referring to the confidence interval you computed in part (b).
(d) ***Interpretation*** 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.

24. ***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 23.) Companies such as Gap, Nordstrom, Dillards, JCPenney, Sears, Wal-Mart, Office Depot, and Toys "Я" Us are included. Assume $\sigma \approx 3.8$ thousand dollars.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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) *Interpretation* 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 this might be low compared with other retail stores? Explain by referring to the confidence interval you computed in part (b).
(d) *Interpretation* 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.

25. *Ballooning: Air Temperature* How hot is the air in the top (crown) of a hot air balloon? Information from *Ballooning: The Complete Guide to Riding the Winds* by Wirth and Young (Random House) claims that the air in the crown should be an average of 100°C for a balloon to be in a state of equilibrium. However, the temperature does not need to be exactly 100°C. What is a reasonable and safe range of temperatures? This range may vary with the size and (decorative) shape of the balloon. All balloons have a temperature gauge in the crown. Suppose that 56 readings (for a balloon in equilibrium) gave a mean temperature of $\bar{x} = 97$°C. For this balloon, $\sigma \approx 17$°C.
(a) Compute a 95% confidence interval for the average temperature at which this balloon will be in a steady-state equilibrium.
(b) *Interpretation* If the average temperature in the crown of the balloon goes above the high end of your confidence interval, do you expect that the balloon will go up or down? Explain.


Images Etc Ltd /Getty images

SECTION 7.2

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 levels.
- Compute confidence intervals for μ when σ is unknown. What does this information tell you?

Student's *t* distribution

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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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.

Degrees of freedom, *d.f.*

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**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad (12)$$

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 (12). 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 say 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.

Table 6 of Appendix II 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

$$d.f. = n - 1 \qquad (13)$$

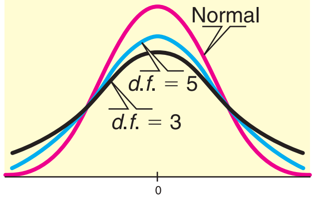
where *d.f.* stands for the degrees of freedom and n is the sample size. Each choice for *d.f.* gives a different t distribution.

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 7-5 shows a standard normal z distribution and Student's t distribution with $d.f. = 3$ and $d.f. = 5$.

FIGURE 7-5

A Standard Normal Distribution and 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.
5. The area under the entire curve is 1.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Using Table 6 to Find Critical Values for Confidence Intervals

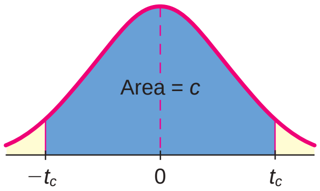
Critical values t_c

Table 6 of Appendix II 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

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

FIGURE 7-6

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



This probability corresponds to the shaded area in Figure 7-6.

Table 6 of Appendix II 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, just 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 7-3 (an excerpt from Table 6 of Appendix II) 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 7-3 Student's *t* Distribution Critical Values (Excerpt from Table 6, Appendix II)

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 6 of Appendix II (or Table 7-3, showing an excerpt from the table) to find t_c for a 0.90 confidence level for a t distribution with sample size $n = 9$.

Continued

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

GUIDED EXERCISE 2 *continued*

(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 $n = 9$.

LOOKING FORWARD

Student's t distributions will be used again in Chapter 8 when testing μ and when testing differences of means. The distributions are also used for confidence intervals and testing of parameters of linear regression (Sections 9.3 and 9.4).

Confidence Intervals for μ When σ Is Unknown

Maximal margin of error, E

In Section 7.1, 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

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

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 7.1 is

$$P\left(-t_c \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_c \frac{s}{\sqrt{n}}\right) = c \qquad (14)$$

COMMENT Comparing Equation (14) with Equation (1) in Section 7.1, it becomes evident that we are using the same basic method on the t distribution that we used on the z distribution.

Likewise, for samples from normal populations with unknown σ, Equation (9) of Section 7.1 becomes

$$P(\bar{x} - E < \mu < \bar{x} + E) = c \qquad (15)$$

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

PROCEDURE

Confidence interval for μ with σ unknown

HOW TO FIND A CONFIDENCE INTERVAL FOR μ WHEN σ IS UNKNOWN

Requirements

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 $\bar{x}$ and the sample standard deviation s.

Continued

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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

$$\bar{x} - E < \mu < \bar{x} + E \tag{16}$$

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 6 of Appendix II.)

Problem 23 presents an alternate method for computing confidence intervals when σ is unknown but $n \geq 30$. The alternate method approximates σ by the sample standard deviation s and utilizes the normal distribution.

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 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: *Check Requirements* We assume that the shoulder heights of the reconstructed skeletons form a random sample of shoulder heights for all the miniature horses of the unknown species. The x distribution is assumed to be approximately normal. Since σ is unknown, it is appropriate to use a Student's t distribution and sample information to compute a confidence interval for μ.

In this case, $n = 7$, so $d.f. = n - 1 = 7 - 1 = 6$. For $c = 0.990$, Table 6 of Appendix II 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

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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Interpretation 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 distribution of weights is mound-shaped and symmetric. 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) *Check Requirements* Is it appropriate to use a Student's t distribution to compute a confidence interval for μ?

→ Yes, we assume that the 37 sapphires constitute a simple random sample of all sapphires produced under the new process. The requirement that the x distribution be approximately normal can be dropped since the sample size is large enough to ensure that the $\bar{x}$ distribution is approximately normal.

(b) What is *d.f.* for this setting?

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

(c) Use Table 6 of Appendix II 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 $t_{0.95} = 2.030$.

(d) Find E.

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

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

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

(f) *Interpretation* What does the confidence interval tell us 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 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.

Continued

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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

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

where degrees of freedom

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

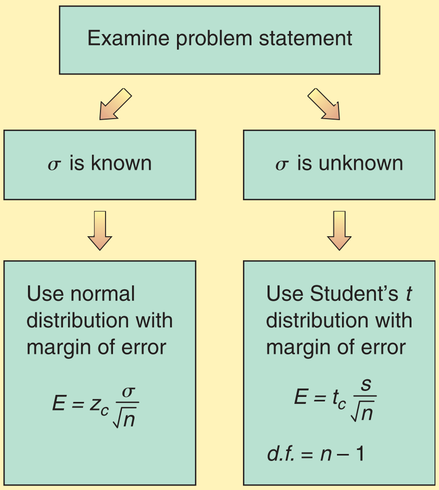
Although a t distribution can be used in many situations, you need to observe some guidelines. If n is less than 30, x should have a distribution that is mound-shaped and approximately symmetric. It's even better if the x distribution is normal. If n is 30 or more, the central limit theorem (Chapter 6) 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 $\bar{x}$?



COMMENT To find confidence intervals for μ based on small samples, we 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. If you read Using Technology at the end of this chapter, you will find one (of many) bootstrap methods (Reference: *An Introduction to the Bootstrap* by B. Efron and R. Tibshirani).

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

TECH NOTES The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab support confidence intervals using the Student's t distribution.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) 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 2007 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 $\bar{x} + E$. On the Home screen, click the **Data** tab. In the Analysis group, click **Data Analysis**. In the dialogue box, select **Descriptive Statistics**. In the dialogue box, check summary statistics and check confidence level for mean. Then set the desired confidence level. Under these 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! Earthquakes often 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://www.cengage.com/statistics/brase** *and find the link to National Earthquake Information Service.*

SECTION 7.2 PROBLEMS

1. Use Table 6 of Appendix II to find t_c for a 0.95 confidence level when the sample size is 18.
2. Use Table 6 of Appendix II to find t_c for a 0.99 confidence level when the sample size is 4.
3. Use Table 6 of Appendix II to find t_c for a 0.90 confidence level when the sample size is 22.
4. Use Table 6 of Appendix II to find t_c for a 0.95 confidence level when the sample size is 12.
5. *Statistical Literacy* Student's t distributions are symmetric about a value of t. What is that t value?
6. *Statistical Literacy* As the degrees of freedom increase, what distribution does the Student's t distribution become more like?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. *Critical Thinking* Consider a 90% confidence interval for μ. Assume σ is not known. For which sample size, $n = 10$ or $n = 20$, is the critical value t_c larger?

8. *Critical Thinking* Consider a 90% confidence interval for μ. Assume σ is not known. For which sample size, $n = 10$ or $n = 20$, is the confidence interval longer?

9. *Critical Thinking* Lorraine computed a confidence interval for μ based on a sample of size 41. Since she did not know σ, she used s in her calculations. Lorraine used the normal distribution for the confidence interval instead of a Student's t distribution. Was her interval longer or shorter than one obtained by using an appropriate Student's t distribution? Explain.

10. *Critical Thinking* Lorraine was in a hurry when she computed a confidence interval for μ. Because σ was not known, she used a Student's t distribution. However, she accidentally used degrees of freedom n instead of $n - 1$. Was her confidence interval longer or shorter than one found using the correct degrees of freedom $n - 1$? Explain.

Answers may vary slightly due to rounding.

11. *Basic Computation: Confidence Interval* Suppose x has a mound-shaped distribution. A random sample of size 16 has sample mean 10 and sample standard deviation 2.
    (a) *Check Requirements* Is it appropriate to use a Student's t distribution to compute a confidence interval for the population mean μ? Explain.
    (b) Find a 90% confidence interval for μ.
    (c) *Interpretation* Explain the meaning of the confidence interval you computed.

12. *Basic Computation: Confidence Interval* A random sample of size 81 has sample mean 20 and sample standard deviation 3.
    (a) *Check Requirements* Is it appropriate to use a Student's t distribution to compute a confidence interval for the population mean μ? Explain.
    (b) Find a 95% confidence interval for μ.
    (c) *Interpretation* Explain the meaning of the confidence interval you computed.

In Problems 13–19, assume that the population of x values has an approximately normal distribution.

13. *Archaeology: Tree Rings* At Burnt Mesa Pueblo, the method of tree-ring dating gave the following years 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 year 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.
    (c) *Interpretation* What does the confidence interval mean in the context of this problem?

14. *Camping: Cost of a Sleeping Bag* How much does a sleeping bag cost? Let's say you want a sleeping bag that should keep you warm in temperatures from 20°F to 45°F. A random sample of prices ($) for sleeping bags in this temperature range was taken from *Backpacker Magazine: Gear Guide* (Vol. 25, Issue 157, No. 2). Brand names include American Camper, Cabela's, Camp 7, Caribou, Cascade, and Coleman.

80	90	100	120	75	37	30	23	100	110
105	95	105	60	110	120	95	90	60	70

    (a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx \$83.75$ and $s \approx \$28.97$.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) Using the given data as representative of the population of prices of all summer sleeping bags, find a 90% confidence interval for the mean price μ of all summer sleeping bags.
(c) *Interpretation* What does the confidence interval mean in the context of this problem?

15. *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 (pounds):

68	104	128	122	60	64

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

16. *Franchise: Candy Store* Do you want to own your own candy store? 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 (in thousands of dollars) 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

(a) 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.
(b) Find a 90% confidence interval for the population average startup costs μ for candy store franchises.
(c) *Interpretation* What does the confidence interval mean in the context of this problem?

17. *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* by 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) *Interpretation* Based on your results in part (b), does it seem that this patient still has a calcium deficiency? Explain.

18. *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

(a) Use a calculator with mean and sample standard deviation keys to verify that $\bar{x} \approx 62.3\%$ and $s \approx 8.0\%$.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) Find a 90% confidence interval for the population average μ of the percentage of hospitals providing at least some charity care.

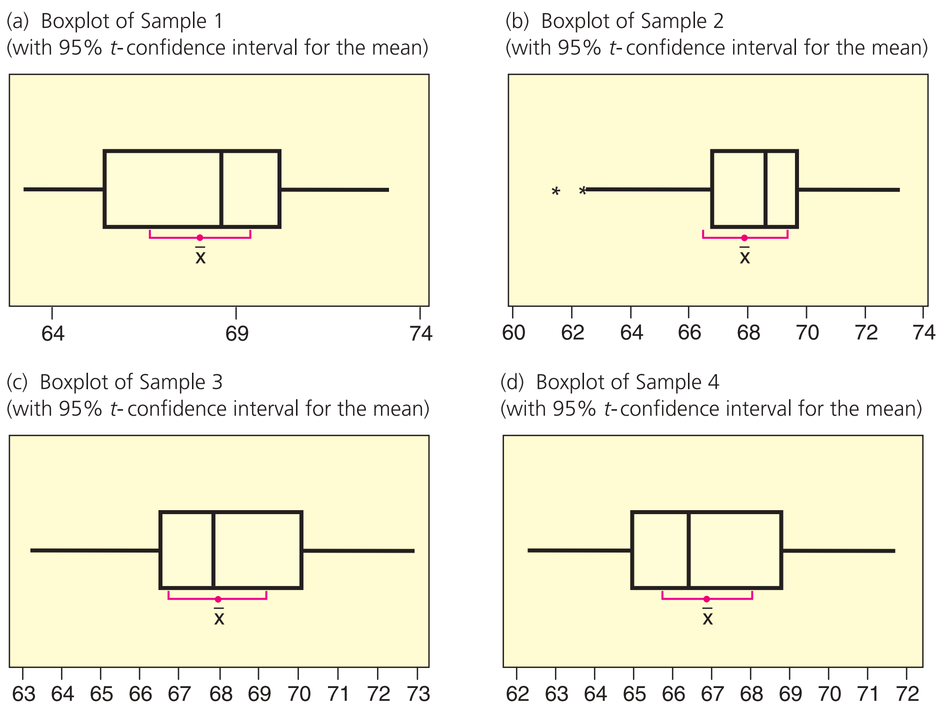
(c) *Interpretation* What does the confidence interval mean in the context of this problem?

19. *Critical Thinking: Boxplots and Confidence Intervals* 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)



Carlos Serrao/Stone/Getty Images

(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?

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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 rates in Denver neighborhoods. Compute an 80% confidence interval for μ, the population mean crime rate for all Denver neighborhoods.

(c) *Interpretation* 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) *Interpretation* 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) *Check Requirement* In previous problems, we assumed 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 6.5.

21. *Finance: P/E Ratio* The price of a share of stock divided by a 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) *Interpretation* 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 the time the sample was taken?

(e) *Check Requirements* In previous problems, we assumed 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 6.5.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

22. ***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 6 of Appendix II, 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) ***Interpretation*** 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 the home run percentages for these players compare to the population average.

(e) ***Check Requirements*** In previous problems, we assumed 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 6.5.

23. ***Expand Your Knowledge: Alternate Method for Confidence Intervals*** When σ is unknown and the sample is of size $n \geq 30$, 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?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

SECTION 7.3

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.

Basic criteria

The binomial distribution is completely determined by the number of trials n and the probability p of success on 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 *both*

Basic criteria

$$np > 5 \quad \text{and} \quad nq > 5, \quad \text{where } q = 1 - p$$

Point estimate for p

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.

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 a shot 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

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

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,

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$P\left(-z_c\sqrt{\frac{pq}{n}} < \hat{p} - p < z_c\sqrt{\frac{pq}{n}}\right) = c \tag{17}$$

Equation (17) says 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 7.1, we call E the *maximal margin of error*.

Optional derivation of Equation (17)

First, we need to show that $\hat{p} = r/n$ has a distribution that is approximately normal, with $\mu = p$ and $\sigma = \sqrt{pq/n}$. From Section 6.6, we know that for sufficiently large n, the binomial distribution can be approximated by a normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. If r is the number of successes out of n trials of a binomial experiment, then r is a binomial random variable with a binomial distribution. When we convert r to standard z units, we obtain

$$z = \frac{r - \mu}{\sigma} = \frac{r - np}{\sqrt{npq}}$$

For sufficiently large n, r will be approximately normally distributed, so z will be too.

If we divide both numerator and denominator of the last expression by n, the value of z will not change.

$$z = \frac{\dfrac{r - np}{n}}{\dfrac{\sqrt{npq}}{n}} \quad \text{Simplified, we find } z = \frac{\dfrac{r}{n} - p}{\sqrt{\dfrac{pq}{n}}}. \tag{18}$$

Equation (18) tells us that the $\hat{p} = r/n$ distribution is approximated by a normal curve with $\mu = p$ and $\sigma = \sqrt{pq/n}$.

The probability is c that z lies in the interval between $-z_c$ and z_c because an area equal to c under the standard normal curve lies between $-z_c$ and z_c. Using the language of probability, we write

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

From Equation (18), we know that

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$$

If we put this expression for z into the preceding equation, we obtain

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

If we multiply all parts of the inequality by $\sqrt{pq/n}$, we obtain the equivalent statement

$$P\left(-z_c\sqrt{\frac{pq}{n}} < \hat{p} - p < z_c\sqrt{\frac{pq}{n}}\right) = c \tag{17}$$

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 (17). Then we get

$$P(-E < \hat{p} - p < E) = c \tag{19}$$

Some algebraic manipulation produces the mathematically equivalent statement

$$P(\hat{p} - E < p < \hat{p} + E) = c \tag{20}$$

Equation (20) 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 reliable 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

Requirements

Consider a binomial experiment with n trials, where p represents the population probability of success on a single trial 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 n should be sufficiently large so that both $n\hat{p} > 5$ 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 5(b) of Appendix II for frequently used values.)

Problem 28 of this section shows the method for computing a "plus four confidence interval for p." This is an alternate method that generally results in a slightly smaller confidence interval than the ones computed with standard methods presented in this section and used in most statistical technology packages.

COMMENT Problem 6 asks you to show that the two conditions $n\hat{p} > 5$ and $n\hat{q} > 5$ are equivalent to the two conditions that the number of successes $r > 5$ and the number of failures $n - r > 5$.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 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.

Stock Boston

(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) *Check Requirements* 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 $q \approx 0.25$, then

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

A normal approximation is certainly justified.

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

SOLUTION:

$$z_{0.99} = 2.58 \text{ (see Table 7-2 or Table 5(b) of Appendix II)}$$

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

The 99% confidence interval is then

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

Interpretation We are 99% confident that the probability a flu shot will be effective for a student selected at random is between 0.71 and 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? → $\hat{p} = \dfrac{r}{n} = \dfrac{66}{188} = 0.35$

Continued

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

GUIDED EXERCISE 4 *continued*

(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) *Interpretation* What does the confidence interval you just computed mean in the context of this application?

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 contain the unknown value p.

(d) *Check Requirements* 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 $np \approx 122.2 > 5$, 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. At the end of this section, we will study the effect of sample size on the reliability of our estimate.

TECH NOTES The TI-84Plus/TI-83Plus/TI-*n*spire calculators and Minitab provide confidence intervals for proportions.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) 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/TI-*n*spire 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 from Chapter 8 material is also shown.

Test and Confidence Interval for One Proportion (Using Binomial)

Test of p = 0.5 vs p not = 0.5

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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

```
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

Margin of error for polls

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 United States. 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 United States 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 excerpt "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.

Continued

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

GUIDED EXERCISE 5 *continued*

(b) The complete 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, "Democrats"?

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 responded, "Democrats" 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

$$n = p(1-p)\left(\frac{z_c}{E}\right)^2 \text{ if you have a preliminary estimate for } p \tag{21}$$

$$n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 \text{ if you do } \textit{not} \text{ have a preliminary estimate for } p \tag{22}$$

where E = specified maximal error of estimate

z_c = critical value from the normal distribution for the desired confidence level c. Commonly used value of z_c can be found in Table 5(b) of Appendix II.

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 (21), 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 $p(1-q)$ is 1/4. See Problem 27.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 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 (22) with $z_{0.95} = 1.96$ (see Table 7-2) 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 he use?

SOLUTION: In this case, we use Equation (21) with $p \approx 0.86$. Again, from Table 7-2, $z_{0.95} = 1.96$, 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 $n = 4626$. 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 through 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 7.3 PROBLEMS

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

1. *Statistical Literacy* For a binomial experiment with r successes out of n trials, what value do we use as a point estimate for the probability of success p on a single trial?

2. *Statistical Literacy* In order to use a normal distribution to compute confidence intervals for p, what conditions on np and nq need to be satisfied?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. *Critical Thinking* Results of a poll of a random sample of 3003 American adults showed that 20% did not know that caffeine contributes to dehydration. The poll was conducted for the Nutrition Information Center and had a margin of error of $\pm 1.4\%$.
   (a) Does the margin of error take into account any problems with the wording of the survey question, interviewer errors, bias from sequence of questions, and so forth?
   (b) What does the margin of error reflect?

4. *Critical Thinking* You want to conduct a survey to determine the proportion of people who favor a proposed tax policy. How does increasing the sample size affect the size of the margin of error?

5. *Critical Thinking* Jerry tested 30 laptop computers owned by classmates enrolled in a large computer science class and discovered that 22 were infected with keystroke-tracking spyware. Is it appropriate for Jerry to use his data to estimate the proportion of all laptops infected with such spyware? Explain.

6. *Critical Thinking: Brain Teaser* A requirement for using the normal distribution to approximate the $\hat{p}$ distribution is that both $np > 5$ and $nq > 5$. Since we usually do not know p, we estimate p by $\hat{p}$ and q by $\hat{q} = 1 - \hat{p}$. Then we require that $n\hat{p} > 5$ and $n\hat{q} > 5$. Show that the conditions $n\hat{p} > 5$ and $n\hat{q} > 5$ are equivalent to the condition that out of n binomial trials, both the number of successes r and the number of failures $n - r$ must exceed 5. *Hint:* In the inequality $n\hat{p} > 5$, replace $\hat{p}$ by r/n and solve for r. In the inequality $n\hat{q} > 5$, replace $\hat{q}$ by $(n - r)/n$ and solve for $n - r$.

7. *Basic Computation: Confidence Interval for p* Consider $n = 100$ binomial trials with $r = 30$ successes.
   (a) *Check Requirements* Is it appropriate to use a normal distribution to approximate the $\hat{p}$ distribution?
   (b) Find a 90% confidence interval for the population proportion of successes p.
   (c) *Interpretation* Explain the meaning of the confidence interval you computed.

8. *Basic Computation: Confidence Interval for p* Consider $n = 200$ binomial trials with $r = 80$ successes.
   (a) *Check Requirements* Is it appropriate to use a normal distribution to approximate the $\hat{p}$ distribution?
   (b) Find a 95% confidence interval for the population proportion of successes p.
   (c) *Interpretation* Explain the meaning of the confidence interval you computed.

9. *Basic Computation: Sample Size* What is the minimal sample size needed for a 95% confidence interval to have a maximal margin of error of 0.1
   (a) if a preliminary estimate for p is 0.25?
   (b) if there is no preliminary estimate for p?

10. *Basic Computation: Sample Size* What is the minimal sample size needed for a 99% confidence interval to have a maximal margin of error of 0.06
    (a) if a preliminary estimate for p is 0.8?
    (b) if there is no preliminary estimate for p?

11. *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) *Check Requirements* Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Explain why this would be an important consideration.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12. ***Myers–Briggs: Judges*** In a random sample of 519 judges, it was found that 285 were introverts (see reference of Problem 11).
    (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) ***Check Requirements*** Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Explain why this would be an important consideration.

13. ***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) ***Check Requirements*** Do you think that $np > 5$ and $nq > 5$ are satisfied for this problem? Explain why this would be an important consideration.

14. ***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) ***Check Requirements*** Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this problem? Why would this be important?

15. ***Health Care: Colorado Physicians*** A random sample of 5792 physicians in Colorado showed that 3139 provide at least some charity care (i.e., treat 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) ***Check Requirements*** Is the normal approximation to the binomial justified in this problem? Explain.

16. ***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) ***Check Requirements*** Is use of the normal approximation to the binomial justified in this problem? Explain.

17. ***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).

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(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) *Check Requirements* Is the normal approximation to the binomial justified in this problem? Explain.

18. *Physicians: Solo Practice* A random sample of 328 medical doctors showed that 171 have a solo practice (Source: *Practice Patterns of General Internal Medicine*, American Medical Association).
(a) Let p represent the proportion of all medical doctors who have a solo practice. 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) *Interpretation* As a news writer, how would you report the survey results regarding the percentage of medical doctors in solo practice? What is the margin of error based on a 95% confidence interval?

19. *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) *Interpretation* As a news writer, how would you report the survey results regarding the percentage of women 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?

20. *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 19.
(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) *Interpretation* As a news writer, how would you report the survey results on the percentage of supermarket shoppers who stock up on real-bargain items? What is the margin of error based on a 95% confidence interval?

21. *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) *Interpretation* As a news writer, how would you report the survey results regarding the proportion of corporations that hire the equally qualified nonsmoker? What is the margin of error based on a 95% confidence interval?

22. *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

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a 95% confidence interval for the percentage of the population that would respond, "Crime and violence" to the question asked by the pollsters.

23. ***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* by F. Fischbach).

24. ***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).

25. ***Campus Life: Coeds*** What percentage of your 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 female. Answer part (a) using this estimate for p.

26. ***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?

27. ***Brain Teaser: Algebra*** Why do we use 1/4 in place of $p(1 - p)$ in formula (22) for sample size when the probability of success p is unknown?
(a) Show that $p(1 - p) = 1/4 - (p - 1/2)^2$.
(b) Why is $p(1 - p)$ never greater than 1/4?

28. ***Expand Your Knowledge: Plus Four Confidence Interval for a Single Proportion*** One of the technical difficulties that arises in the computation of confidence intervals for a single proportion is that the exact formula for the maximal margin of error requires knowledge of the population proportion of success p. Since p is usually not known, we use the sample estimate $\hat{p} = r/n$ in place of p. As discussed in the article "How Much Confidence Should You Have in Binomial Confidence Intervals?" appearing in issue no. 45 of the magazine *STATS* (a publication of the American Statistical Association), use of $\hat{p}$ as an estimate for p means that the actual confidence level for the intervals may in fact be smaller than the specified level c. This problem arises even when n is large, especially if p is not near 1/2.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A simple adjustment to the formula for the confidence intervals is the *plus four estimate*, first suggested by Edwin Bidwell Wilson in 1927. It is also called the Agresti–Coull confidence interval. This adjustment works best for 95% confidence intervals.

The plus four adjustment has us add two successes and two failures to the sample data. This means that r, the number of successes, is increased by 2, and n, the sample size, is increased by 4. We use the symbol $\tilde{p}$, read "p tilde," for the resulting sample estimate of p. So, $\tilde{p} = (r + 2)/(n + 4)$.

PROCEDURE

HOW TO COMPUTE A PLUS FOUR CONFIDENCE INTERVAL FOR p

Requirements

Consider a binomial experiment with n trials, where 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 plus four point estimates for p and q are

$$\tilde{p} = \frac{r + 2}{n + 4} \quad \text{and} \quad \tilde{q} = 1 - \tilde{p}$$

The number of trials n should be at least 10.

Approximate confidence interval for p

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

where $E \approx z_c\sqrt{\dfrac{\tilde{p}\tilde{q}}{n + 4}} = z_c\sqrt{\dfrac{\tilde{p}(1 - \tilde{p})}{n + 4}}$

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

z_c = critical value for confidence level c based on the standard normal distribution

(a) Consider a random sample of 50 trials with 20 successes. Compute a 95% confidence interval for p using the plus four method.

(b) Compute a traditional 95% confidence interval for p using a random sample of 50 trials with 20 successes.

(c) Compare the lengths of the intervals obtained using the two methods. Is the point estimate closer to 1/2 when using the plus four method? Is the margin of error smaller when using the plus four method?

SECTION 7.4

Estimating $\mu_1 - \mu_2$ and $p_1 - p_2$

FOCUS POINTS

- Distinguish between independent and dependent samples.
- Compute confidence intervals for $\mu_1 - \mu_2$ when σ_1 and σ_2 are known.
- Compute confidence intervals for $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown.
- Compute confidence intervals for $p_1 - p_2$ using the normal approximation.
- Interpret the meaning and implications of an all-positive, all-negative, or mixed confidence interval.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Independent Samples and Dependent Samples

How can we tell if two populations are different? One way is to compare the difference in population means or the difference in population proportions. In this section, we will use samples from two populations to create confidence intervals for the difference between population parameters.

To make a statistical estimate about the difference between two population parameters, we need to have a sample from each population. Samples may be *independent* or *dependent* according to how they are selected.

Independent samples

Two samples are **independent** if sample data drawn from one population are completely unrelated to the selection of sample data from the other population.

Dependent samples

Two samples are **dependent** if each data value in one sample can be paired with a corresponding data value in the other sample.

Dependent samples and data pairs occur very naturally in "before and after" situations in which the *same object* or item is *measured twice*. We will devote an entire section (8.4) to the study of dependent samples and paired data. However, in this section, we will confine our interest to independent samples.

Independent samples occur very naturally when we draw *two random samples*, one from the first population and one from the second population. Because *both* samples are random samples, there is no pairing of measurements between the two populations. All the examples of this section will involve independent random samples.

GUIDED EXERCISE 6 Distinguishing between independent and dependent samples

For each experiment, categorize the sampling as independent or dependent, and explain your choice.

(a) In many medical experiments, a sample of subjects is randomly divided into two groups. One group is given a specific treatment, and the other group is given a placebo. After a certain period of time, both groups are measured for the same condition. Do the measurements from these two groups constitute independent or dependent samples?

→ Since the subjects are *randomly assigned* to the two treatment groups (one receives a treatment, the other a placebo), the resulting measurements form independent samples.

(b) In an accountability study, a group of students in an English composition course is given a pretest. After the course, the same students are given a posttest covering similar material. Are the two groups of scores independent or dependent?

→ Since the pretest scores and the posttest scores are from the same students, the samples are dependent. Each student has both a pretest score and a posttest score, so there is a natural pairing of data values.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Confidence Intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 known)

The $\bar{x}_1 - \bar{x}_2$ sampling distribution

In this section, we will use probability distributions that arise from a difference of means (or proportions). How do we obtain such distributions? Suppose we have two statistical variables x_1 and x_2, each with its own distribution. We take *independent* random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution. Then we compute the respective means $\bar{x}_1$ and $\bar{x}_2$. Now consider the difference $\bar{x}_1 - \bar{x}_2$. This expression represents a difference of means. If we repeat this sampling process over and over, we will create lots of $\bar{x}_1 - \bar{x}_2$ values. Figure 7-7 illustrates the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

The values of $\bar{x}_1 - \bar{x}_2$ that come from repeated (independent) sampling of populations 1 and 2 can be arranged in a relative-frequency table and a relative-frequency histogram (see Section 2.1). This would give us an experimental idea of the theoretical probability distribution of $\bar{x}_1 - \bar{x}_2$.

Fortunately, it is not necessary to carry out this lengthy process for each example. The results have been worked out mathematically. The next theorem presents the main results.

THEOREM 7.1 Let x_1 and x_2 have normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively. If we take independent random samples of size n_1 from the x_1 distribution and of size n_2 from the x_2 distribution, then the variable $\bar{x}_1 - \bar{x}_2$ has

1. a normal distribution
2. mean $\mu_1 - \mu_2$
3. standard deviation $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

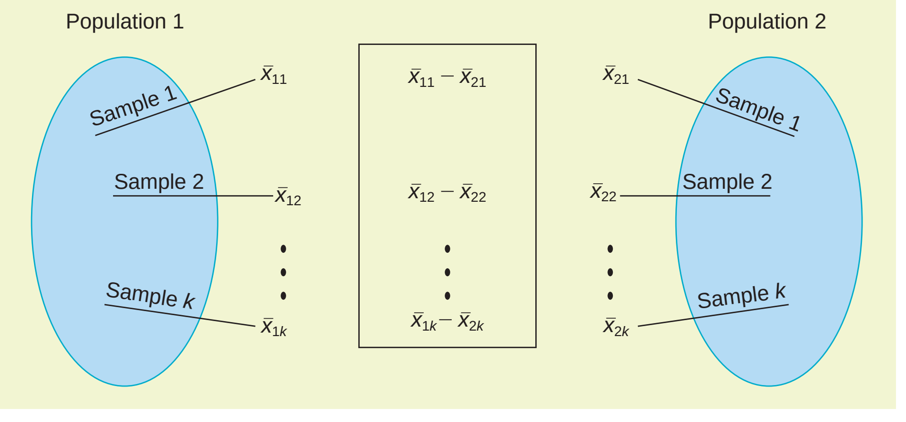
COMMENT The theorem requires that x_1 and x_2 have *normal* distributions. However, if *both* n_1 and n_2 are 30 or larger, then the central limit theorem (Section 6.5) assures us that $\bar{x}_1$ and $\bar{x}_2$ are approximately normally distributed. In this case, the conclusions of the theorem are again valid even if the original x_1 and x_2 distributions are not exactly normal.

Confidence intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 known)

If we use Theorem 7.1, then a discussion similar to that of Section 7.1 gives the following information.

FIGURE 7-7

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$



Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

PROCEDURE

How to find a confidence interval for $\mu_1 - \mu_2$ when both σ_1 and σ_2 are known

Requirements

Let σ_1 and σ_2 be the population standard deviations of populations 1 and 2. Obtain two *independent* random samples from populations 1 and 2, where

$\bar{x}_1$ and $\bar{x}_2$ are sample means from populations 1 and 2

n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal, any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Confidence interval for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where $E = z_c\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

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

z_c = critical value for confidence level c based on the standard normal distribution (See Table 5(b) of Appendix II for commonly used values.)

Problem 28 of this section shows how to determine minimal sample sizes for a specified maximal error of estimate E.

EXAMPLE 8

Confidence interval for $\mu_1 - \mu_2$, σ_1 and σ_2 known

In the summer of 1988, Yellowstone National Park had some major fires that destroyed large tracts of old timber near many famous trout streams. Fishermen were concerned about the long-term effects of the fires on these streams. However, biologists claimed that the new meadows that would spring up under dead trees would produce a lot more insects, which would in turn mean better fishing in the years ahead. Guide services registered with the park provided data about the daily catch for fishermen over many years. Ranger checks on the streams also provided data about the daily number of fish caught by fishermen.

Yellowstone Today (a national park publication) indicates that the biologists' claim is basically correct and that Yellowstone anglers are delighted by their average increased catch.

Suppose you are a biologist studying fishing data from Yellowstone streams before and after the fire. Fishing reports include the number of trout caught per day per fisherman. A random sample of $n_1 = 167$ reports from the period before the fire showed that the average catch was $\bar{x}_1 = 5.2$ trout per day. Assume that the standard deviation of daily catch per fisherman during this period was $\sigma_1 = 1.9$. Another random sample of $n_2 = 125$ fishing reports 5 years after the fire showed that the average catch per day was $\bar{x}_2 = 6.8$ trout. Assume that the standard deviation during this period was $\sigma_2 = 2.3$.

Mark R,2010/Used under license from Shutterstock.Com

Yellowstone National Park

(a) *Check Requirements* For each sample, what is the population? Are the samples dependent or independent? Explain. Is it approriate to use a normal distribution for the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.

SOLUTION: The population for the first sample is the number of trout caught per day by fishermen before the fire. The population for the second sample is the number of trout caught per day after the fire. Both samples were random

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

samples taken in their respective time periods. There was no effort to pair individual data values. Therefore, the samples can be thought of as independent samples.

A normal distribution is appropriate for the $\bar{x}_1 - \bar{x}_2$ distribution because sample sizes are sufficiently large and we know both σ_1 and σ_2.

(b) Compute a 95% confidence interval for $\mu_1 - \mu_2$, the difference of population means.

SOLUTION: Since $n_1 = 167$, $\bar{x}_1 = 5.2$, $\sigma_1 = 1.9$, $n_2 = 125$, $\bar{x}_2 = 6.8$, $\sigma_2 = 2.3$, and $z_{0.95} = 1.96$ (see Table 7-2), then

$$E = z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= 1.96\sqrt{\frac{(1.9)^2}{167} + \frac{(2.3)^2}{125}} \approx 1.96\sqrt{0.0639} \approx 0.4955 \approx 0.50$$

The 95% confidence interval is

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

$$(5.2 - 6.8) - 0.50 < \mu_1 - \mu_2 < (5.2 - 6.8) + 0.50$$

$$-2.10 < \mu_1 - \mu_2 < -1.10$$

(c) *Interpretation* What is the meaning of the confidence interval computed in part (b)?

SOLUTION: We are 95% confident that the interval -2.10 to -1.10 fish per day is one of the intervals containing the population difference $\mu_1 - \mu_2$, where μ_1 represents the population average daily catch before the fire and μ_2 represents the population average daily catch after the fire. Put another way, since the confidence interval contains only *negative values*, we can be 95% sure that $\mu_1 - \mu_2 < 0$. This means we are 95% sure that $\mu_1 < \mu_2$. In words, we are 95% sure that the average catch before the fire was less than the average catch after the fire.

COMMENT In the case of large samples ($n_1 \geq 30$ and $n_2 \geq 30$), it is not unusual to see σ_1 and σ_2 approximated by s_1 and s_2. Then Theorem 7.1 is used as a basis for approximating confidence intervals for $\mu_1 - \mu_2$. In other words, when samples are large, sample estimates for σ_1 and σ_2 can be used together with the standard normal distribution to find confidence intervals for $\mu_1 - \mu_2$. However, in this text, we follow the more common convention of using a Student's t distribution whenever σ_1 and σ_2 are unknown.

Confidence Intervals for $\mu_1 - \mu_2$ When σ_1 and σ_2 Are Unknown

When σ_1 and σ_2 are unknown, we turn to a Student's t distribution. As before, when we use a Student's t distribution, we require that our populations be normal or approximately normal (mound-shaped and symmetric) when the sample sizes n_1 and n_2 are less than 30. We also replace σ_1 by s_1 and σ_2 by s_2. Then we consider the approximate t value attributed to Welch (*Biometrika*, Vol. 29, pp. 350–362).

$$t \approx \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

Unfortunately, this approximation is *not exactly* a Student's t distribution. However, it will be a good approximation provided we adjust the degrees of freedom by one of the following methods.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problem 30 of this section shows the formula for Satterthwaite's approximation for degrees of freedom.

1. The adjustment for the degrees of freedom is calculated from sample data. The formula, called *Satterthwaite's approximation*, is rather complicated. Satterthwaite's approximation is used in statistical software packages such as Minitab and in the TI-84Plus/TI-83Plus/TI-*n*spire calculators. See Problem 30 for the formula.
2. An alternative method, which is much simpler, is to approximate the degrees of freedom using the *smaller* of $n_1 - 1$ and $n_2 - 1$.

For confidence intervals, we take the degrees of freedom $d.f.$ to be the smaller of $n_1 - 1$ and $n_2 - 1$. This commonly used choice for the degrees of freedom is more conservative than Satterthwaite's approximation in the sense that the former produces a slightly larger margin of error. The resulting confidence interval will be *at least* at the c level, or a little higher.

Confidence intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 unknown)

Applying methods similar to those used to find confidence intervals for μ when σ is unknown, and using the Welch approximation for t, we obtain the following results.

PROCEDURE

How to find a confidence interval for $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown

Requirements

Obtain two independent random samples from populations 1 and 2, where

$\bar{x}_1$ and $\bar{x}_2$ are sample means from populations 1 and 2

s_1 and s_2 are sample standard deviations from populations 1 and 2

n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal or at least mound-shaped and symmetric, then any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Confidence interval for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where $E \approx t_c\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

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

t_c = critical value for confidence level c

$d.f.$ = *smaller* of $n_1 - 1$ and $n_2 - 1$. Note that statistical software gives a more accurate and larget d.f. based on Satterthwaite's approximation.

EXAMPLE 9

Confidence Interval for $\mu_1 - \mu_2$, σ_1 and σ_2 unknown

Alexander Borbely is a professor at the Medical School of the University of Zurich, where he is director of the Sleep Laboratory. Dr. Borbely and his colleagues are experts on sleep, dreams, and sleep disorders. In his book *Secrets of Sleep*, Dr. Borbely discusses brain waves, which are measured in hertz, the number of oscillations per second. Rapid brain waves (wakefulness) are in the range of 16 to 25 hertz. Slow brain waves (sleep) are in the range of 4 to 8 hertz. During normal sleep, a person goes through several cycles (each cycle is about 90 minutes) of brain waves, from rapid to slow and back to rapid. During deep sleep, brain waves are at their slowest.

In his book, Professor Borbely comments that alcohol is a *poor* sleep aid. In one study, a number of subjects were given 1/2 liter of red wine before they went

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to sleep. The subjects fell asleep quickly but did not remain asleep the entire night. Toward morning, between 4 and 6 A.M., they tended to wake up and have trouble going back to sleep.

Suppose that a random sample of 29 college students was randomly divided into two groups. The first group of $n_1 = 15$ people was given 1/2 liter of red wine before going to sleep. The second group of $n_2 = 14$ people was given no alcohol before going to sleep. Everyone in both groups went to sleep at 11 P.M. The average brain wave activity (4 to 6 A.M.) was determined for each individual in the groups. Assume the average brain wave distribution in each group is mound-shaped and symmetric. The results follow:

Group 1 (x_1 values): $n_1 = 15$ (with alcohol)
Average brain wave activity in the hours 4 to 6 A.M.

16.0	19.6	19.9	20.9	20.3	20.1	16.4	20.6
20.1	22.3	18.8	19.1	17.4	21.1	22.1	

For group 1, we have the sample mean and standard deviation of

$\bar{x}_1 \approx 19.65$ and $s_1 \approx 1.86$

Group 2 (x_2 values): $n_2 = 14$ (no alcohol)
Average brain wave activity in the hours 4 to 6 A.M.

8.2	5.4	6.8	6.5	4.7	5.9	2.9
7.6	10.2	6.4	8.8	5.4	8.3	5.1

For group 2, we have the sample mean and standard deviation of

$\bar{x}_2 \approx 6.59$ and $s_2 \approx 1.91$

(a) *Check Requirements* Are the samples independent or dependent? Explain. Is it appropriate to use a Student's t distribution to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.

SOLUTION: Since the original random sample of 29 students was randomly divided into two groups, it is reasonable to say that the samples are independent. A Student's t distribution is appropriate for the $\bar{x}_1 - \bar{x}_2$ distribution because both original distributions are mound-shaped and symmetric. We don't know population standard deviations, but we can compute s_1 and s_2.

(b) Compute a 90% confidence interval for $\mu_1 - \mu_2$, the difference of population means.

SOLUTION: First we find $t_{0.90}$. We approximate the degrees of freedom $d.f.$ by using the smaller of $n_1 - 1$ and $n_2 - 1$. Since n_2 is smaller, $d.f. = n_2 - 1 = 14 - 1 = 13$. This gives us $t_{0.90} \approx 1.771$. The margin of error is then

$$E \approx t_c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.771\sqrt{\frac{1.86^2}{15} + \frac{1.91^2}{14}} \approx 1.24$$

The c confidence interval is

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$
$$(19.65 - 6.59) - 1.24 < \mu_1 - \mu_2 < (19.65 - 6.59) + 1.24$$
$$11.82 < \mu_1 - \mu_2 < 14.30$$

After further rounding we have

$$11.8 \text{ hertz} < \mu_1 - \mu_2 < 14.3 \text{ hertz}$$

(c) *Interpretation* What is the meaning of the confidence interval you computed in part (b)?

SOLUTION: μ_1 represents the population average brain wave activity for people who drank 1/2 liter of wine before sleeping. μ_2 represents the population

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

average brain wave activity for people who took no alcohol before sleeping. Both periods of measurement are from 4 to 6 A.M. We are 90% confident that the interval between 11.8 and 14.3 hertz is one that contains the difference $\mu_1 - \mu_2$. It would seem reasonable to conclude that people who drink before sleeping might wake up in the early morning and have trouble going back to sleep. Since the confidence interval from 11.8 to 14.3 contains only *positive values*, we could express this by saying that we are 90% confident that $\mu_1 - \mu_2$ is *positive*. This means that $\mu_1 - \mu_2 > 0$. Thus, we are 90% confident that $\mu_1 > \mu_2$ (that is, average brain wave activity from 4 to 6 A.M. for the group drinking wine was more than average brain wave activity for the group not drinking).

There is another method of constructing confidence intervals for $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown. Suppose the sample values s_1 and s_2 are sufficiently close and there is reason to believe that $\sigma_1 = \sigma_2$. Methods shown in Section 10.4 use sample standard deviations s_1 and s_2 to determine if $\sigma_1 = \sigma_2$. When you can assume that $\sigma_1 = \sigma_2$, it is best to use a *pooled standard deviation* to compute the margin of error. The $\bar{x}_1 - \bar{x}_2$ distribution has an *exact* Student's t distribution with $d.f. = n_1 + n_2 - 2$. Problem 31 of this section gives the details.

The formulas and requirements for using a pooled standard deviation to compute a confidence interval for $\mu_1 - \mu_2$ can be found in Problem 31 at the end of this section.

Sometimes we want several confidence intervals to hold concurrently. Problem 19 in the Chapter Review Problems discusses this situation.

Summary

What should a person do? You have independent random samples from two populations. You can compute $\bar{x}_1$, $\bar{x}_2$, s_1, and s_2 and you have the sample sizes n_1 and n_2. In any case, a confidence interval for the difference $\mu_1 - \mu_2$ of population means is

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 < \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where E is the margin of error. How do you compute E? The answer depends on how much you know about the x_1 and x_2 distributions.

Situation I (**the usual case**)

You simply don't know the population values of σ_1 and σ_2. In this situation, you use a t distribution with margin of error

$$E = t_c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where a conservative estimate for the degrees of freedom is

$$d.f. = \text{minimum of } n_1 - 1 \text{ and } n_2 - 1$$

Like a good friend, the t distribution has a reputation for being robust and forgiving. Nevertheless, some guidelines should be observed. If n_1 and n_2 are both less than 30, then x_1 and x_2 should have distributions that are mound-shaped and approximately symmetric (or, even better, normal). If both n_1 and n_2 are 30 or more, the central limit theorem (Chapter 6) implies that these restrictions can be relaxed.

Situation II (**almost never happens**)

You actually know the population values of σ_1 and σ_2. In addition, you know that x_1 and x_2 have normal distributions. If you know σ_1 and σ_2 but are not sure about the x_1 and x_2 distributions, then you must have $n_1 \geq 30$ and $n_2 \geq 30$. In this situation, you use a z distribution with margin of error

$$E = z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

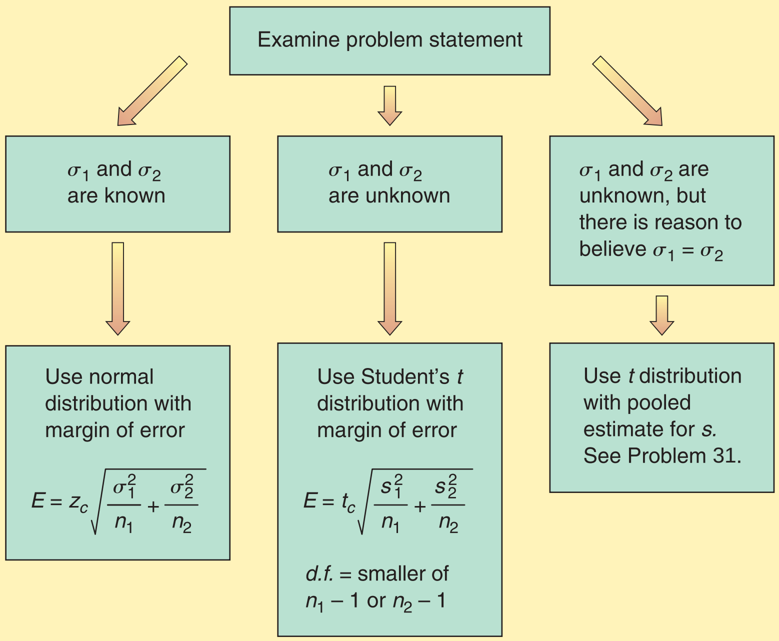
Continued

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Situation III (yes, this does sometimes occur)

You don't know σ_1 and σ_2, but the sample values s_1 and s_2 are close to each other and there is reason to believe that $\sigma_1 = \sigma_2$. This can happen when you make a slight change or alteration to a known process or method of production. The standard deviation may not change much, but the outputs or means could be very different. In this situation, you are advised to use a t distribution with a pooled standard deviation. See Problem 31 at the end of this section.

Which distribution should you use for $\bar{x}_1 - \bar{x}_2$



Estimating the Difference of Proportions $p_1 - p_2$

We conclude this section with a discussion of confidence intervals for $p_1 - p_2$, the difference of two proportions from binomial probability distributions. The main result on this topic is the following theorem.

THEOREM 7.2 Consider two binomial probability distributions

Distribution 1

n_1 = number of trials

r_1 = number of successes out of n_1 trials

p_1 = probability of success on each trial

$q_1 = 1 - p_1$ = probability of failure on each trial

$\hat{p}_1 = \dfrac{r_1}{n_1}$ = point estimate for p_1

$\hat{q}_1 = 1 - \dfrac{r_1}{n_1}$ = point estimate for q_1

Distribution 2

n_2 = number of trials

r_2 = number of successes out of n_2 trials

p_2 = probability of success on each trial

$q_2 = 1 - p_2$ = probability of failure on each trial

$\hat{p}_2 = \dfrac{r_2}{n_2}$ = point estimate for p_2

$\hat{q}_2 = 1 - \dfrac{r_2}{n_2}$ = point estimate for q_2

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For most practical applications, if the four quantities

$$n_1\hat{p}_1 \qquad n_1\hat{q}_1 \qquad n_2\hat{p}_2 \qquad n_2\hat{q}_2$$

are all larger than 5 (see Section 6.6), then the following statements are true about the random variable $\frac{r_1}{n_1} - \frac{r_2}{n_2}$:

1. $\frac{r_1}{n_1} - \frac{r_2}{n_2}$ has an approximately normal distribution.
2. The mean is $p_1 - p_2$.
3. The standard deviation is approximately

$$\hat{\sigma} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Confidence intervals for $p_1 - p_2$

Based on the preceding theorem, we can find confidence intervals for $p_1 - p_2$ in the following way:

PROCEDURE

How to find a confidence interval for $p_1 - p_2$

Requirements

Consider two independent binomial experiments.

Binomial Experiment 1	***Binomial Experiment 2***
n_1 = number of trials	n_2 = number of trials
r_1 = number of successes out of n_1 trials	r_2 = number of successes out of n_2 trials
$\hat{p}_1 = \frac{r_1}{n_1}$; $\hat{q}_1 = 1 - \hat{p}_1$	$\hat{p}_2 = \frac{r_2}{n_2}$; $\hat{q}_2 = 1 - \hat{p}_2$
p_1 = population probability of success	p_2 = population probability of success

The number of trials should be sufficiently large so that all four of the following inequalities are true:

$$n_1\hat{p}_1 > 5; \qquad n_1\hat{q}_1 > 5; \qquad n_2\hat{p}_2 > 5; \qquad n_2\hat{q}_2 > 5$$

Confidence interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) - E \le p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + E$$

where

$$E = z_c\hat{\sigma} = z_c\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

c = confidence level, $0 < c < 1$

z_c = critical value for confidence level c based on the standard normal distribution (See Table 5(b) of Appendix II for commonly used values.)

Problem 29 shows how to compute minimal sample sizes for a specified margin of error E.

EXAMPLE 10

Confidence interval for $p_1 - p_2$

In his book *Secrets of Sleep*, Professor Borbely describes research on dreams in the Sleep Laboratory at the University of Zurich Medical School. During normal sleep, there is a phase known as *REM* (rapid eye movement). For most people, REM sleep occurs about every 90 minutes or so, and it is thought that dreams occur just before or during the REM phase. Using electronic equipment in the Sleep Laboratory, it is possible to detect the REM phase in a sleeping person. If a person is wakened immediately after the REM phase, he or she usually can

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

describe a dream that has just taken place. Based on a study of over 650 people in the Zurich Sleep Laboratory, it was found that about one-third of all dream reports contain feelings of fear, anxiety, or aggression. There is a conjecture that if a person is in a good mood when going to sleep, the proportion of "bad" dreams (fear, anxiety, aggression) might be reduced.

Suppose that two groups of subjects were randomly chosen for a sleep study. In group I, before going to sleep, the subjects spent 1 hour watching a comedy movie. In this group, there were a total of $n_1 = 175$ dreams recorded, of which $r_1 = 49$ were dreams with feelings of anxiety, fear, or aggression. In group II, the subjects did not watch a movie but simply went to sleep. In this group, there were a total of $n_2 = 180$ dreams recorded, of which $r_2 = 63$ were dreams with feelings of anxiety, fear, or aggression.

(a) *Check Requirements* Why could groups I and II be considered independent binomial distributions? Why do we have a "large-sample" situation?

SOLUTION: Since the two groups were chosen randomly, it is reasonable to assume that neither group's responses would be related to the other's. In both groups, each recorded dream could be thought of as a trial, with success being a dream with feelings of fear, anxiety, or aggression.

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{49}{175} = 0.28 \quad \text{and} \quad \hat{q}_1 = 1 - \hat{p}_1 = 0.72$$

$$\hat{p}_2 = \frac{r_2}{n_2} = \frac{63}{180} = 0.35 \quad \text{and} \quad \hat{q}_2 = 1 - \hat{p}_2 = 0.65$$

Since

$$n_1\hat{p}_1 = 49 > 5 \qquad n_1\hat{q}_1 = 126 > 5$$
$$n_2\hat{p}_2 = 63 > 5 \qquad n_2\hat{q}_2 = 117 > 5$$

large-sample theory is appropriate.

(b) What is $p_1 - p_2$? Compute a 95% confidence interval for $p_1 - p_2$.

SOLUTION: p_1 is the population proportion of successes (bad dreams) for all people who watched comedy movies before bed. Thus, p_1 can be thought of as the percentage of bad dreams for all people who were in a "good mood" when they went to bed. Likewise, p_2 is the percentage of bad dreams for the population of all people who just went to bed (no movie). The difference $p_1 - p_2$ is the population difference.

To find a confidence interval for $p_1 - p_2$, we need the values of z_c, $\hat{\sigma}$, and then E. From Table 7-2, we see that $z_{0.95} = 1.96$, so

$$\hat{\sigma} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{(0.28)(0.72)}{175} + \frac{(0.35)(0.65)}{180}}$$
$$\approx \sqrt{0.0024} \approx 0.0492$$
$$E = z_c\hat{\sigma} = 1.96(0.0492) \approx 0.096$$

$$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$$
$$(0.28 - 0.35) - 0.096 < p_1 - p_2 < (0.28 - 0.35) + 0.096$$
$$-0.166 < p_1 - p_2 < 0.026$$

(c) *Interpretation* What is the meaning of the confidence interval constructed in part (b)?

SOLUTION: We are 95% sure that the interval between −16.6% and 2.6% is one that contains the percentage difference of "bad" dreams for group I and group II. Since the interval −0.166 to 0.026 is not all negative (or all positive), we cannot say that $p_1 - p_2 < 0$ (or $p_1 - p_2 > 0$). Thus, at the 95% confidence level, we *cannot* conclude that $p_1 < p_2$ or $p_1 > p_2$. The comedy

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

movies before bed helped some people reduce the percentage of "bad" dreams, but at the 95% confidence level, we cannot say that the *population difference* is reduced.

CRITICAL THINKING

Interpreting Confidence Intervals for Differences

As we have seen in the preceding examples, at the c confidence level we can determine how two means or proportions from independent random samples are related. The next procedure summarizes the results.

PROCEDURE

How to interpret confidence intervals for differences

Suppose we construct a c% confidence interval for $\mu_1 - \mu_2$ (or $p_1 - p_2$). Then three cases arise:

1. The c% confidence interval contains only *negative values* (see Example 8). In this case, we conclude that $\mu_1 - \mu_2 < 0$ (or $p_1 - p_2 < 0$), and we are therefore c% confident that $\mu_1 < \mu_2$ (or $p_1 < p_2$).
2. The c% confidence interval contains only *positive values* (see Example 9). In this case, we conclude that $\mu_1 - \mu_2 > 0$ (or $p_1 - p_2 > 0$), and we can be c% confident that $\mu_1 > \mu_2$ (or $p_1 > p_2$).
3. The c% confidence interval contains *both positive and negative values* (see Example 10). In this case, we cannot at the c% confidence level conclude that either μ_1 or μ_2 (or p_1 or p_2) is larger. However, if we *reduce* the confidence level c to a *smaller value*, then the confidence interval will, in general, be shorter (explain why). Another approach (when possible) is to increase the sample sizes n_1 and n_2. This would also tend to make the confidence interval shorter (explain why). A shorter confidence interval *might* put us back into case 1 or case 2 above (again, explain why).

In Section 8.5, we will see another method to determine if two means or proportions from independent random samples are equal.

GUIDED EXERCISE 7 *Interpreting a confidence interval*

(a) A study reported a 90% confidence interval for the difference of means to be

$$10 < \mu_1 - \mu_2 < 20$$

For this interval, what can you conclude about the respective values of μ_1 and μ_2?

At a 90% level of confidence, we can say that the difference $\mu_1 - \mu_2$ is positive, so $\mu_1 - \mu_2 > 0$ and $\mu_1 > \mu_2$.

(b) A study reported a 95% confidence interval for the difference of proportions to be

$$-0.32 < p_1 - p_2 < 0.16$$

For this interval, what can you conclude about the respective values of p_1 and p_2?

At the 95% confidence level, we see that the difference of proportions ranges from negative to positive values. We cannot tell from this interval if p_1 is greater than p_2 or p_1 is less than p_2.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

TECH NOTES The TI-84Plus/TI-83Plus/TI-*n*spire calculators and Minitab supply confidence intervals for the difference of means and for the difference of proportions.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Use the **STAT** key and highlight **TESTS**. The choice **9:2-SampZInt** finds confidence intervals for differences of means when σ_1 and σ_2 are known. Choice **0:2-SampTInt** finds confidence intervals for differences of means when σ_1 and σ_2 are unknown. In general, use No for Pooled. However, if $\sigma_1 \approx \sigma_2$, use Yes for Pooled. Choice **B:2-PropZInt** provides confidence intervals for proportions.

Minitab Use the menu choice **STAT** ➤ **Basic Statistics** ➤ **2 sample t** or **2 proportions**. Minitab always uses the Student's t distribution for $\mu_1 - \mu_2$ confidence intervals. If the variances are equal, check "assume equal variances."

VIEWPOINT What's the Difference?

Will two 15-minute piano lessons a week significantly improve a child's analytical reasoning skills? Why piano? Why not computer keyboard instruction or maybe voice lessons? Professor Frances Rauscher, University of Wisconsin, and Professor Gordon Shaw, University of California at Irvine, claim there is a difference! How could this be measured? A large number of piano students were given complicated tests of mental ability. Independent control groups of other students were given the same tests. Techniques involving the study of differences of means were used to draw the conclusion that students taking piano lessons did better on tests measuring analytical reasoning skills. (Reported in The Denver Post.*)*

SECTION 7.4 PROBLEMS

Answers may vary slightly due to rounding.

1. *Statistical Literacy* When are two random samples independent?
2. *Statistical Literacy* When are two random samples dependent?
3. *Critical Thinking* Josh and Kendra each calculated a 90% confidence interval for the difference of means using a Student's t distribution for random samples of size $n_1 = 20$ and $n_2 = 31$. Kendra followed the convention of using the smaller sample size to compute $d.f. = 19$. Josh used his calculator and Satterthwaite's approximation and obtained $d.f. \approx 36.3$. Which confidence interval is shorter? Which confidence interval is more conservative in the sense that the margin of error is larger?
4. *Critical Thinking* If a 90% confidence interval for the difference of means $\mu_1 - \mu_2$ contains all positive values, what can we conclude about the relationship between μ_1 and μ_2 at the 90% confidence level?
5. *Critical Thinking* If a 90% confidence interval for the difference of means $\mu_1 - \mu_2$ contains all negative values, what can we conclude about the relationship between μ_1 and μ_2 at the 90% confidence level?
6. *Critical Thinking* If a 90% confidence interval for the difference of proportions contains some positive and some negative values, what can we conclude about the relationship between p_1 and p_2 at the 90% confidence level?
7. *Basic Computation: Confidence Interval for $\mu_1 - \mu_2$* Consider two independent normal distributions. A random sample of size $n_1 = 20$ from the first distribution showed $\bar{x}_1 = 12$ and a random sample of size $n_2 = 25$ from the second distribution showed $\bar{x}_2 = 14$.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(a) ***Check Requirements*** If σ_1 and σ_2 are known, what distribution does $\bar{x}_1 - \bar{x}_2$ follow? Explain.
(b) Given $\sigma_1 = 3$ and $\sigma_2 = 4$, find a 90% confidence interval for $\mu_1 - \mu_2$
(c) ***Check Requirements*** Suppose σ_1 and σ_2 are both unknown, but from the random samples, you know $s_1 = 3$ and $s_2 = 4$. What distribution approximates the $\bar{x}_1 - \bar{x}_2$ distribution? What are the degrees of freedom? Explain.
(d) With $s_1 = 3$ and $s_2 = 4$, find a 90% confidence interval for $\mu_1 - \mu_2$.
(e) If you have an appropriate calculator or computer software, find a 90% confidence interval for $\mu_1 - \mu_2$ using degrees of freedom based on Satterthwaite's approximation.
(f) ***Interpretation*** Based on the confidence intervals you computed, can you be 90% confident that μ_1 is smaller than μ_2? Explain.

8. ***Basic Computation: Confidence Interval for $\mu_1 - \mu_2$*** Consider two independent distributions that are mound-shaped. A random sample of size $n_1 = 36$ from the first distribution showed $\bar{x}_1 = 15$, and a random sample of size $n_2 = 40$ from the second distribution showed $\bar{x}_2 = 14$.
(a) ***Check Requirements*** If σ_1 and σ_2 are known, what distribution does $\bar{x}_1 - \bar{x}_2$ follow? Explain.
(b) Given $\sigma_1 = 3$ and $\sigma_2 = 4$, find a 95% confidence interval for $\mu_1 - \mu_2$.
(c) ***Check Requirements*** Suppose σ_1 and σ_2 are both unknown, but from the random samples, you know $s_1 = 3$ and $s_2 = 4$. What distribution approximates the $\bar{x}_1 - \bar{x}_2$ distribution? What are the degrees of freedom? Explain.
(d) With $s_1 = 3$ and $s_2 = 4$, find a 95% confidence interval for $\mu_1 - \mu_2$.
(e) If you have an appropriate calculator or computer software, find a 95% confidence interval for $\mu_1 - \mu_2$ using degrees of freedom based on Satterthwaite's approximation.
(f) ***Interpretation*** Based on the confidence intervals you computed, can you be 95% confident that μ_1 is larger than μ_2? Explain.

9. ***Basic Computation: Confidence Interval for $p_1 - p_2$*** Consider two independent binomial experiments. In the first one, 40 trials had 10 successes. In the second one, 50 trials had 15 successes.
(a) ***Check Requirements*** Is it appropriate to use a normal distribution to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
(b) Find a 90% confidence interval for $p_1 - p_2$.
(c) ***Interpretation*** Based on the confidence interval you computed, can you be 90% confident that p_1 is less than p_2? Explain.

10. ***Basic Computation: Confidence Interval for $p_1 - p_2$*** Consider two independent binomial experiments. In the first one, 40 trials had 15 successes. In the second one, 60 trials had 6 successes.
(a) ***Check Requirements*** Is it appropriate to use a normal distribution to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
(b) Find a 95% confidence interval for $p_1 - p_2$.
(c) ***Interpretation*** Based on the confidence interval you computed, can you be 95% confident that p_1 is more than p_2? Explain.

11. ***Archaeology: Ireland*** Inorganic phosphorous is a naturally occurring element in all plants and animals, with concentrations increasing progressively up the food chain (fruit < vegetables < cereals < nuts < corpse). Geochemical surveys take soil samples to determine phosphorous content (in ppm, parts per million). A high phosphorous content may or may not indicate an ancient burial site, food storage site, or even a garbage dump. The Hill of Tara is a very important archaeological site in Ireland. It is by legend the seat of Ireland's ancient high kings (Reference: *Tara, An Archaeological Survey* by Conor Newman, Royal Irish Academy, Dublin). Independent random samples from two regions in Tara gave the following phosphorous measurements (in ppm). Assume the population distributions of phosphorous are mound-shaped and symmetric for these two regions.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Region I: x_1; $n_1 = 12$

540	810	790	790	340	800
890	860	820	640	970	720

Region II: x_2; $n_2 = 16$

750	870	700	810	965	350	895	850
635	955	710	890	520	650	280	993

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 747.5$, $s_1 \approx 170.4$, $\bar{x}_2 \approx 738.9$, and $s_2 \approx 212.1$.
(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find a 90% confidence interval for $\mu_1 - \mu_2$.
(c) *Interpretation* Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 90% level of confidence, is one region more interesting than the other from a geochemical perspective?
(d) *Check Requirements* Which distribution (standard normal or Student's t) did you use? Why?

12. *Archaeology: Ireland* Please see the setting and reference in Problem 11. Independent random samples from two regions (not those cited in Problem 11) gave the following phosphorous measurements (in ppm). Assume the distribution of phosphorous is mound-shaped and symmetric for these two regions.

Region I: x_1; $n_1 = 15$

855	1550	1230	875	1080	2330	1850	1860
2340	1080	910	1130	1450	1260	1010	

Region II: x_2; $n_2 = 14$

540	810	790	1230	1770	960	1650	860
890	640	1180	1160	1050	1020		

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 1387.3$, $s_1 \approx 498.3$, $\bar{x}_2 \approx 1039.3$, and $s_2 \approx 346.7$.
(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find an 80% confidence interval for $\mu_1 - \mu_2$.
(c) *Interpretation* Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 80% level of confidence, is one region more interesting than the other from a geochemical perspective?
(d) *Check Requirements* Which distribution (standard normal or Student's t) did you use? Why?

13. *Large U.S. Companies: Foreign Revenue* For large U.S. companies, what percentage of their total income comes from foreign sales? A random sample of technology companies (IBM, Hewlett-Packard, Intel, and others) gave the following information.

Technology companies, % foreign revenue: x_1; $n_1 = 16$

62.8	55.7	47.0	59.6	55.3	41.0	65.1	51.1
53.4	50.8	48.5	44.6	49.4	61.2	39.3	41.8

Another independent random sample of basic consumer product companies (Goodyear, Sarah Lee, H.J. Heinz, Toys "Я" Us) gave the following information.

Basic consumer product companies, % foreign revenue: x_2; $n_2 = 17$

28.0	30.5	34.2	50.3	11.1	28.8	40.0	44.9
40.7	60.1	23.1	21.3	42.8	18.0	36.9	28.0
32.5							

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(Reference: *Forbes Top Companies*.) Assume that the distributions of percentage foreign revenue are mound-shaped and symmetric for these two company types.

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 51.66$, $s_1 \approx 7.93$, $\bar{x}_2 \approx 33.60$, and $s_2 \approx 12.26$.

(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find an 85% confidence interval for $\mu_1 - \mu_2$.

(c) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 85% level of confidence, do technology companies have a greater percentage foreign revenue than basic consumer product companies?

(d) *Check Requirements* Which distribution (standard normal or Student's t) did you use? Why?

14. *Pro Football and Basketball: Weights of Players* Independent random samples of professional football and basketball players gave the following information (References: *Sports Encyclopedia of Pro Football* and *Official NBA Basketball Encyclopedia*). *Note:* These data are also available for download at the Online Study Center. Assume that the weight distributions are mound-shaped and symmetric.

Weights (in lb) of pro football players: x_1; $n_1 = 21$

245	262	255	251	244	276	240	265	257	252	282
256	250	264	270	275	245	275	253	265	270	

Weights (in lb) of pro basketball players: x_2; $n_2 = 19$

205	200	220	210	191	215	221	216	228	207
225	208	195	191	207	196	181	193	201	

Richard Paul Kane,2010/Used under license from Shutterstock.Com

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 259.6$, $s_1 \approx 12.1$, $\bar{x}_2 \approx 205.8$, and $s_2 \approx 12.9$.

(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find a 99% confidence interval for $\mu_1 - \mu_2$.

(c) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 99% level of confidence, do professional football players tend to have a higher population mean weight than professional basketball players?

(d) Which distribution (standard normal or Student's t) did you use? Why?

15. *Pro Football and Basketball: Heights of Players* Independent random samples of professional football and basketball players gave the following information (References: *Sports Encyclopedia of Pro Football* and *Official NBA Basketball Encyclopedia*). *Note:* These data are also available for download at the Online Study Center.

Heights (in ft) of pro football players: x_1; $n_1 = 45$

6.33	6.50	6.50	6.25	6.50	6.33	6.25	6.17	6.42	6.33
6.42	6.58	6.08	6.58	6.50	6.42	6.25	6.67	5.91	6.00
5.83	6.00	5.83	5.08	6.75	5.83	6.17	5.75	6.00	5.75
6.50	5.83	5.91	5.67	6.00	6.08	6.17	6.58	6.50	6.25
6.33	5.25	6.67	6.50	5.83					

Heights (in ft) of pro basketball players: x_2; $n_2 = 40$

6.08	6.58	6.25	6.58	6.25	5.92	7.00	6.41	6.75	6.25
6.00	6.92	6.83	6.58	6.41	6.67	6.67	5.75	6.25	6.25
6.50	6.00	6.92	6.25	6.42	6.58	6.58	6.08	6.75	6.50
6.83	6.08	6.92	6.00	6.33	6.50	6.58	6.83	6.50	6.58

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 6.179$, $s_1 \approx 0.366$, $\bar{x}_2 \approx 6.453$, and $s_2 \approx 0.314$.
(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find a 90% confidence interval for $\mu_1 - \mu_2$.
(c) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 90% level of confidence, do professional football players tend to have a higher population mean height than professional basketball players?
(d) *Check Requirements* Which distribution (standard normal or Student's t) did you use? Why? Do you need information about the height distributions? Explain.

16. *Botany: Iris* The following data represent petal lengths (in cm) for independent random samples of two species of iris (Reference: E. Anderson, *Bulletin American Iris Society*). *Note:* These data are also available for download at the Online Study Center.

Petal length (in cm) of *Iris virginica*: x_1; $n_1 = 35$

5.1 5.8 6.3 6.1 5.1 5.5 5.3 5.5 6.9 5.0 4.9 6.0 4.8 6.1 5.6 5.1
5.6 4.8 5.4 5.1 5.1 5.9 5.2 5.7 5.4 4.5 6.1 5.3 5.5 6.7 5.7 4.9
4.8 5.8 5.1

Petal length (in cm) of *Iris setosa*: x_2; $n_2 = 38$

1.5 1.7 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.4 1.7 1.0 1.7 1.9 1.6 1.4
1.5 1.4 1.2 1.3 1.5 1.3 1.6 1.9 1.4 1.6 1.5 1.4 1.6 1.2 1.9 1.5
1.6 1.4 1.3 1.7 1.5 1.7

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 5.48$, $s_1 \approx 0.55$, $\bar{x}_2 \approx 1.49$, and $s_2 \approx 0.21$.
(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find a 99% confidence interval for $\mu_1 - \mu_2$.
(c) *Interpretation* Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 99% level of confidence, is the population mean petal length of *Iris virginica* longer than that of *Iris setosa*?
(d) *Check Requirements* Which distribution (standard normal or Student's t) did you use? Why? Do you need information about the petal length distributions? Explain.

17. *Myers–Briggs: Marriage Counseling* Isabel Myers was a pioneer in the study of personality types. She identified four basic personality preferences, which are described at length in the book *A Guide to the Development and Use of the Myers–Briggs Type Indicator* by Myers and McCaulley (Consulting Psychologists Press). Marriage counselors know that couples who have none of the four preferences in common may have a stormy marriage. Myers took a random sample of 375 married couples and found that 289 had two or more personality preferences in common. In another random sample of 571 married couples, it was found that only 23 had no preferences in common. Let p_1 be the population proportion of all married couples who have two or more personality preferences in common. Let p_2 be the population proportion of all married couples who have no personality perferences in common.
(a) *Check Requirements* Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
(b) Find a 99% confidence interval for $p_1 - p_2$.
(c) *Interpretation* Explain the meaning of the confidence interval in part (a) in the context of this problem. Does the confidence interval contain all positive, all negative, or both positive and negative numbers? What does this tell

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

you (at the 99% confidence level) about the proportion of married couples with two or more personality preferences in common compared with the proportion of married couples sharing no personality preferences in common?

18. *Myers–Briggs: Marriage Counseling* Most married couples have two or three personality preferences in common (see reference in Problem 17). Myers used a random sample of 375 married couples and found that 132 had three preferences in common. Another random sample of 571 couples showed that 217 had two personality preferences in common. Let p_1 be the population proportion of all married couples who have three personality preferences in common. Let p_2 be the population proportion of all married couples who have two personality preferences in common.
    (a) *Check Requirements* Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
    (b) Find a 90% confidence interval for $p_1 - p_2$.
    (c) *Interpretation* Examine the confidence interval in part (a) and explain what it means in the context of this problem. Does the confidence interval contain all positive, all negative, or both positive and negative numbers? What does this tell you about the proportion of married couples with three personality preferences in common compared with the proportion of couples with two preferences in common (at the 90% confidence level)?

19. *Yellowstone National Park: Old Faithful Geyser* The U.S. Geological Survey compiled historical data about Old Faithful Geyser (Yellowstone National Park) from 1870 to 1987. Some of these data are published in the book *The Story of Old Faithful*, by G. D. Marler (Yellowstone Association Press). Let x_1 be a random variable that represents the time interval (in minutes) between Old Faithful's eruptions for the years 1948 to 1952. Based on 9340 observations, the sample mean interval was $\bar{x}_1 = 63.3$ minutes. Let x_2 be a random variable that represents the time interval in minutes between Old Faithful's eruptions for the years 1983 to 1987. Based on 25,111 observations, the sample mean time interval was $\bar{x}_2 = 72.1$ minutes. Historical data suggest that $\sigma_1 = 9.17$ minutes and $\sigma_2 = 12.67$ minutes. Let μ_1 be the population mean of x_1 and let μ_2 be the population mean of x_2.
    (a) *Check Requirements* Which distribution, normal or Student's t, do we use to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.
    (b) Compute a 99% confidence interval for $\mu_1 - \mu_2$.
    (c) *Interpretation* Comment on the meaning of the confidence interval in the context of this problem. Does the interval consist of positive numbers only? negative numbers only? a mix of positive and negative numbers? Does it appear (at the 99% confidence level) that a change in the interval length between eruptions has occurred? Many geologic experts believe that the distribution of eruption times of Old Faithful changed after the major earthquake that occurred in 1959.

20. *Psychology: Parental Sensitivity* "Parental Sensitivity to Infant Cues: Similarities and Differences Between Mothers and Fathers" by M. V. Graham (*Journal of Pediatric Nursing*, Vol. 8, No. 6) reports a study of parental empathy for sensitivity cues and baby temperament (higher scores mean more empathy). Let x_1 be a random variable that represents the score of a mother on an empathy test (as regards her baby). Let x_2 be the empathy score of a father. A random sample of 32 mothers gave a sample mean of $\bar{x}_1 = 69.44$. Another random sample of 32 fathers gave $\bar{x}_2 = 59$. Assume that $\sigma_1 = 11.69$ and $\sigma_2 = 11.60$.
    (a) *Check Requirements* Which distribution, normal or Student's t, do we use to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.
    (b) Let μ_1 be the population mean of x_1 and let μ_2 be the population mean of x_2. Find a 99% confidence interval for $\mu_1 - \mu_2$.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(c) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the confidence interval contain all positive, all negative, or both positive and negative numbers? What does this tell you about the relationship between average empathy scores for mothers compared with those for fathers at the 99% confidence level?

21. *Navajo Culture: Traditional Hogans* S. C. Jett is a professor of geography at the University of California, Davis. He and a colleague, V. E. Spencer, are experts on modern Navajo culture and geography. The following information is taken from their book *Navajo Architecture: Forms, History, Distributions* (University of Arizona Press). On the Navajo Reservation, a random sample of 210 permanent dwellings in the Fort Defiance region showed that 65 were traditional Navajo hogans. In the Indian Wells region, a random sample of 152 permanent dwellings showed that 18 were traditional hogans. Let p_1 be the population proportion of all traditional hogans in the Fort Defiance region, and let p_2 be the population proportion of all traditional hogans in the Indian Wells region.
(a) *Check Requirements* Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
(b) Find a 99% confidence interval for $p_1 - p_2$.
(c) *Interpretation* Examine the confidence interval and comment on its meaning. Does it include numbers that are all positive? all negative? mixed? What if it is hypothesized that Navajo who follow the traditional culture of their people tend to occupy hogans? Comment on the confidence interval for $p_1 - p_2$ in this context.

22. *Archaeology: Cultural Affiliation* "Unknown cultural affiliations and loss of identity at high elevations." These words are used to propose the hypothesis that archaeological sites tend to lose their identity as altitude extremes are reached. This idea is based on the notion that prehistoric people tended *not* to take trade wares to temporary settings and/or isolated areas (Source: *Prehistoric New Mexico: Background for Survey*, by D. E. Stuart and R. P. Gauthier, University of New Mexico Press). As elevation zones of prehistoric people (in what is now the state of New Mexico) increased, there seemed to be a loss of artifact identification. Consider the following information.

Elevation Zone	Number of Artifacts	Number Unidentified
7000–7500 ft	112	69
5000–5500 ft	140	26

Let p_1 be the population proportion of unidentified archaeological artifacts at the elevation zone 7000–7500 feet in the given archaeological area. Let p_2 be the population proportion of unidentified archaeological artifacts at the elevation zone 5000–5500 feet in the given archaeological area.
(a) *Check Requirements* Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
(b) Find a 99% confidence interval for $p_1 - p_2$.
(c) *Interpretation* Explain the meaning of the confidence interval in the context of this problem. Does the confidence interval contain all positive numbers? all negative numbers? both positive and negative numbers? What does this tell you (at the 99% confidence level) about the comparison of the population proportion of unidentified artifacts at high elevations (7000–7500 feet) with the population proportion of unidentified artifacts at lower elevations (5000–5500 feet)? How does this relate to the stated hypothesis?

23. *Wildlife: Wolves* David E. Brown is an expert in wildlife conservation. In his book *The Wolf in the Southwest: The Making of an Endangered Species* (University of Arizona Press), he lists the following weights of adult gray wolves from two regions in Old Mexico.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chihuahua region: x_1 variable in pounds

86	75	91	70	79
80	68	71	74	64

Durango region: x_2 variable in pounds

68	72	79	68	77	89	62	55	68
68	59	63	66	58	54	71	59	67

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 = 75.80$ pounds, $s_1 = 8.32$ pounds, $\bar{x}_2 = 66.83$ pounds, and $s_2 = 8.87$ pounds.
(b) *Check Requirements* Assuming that the original distribution of the weights of wolves are mound-shaped and symmetric, what distribution can be used to approximate the $\bar{x}_1 - \bar{x}_2$ distribution? Explain.
(c) Let μ_1 be the mean weight of the population of all gray wolves in the Chihuahua region. Let μ_2 be the mean weight of the population of all gray wolves in the Durango region. Find an 85% confidence interval for $\mu_1 - \mu_2$.
(d) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 85% level of confidence, what can you say about the comparison of the average weight of gray wolves in the Chihuahua region with the average weight of gray wolves in the Durango region?

24. *Medical: Plasma Compress* At Community Hospital, the burn center is experimenting with a new plasma compress treatment. A random sample of $n_1 = 316$ patients with minor burns received the plasma compress treatment. Of these patients, it was found that 259 had no visible scars after treatment. Another random sample of $n_2 = 419$ patients with minor burns received no plasma compress treatment. For this group, it was found that 94 had no visible scars after treatment. Let p_1 be the population proportion of all patients with minor burns receiving the plasma compress treatment who have no visible scars. Let p_2 be the population proportion of all patients with minor burns not receiving the plasma compress treatment who have no visible scars.
(a) *Check Requirements* Can a normal distribution be used to approximate the $\hat{p}_1 - \hat{p}_2$ distribution? Explain.
(b) Find a 95% confidence interval for $p_1 - p_2$.
(c) *Interpretation* Explain the meaning of the confidence interval found in part (b) in the context of the problem. Does the interval contain numbers that are all positive? all negative? both positive and negative? At the 95% level of confidence, does treatment with plasma compresses seem to make a difference in the proportion of patients with visible scars from minor burns?

25. *Psychology: Self-Esteem* Female undergraduates in randomized groups of 15 took part in a self-esteem study ("There's More to Self-Esteem than Whether It Is High or Low: The Importance of Stability of Self-Esteem," by M. H. Kernis et al., *Journal of Personality and Social Psychology*, Vol. 65, No. 6). The study measured an index of self-esteem from the points of view competence, social acceptance, and physical attractiveness. Let x_1, x_2, and x_3 be random variables representing the measure of self-esteem through x_1 (competence), x_2 (social acceptance), and x_3 (attractiveness). Higher index values mean a more positive influence on self-esteem.

Variable	Sample Size	Mean $\bar{x}$	Standard Deviation s	Population Mean
x_1	15	19.84	3.07	μ_1
x_2	15	19.32	3.62	μ_2
x_3	15	17.88	3.74	μ_3

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(a) Find an 85% confidence interval for $\mu_1 - \mu_2$.
(b) Find an 85% confidence interval for $\mu_1 - \mu_3$.
(c) Find an 85% confidence interval for $\mu_2 - \mu_3$.
(d) *Interpretation* Comment on the meaning of each of the confidence intervals found in parts (a), (b), and (c). At the 85% confidence level, what can you say about the average differences in influence on self-esteem between competence and social acceptance? between competence and attractiveness? between social acceptance and attractiveness?

26. *Focus Problem: Wood Duck Nests* In the Focus Problem at the beginning of this chapter, a study was described comparing the hatch ratios of wood duck nesting boxes. Group I nesting boxes were well separated from each other and well hidden by available brush. There were a total of 474 eggs in group I boxes, of which a field count showed about 270 had hatched. Group II nesting boxes were placed in highly visible locations and grouped closely together. There were a total of 805 eggs in group II boxes, of which a field count showed about 270 had hatched.
(a) Find a point estimate $\hat{p}_1$ for p_1, the proportion of eggs that hatched in group I nest box placements. Find a 95% confidence interval for p_1.
(b) Find a point estimate $\hat{p}_2$ for p_2, the proportion of eggs that hatched in group II nest box placements. Find a 95% confidence interval for p_2.
(c) Find a 95% confidence interval for $p_1 - p_2$. Does the interval indicate that the proportion of eggs hatched from group I nest boxes is higher than, lower than, or equal to the proportion of eggs hatched from group II nest boxes?
(d) *Interpretation* What conclusions about placement of nest boxes can be drawn? In the article discussed in the Focus Problem, additional concerns are raised about the higher cost of placing and maintaining group I nest box placements. Also at issue is the cost efficiency per successful wood duck hatch.

27. *Critical Thinking: Different Confidence Levels*
(a) Suppose a 95% confidence interval for the difference of means contains both positive and negative numbers. Will a 99% confidence interval based on the same data necessarily contain both positive and negative numbers? Explain. What about a 90% confidence interval? Explain.
(b) Suppose a 95% confidence interval for the difference of proportions contains all positive numbers. Will a 99% confidence interval based on the same data necessarily contain all positive numbers as well? Explain. What about a 90% confidence interval? Explain.

28. *Expand Your Knowledge: Sample Size, Difference of Means* What about sample size? If we want a confidence interval with maximal margin of error E and level of confidence c, then Section 7.1 shows us which formulas to apply for a *single* mean μ and Section 7.3 shows us formulas for a *single* proportion p.
(a) How about a *difference of means*? When σ_1 and σ_2 are known, the margin of error E for a c% confidence interval is

$$E = z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Let us make the simplifying assumption that we have *equal sample sizes* n so that $n = n_1 = n_2$. We also assume that $n \geq 30$. In this context, we get

$$E = z_c\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = \frac{z_c}{\sqrt{n}}\sqrt{\sigma_1^2 + \sigma_2^2}$$

Solve this equation for n and show that

$$n = \left(\frac{z_c}{E}\right)^2(\sigma_1^2 + \sigma_2^2)$$

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) In Problem 15 (football and basketball player heights), suppose we want to be 95% sure that our estimate $\bar{x}_1 - \bar{x}_2$ for the difference $\mu_1 - \mu_2$ has a margin of error $E = 0.05$ foot. How large should the sample size be (assuming equal sample size— i.e., $n = n_1 = n_2$)? Since we do not know σ_1 or σ_2 and $n \geq 30$, use s_1 and s_2, respectively, from the preliminary sample of Problem 15.

(c) In Problem 16 (petal lengths of two iris species), suppose we want to be 90% sure that our estimate $\bar{x}_1 - \bar{x}_2$ for the difference $\mu_1 - \mu_2$ has a margin of error $E = 0.1$ cm. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)? Since we do not know σ_1 or σ_2 and $n \geq 30$, use s_1 and s_2, respectively, from the preliminary sample of Problem 16.

29. *Expand Your Knowledge: Sample Size, Difference of Proportions* What about the sample size n for confidence intervals for the difference of proportions $p_1 - p_2$? Let us make the following assumptions: *equal sample sizes* $n = n_1 = n_2$ and *all four quantities* $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ *are greater than 5*. Those readers familiar with algebra can use the procedure outlined in Problem 28 to show that if we have preliminary estimates $\hat{p}_1$ and $\hat{p}_2$ and a given maximal margin of error E for a specified confidence level c, then the sample size n should be at least

$$n = \left(\frac{z_c}{E}\right)^2 (\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2)$$

However, if we have no preliminary estimates for $\hat{p}_1$ and $\hat{p}_2$, then the theory similar to that used in this section tells us that the sample size n should be at least

$$n = \frac{1}{2}\left(\frac{z_c}{E}\right)^2$$

(a) In Problem 17 (Myers–Briggs personality type indicators in common for married couples), suppose we want to be 99% confident that our estimate $\hat{p}_1 - \hat{p}_2$ for the difference $p_1 - p_2$ has a maximal margin of error $E = 0.04$. Use the preliminary estimates $\hat{p}_1 = 289/375$ for the proportion of couples sharing two personality traits and $\hat{p}_2 = 23/571$ for the proportion having no traits in common. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)?

(b) Suppose that in Problem 17 we have no preliminary estimates for $\hat{p}_1$ and $\hat{p}_2$ and we want to be 95% confident that our estimate $\hat{p}_1 - \hat{p}_2$ for the difference $p_1 - p_2$ has a maximal margin of error $E = 0.05$. How large should the sample size be (assuming equal sample size—i.e., $n = n_1 = n_2$)?

30. *Expand Your Knowledge: Software Approximation for Degrees of Freedom* Given x_1 and x_2 distributions that are normal or approximately normal with unknown σ_1 and σ_2, the value of t corresponding to $\bar{x}_1 - \bar{x}_2$ has a distribution that is approximated by a Student's t distribution. We use the convention that the degrees of freedom are approximately the smaller of $n_1 - 1$ and $n_2 - 1$. However, a more accurate estimate for the appropriate degrees of freedom is given by *Satterthwaite's formula*

$$d.f. \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

where s_1, s_2, n_1, and n_2 are the respective sample standard deviations and sample sizes of independent random samples from the x_1 and x_2 distributions. This is the approximation used by most statistical software. When both n_1 and n_2 are 5 or larger, it is quite accurate. The degrees of freedom computed from this formula are either truncated or not rounded.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(a) Use the data of Problem 14 (weights of pro football and pro basketball players) to compute *d.f.* using the formula. Compare the result to 36, the value generated by Minitab. Did Minitab truncate?
(b) Compute a 99% confidence interval using $d.f. \approx 36$. (Using Table 6 requires using $d.f. = 35$.) Compare this confidence interval to the one you computed in Problem 14. Which *d.f.* gives the longer interval?

31. ***Expand Your Knowledge: Pooled Two-Sample Procedures*** Under the condition that both populations have equal standard deviations ($\sigma_1 = \sigma_2$), we can pool the standard deviations and use a Student's t distribution with degrees of freedom $d.f. = n_1 + n_2 - 2$ to find the margin of error of a c confidence interval for $\mu_1 - \mu_2$. This technique demonstrates another commonly used method of computing confidence intervals for $\mu_1 - \mu_2$.

PROCEDURE

HOW TO FIND A CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$ WHEN $\sigma_1 = \sigma_2$

Requirements

Consider two *independent* random samples, where

$\bar{x}_1$ and $\bar{x}_2$ are sample means from populations 1 and 2
s_1 and s_2 are sample standard deviations from populations 1 and 2
n_1 and n_2 are sample sizes from populations 1 and 2

If you can assume that both population distributions 1 and 2 are normal or at least mound-shaped and symmetric, then any sample sizes n_1 and n_2 will work. If you cannot assume this, then use sample sizes $n_1 \geq 30$ and $n_2 \geq 30$.

Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$

$$(\hat{x}_1 - \hat{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

where

$$E = t_c s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \qquad \text{(pooled standard deviation)}$$

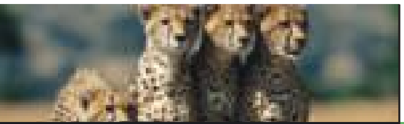
c = confidence level $(0 < c < 1)$

t_c = critical value for confidence level c and degrees of freedom $d.f. = n_1 + n_2 - 2$(See Table 6 of Appendix II.)

Note: With statistical software, select pooled variance or equal variance options.

(a) There are many situations in which we want to compare means from populations having standard deviations that are equal. The pooled standard deviation method applies even if the standard deviations are known to be only approximately equal. (See Section 10.4 for methods to test that $\sigma_1 = \sigma_2$.) Consider Problem 23 regarding weights of grey wolves in two regions. Notice that $s_1 = 8.32$ pounds and $s_2 = 8.87$ pounds are fairly close. Use the method of pooled standard deviation to find an 85% confidence interval for the difference in population mean weights of grey wolves in the Chihuahua region compared with those in the Durango region.
(b) Compare the confidence interval computed in part (a) with that computed in Problem 23. Which method has the larger degrees of freedom? Which method has the longer confidence interval?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Chapter Review

SUMMARY

How do you get information about a population by looking at a random sample? One way is to use point estimates and confidence intervals.

- Point estimates and their corresponding parameters are

  $\bar{x}$ for μ $\quad$ $\bar{x}_1 - \bar{x}_2$ for $\mu_1 - \mu_2$

  $\hat{p}$ for p $\quad$ $\hat{p}_1 - \hat{p}_2$ for $p_1 - p_2$

- Confidence intervals are of the form

  point estimate $- E <$ parameter $<$ point estimate $+ E$

- E is the maximal margin of error. Specific values of E depend on the parameter, level of confidence, whether population standard deviations are known, sample size, and the shapes of the original population distributions.

  For μ: $E = z_c \dfrac{\sigma}{\sqrt{n}}$ when σ is known;

  $E = t_c \dfrac{s}{\sqrt{n}}$ with $d.f. = n - 1$ when σ is unknown

  For p: $E = z_c\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ when $n\hat{p} > 5$ and $n\hat{q} > 5$.

  For $\mu_1 - \mu_2$: $E = z_c\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$: when σ_1 and σ_2 are known

  $E = t_c\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ when σ_1 or σ_2 is unknown with $d.f.$ = smaller of $n_1 - 1$ or $n_2 - 1$

  Software uses Satterthwaite's approximation for $d.f.$

  For $p_1 - p_2$: $E = z_c\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}$ for sufficiently large n

- Confidence intervals have an associated probability c called the confidence level. For a given sample size, the proportion of all corresponding confidence intervals that contain the parameter in question is c.

IMPORTANT WORDS & SYMBOLS

Section 7.1

Point estimate for μ 335
Confidence level c 335
Critical values z_c 335
Maximal margin of error E 337
c confidence interval 338
Sample size for estimating μ 342

Section 7.2

Student's t distribution 347
Degrees of freedom $d.f.$ 348
Critical values t_c 349

Section 7.3

Point estimate for p, $\hat{p}$ 360
Confidence interval for p 362
Margin of error for polls 365
Sample size for estimating p 366

Section 7.4

Independent samples 373
Dependent samples 373
$\bar{x}_1 - \bar{x}_2$ sampling distribution 374
Confidence interval for $\mu_1 - \mu_2$ (σ_1 and σ_2 known) 375
Confidence interval for $\mu_1 - \mu_2$ (σ_1 and σ_2 unknown) 377
Confidence interval for $p_1 - p_2$ 381
Satterthwaite's Formula for $d.f.$ 393

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 had gone 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. (See Chapter Review Problem 19.) Diane Vaughn is a professor of sociology at Boston College and author of the book* The Challenger Launch Decision

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(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. *Statistical Literacy* In your own words, carefully explain the meanings of the following terms: *point estimate*, *critical value*, *maximal margin of error*, *confidence level*, and *confidence interval*.

2. *Critical Thinking* Suppose you are told that a 95% confidence interval for the average price of a gallon of regular gasoline in your state is from \$3.15 to \$3.45. Use the fact that the confidence interval for the mean has the form $\bar{x} - E$ to $\bar{x} + E$ to compute the sample mean and the maximal margin of error E.

3. *Critical Thinking* If you have a 99% confidence interval for μ based on a simple random sample,
   (a) is it correct to say that the *probability* that μ is in the specified interval is 99%? Explain.
   (b) is it correct to say that in the long run, if you computed many, many confidence intervals using the prescribed method, about 99% of such intervals would contain μ? Explain.

For Problems 4–19, categorize each problem according to the parameter being estimated: proportion p, mean μ, difference of means $\mu_1 - \mu_2$, or difference of proportions $p_1 - p_2$. Then solve the problem.

4. *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.

5. *Psychology: Closure* Three experiments investigating the relationship 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 a population standard deviation of $\sigma = 7.81$. 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.

6. *Psychology: Closure* How large a sample is needed in Problem 5 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?

7. *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	

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(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.

8. ***Archaeology: Pottery*** Shards of clay vessels were put together to reconstruct rim diameters of the original ceramic vessels found at the Wind Mountain archaeological site (see source in Problem 7). 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.

9. ***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 relationships 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.

10. ***Telephone Interviews: Survey*** How large a sample is needed in Problem 9 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.

11. ***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 7). 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.

12. ***Archaeology: Pottery*** Consider the three-circle, red-on-white pattern discussed in Problem 11. 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 11 as a preliminary estimate.

13. ***Agriculture: Bell Peppers*** The following data represent soil water content (percent water by volume) for independent random samples of soil taken from two experimental fields growing bell peppers (Reference: *Journal of Agricultural, Biological, and Environmental Statistics*). *Note:* These data are also available for download at the Online Study Center.

Soil water content from field I: x_1; $n_1 = 72$

15.1	11.2	10.3	10.8	16.6	8.3	9.1	12.3	9.1	14.3
10.7	16.1	10.2	15.2	8.9	9.5	9.6	11.3	14.0	11.3
15.6	11.2	13.8	9.0	8.4	8.2	12.0	13.9	11.6	16.0
9.6	11.4	8.4	8.0	14.1	10.9	13.2	13.8	14.6	10.2
11.5	13.1	14.7	12.5	10.2	11.8	11.0	12.7	10.3	10.8
11.0	12.6	10.8	9.6	11.5	10.6	11.7	10.1	9.7	9.7
11.2	9.8	10.3	11.9	9.7	11.3	10.4	12.0	11.0	10.7
8.8	11.1								

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Soil water content from field II: x_2; $n_2 = 80$

12.1	10.2	13.6	8.1	13.5	7.8	11.8	7.7	8.1	9.2
14.1	8.9	13.9	7.5	12.6	7.3	14.9	12.2	7.6	8.9
13.9	8.4	13.4	7.1	12.4	7.6	9.9	26.0	7.3	7.4
14.3	8.4	13.2	7.3	11.3	7.5	9.7	12.3	6.9	7.6
13.8	7.5	13.3	8.0	11.3	6.8	7.4	11.7	11.8	7.7
12.6	7.7	13.2	13.9	10.4	12.8	7.6	10.7	10.7	10.9
12.5	11.3	10.7	13.2	8.9	12.9	7.7	9.7	9.7	11.4
11.9	13.4	9.2	13.4	8.8	11.9	7.1	8.5	14.0	14.2

(a) Use a calculator with mean and standard deviation keys to verify that $\bar{x}_1 \approx 11.42$, $s_1 \approx 2.08$, $\bar{x}_2 \approx 10.65$, and $s_2 \approx 3.03$.
(b) Let μ_1 be the population mean for x_1 and let μ_2 be the population mean for x_2. Find a 95% confidence interval for $\mu_1 - \mu_2$.
(c) *Interpretation* Explain what the confidence interval means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 95% level of confidence, is the population mean soil water content of the first field higher than that of the second field?
(d) Which distribution (standard normal or Student's t) did you use? Why? Do you need information about the soil water content distributions?

14. *Stocks: Retail and Utility* How profitable are different sectors of the stock market? One way to answer such a question is to examine profit as a percentage of stockholder equity. A random sample of 32 retail stocks such as Toys "Я" Us, Best Buy, and Gap was studied for x_1, profit as a percentage of stockholder equity. The result was $\bar{x}_1 = 13.7$. A random sample of 34 utility (gas and electric) stocks such as Boston Edison, Wisconsin Energy, and Texas Utilities was studied for x_2, profit as a percentage of stockholder equity. The result was $\bar{x}_2 = 10.1$ (Source: *Fortune 500*, Vol. 135, No. 8). Assume that $\sigma_1 = 4.1$ and $\sigma_2 = 2.7$.
(a) Let μ_1 represent the population mean profit as a percentage of stockholder equity for retail stocks, and let μ_2 represent the population mean profit as a percentage of stockholder equity for utility stocks. Find a 95% confidence interval for $\mu_1 - \mu_2$.
(b) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 95% level of confidence, does it appear that the profit as a percentage of stockholder equity for retail stocks is higher than that for utility stocks?

15. *Wildlife: Wolves* A random sample of 18 adult male wolves from the Canadian Northwest Territories gave an average weight $\bar{x}_1 = 98$ pounds, with estimated sample standard deviation $s_1 = 6.5$ pounds. Another sample of 24 adult male wolves from Alaska gave an average weight $\bar{x}_2 = 90$ pounds, with estimated sample standard deviation $s_2 = 7.3$ pounds (Source: *The Wolf* by L. D. Mech, University of Minnesota Press).
(a) Let μ_1 represent the population mean weight of adult male wolves from the Northwest Territories, and let μ_2 represent the population mean weight of adult male wolves from Alaska. Find a 75% confidence interval for $\mu_1 - \mu_2$.
(b) *Interpretation* Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 75% level of confidence, does it appear that the average weight of adult male wolves from the Northwest Territories is greater than that of the Alaska wolves?

16. *Wildlife: Wolves* A random sample of 17 wolf litters in Ontario, Canada, gave an average of $\bar{x}_1 = 4.9$ wolf pups per litter, with estimated sample standard deviation $s_1 = 1.0$. Another random sample of 6 wolf litters in Finland gave an average of $\bar{x}_2 = 2.8$ wolf pups per litter, with sample standard deviation $s_2 = 1.2$ (see source for Problem 15).

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(a) Find an 85% confidence interval for $\mu_1 - \mu_2$, the difference in population mean litter size between Ontario and Finland.

(b) ***Interpretation*** Examine the confidence interval and explain what it means in the context of this problem. Does the interval consist of numbers that are all positive? all negative? of different signs? At the 85% level of confidence, does it appear that the average litter size of wolf pups in Ontario is greater than the average litter size in Finland?

17. ***Survey Response: Validity*** The book *Survey Responses: An Evaluation of Their Validity* by E. J. Wentland and K. Smith (Academic Press), includes studies reporting accuracy of answers to questions from surveys. A study by Locander et al. considered the question "Are you a registered voter?" Accuracy of response was confirmed by a check of city voting records. Two methods of survey were used: a face-to-face interview and a telephone interview. A random sample of 93 people were asked the voter registration question face-to-face. Seventy-nine respondents gave accurate answers (as verified by city records). Another random sample of 83 people were asked the same question during a telephone interview. Seventy-four respondents gave accurate answers. Assume the samples are representative of the general population.

(a) Let p_1 be the population proportion of all people who answer the voter registration question accurately during a face-to-face interview. Let p_2 be the population proportion of all people who answer the question accurately during a telephone interview. Find a 95% confidence interval for $p_1 - p_2$.

(b) ***Interpretation*** Does the interval contain numbers that are all positive? all negative? mixed? Comment on the meaning of the confidence interval in the context of this problem. At the 95% level, do you detect any difference in the proportion of accurate responses from face-to-face interviews compared with the proportion of accurate responses from telephone interviews?

18. ***Survey Response: Validity*** Locander et al. (see reference in Problem 17) also studied the accuracy of responses on questions involving more sensitive material than voter registration. From public records, individuals were identified as having been charged with drunken driving not less than 6 months or more than 12 months from the starting date of the study. Two random samples from this group were studied. In the first sample of 30 individuals, the respondents were asked in a face-to-face interview if they had been charged with drunken driving in the last 12 months. Of these 30 people interviewed face-to-face, 16 answered the question accurately. The second random sample consisted of 46 people who had been charged with drunken driving. During a telephone interview, 25 of these responded accurately to the question asking if they had been charged with drunken driving during the past 12 months. Assume the samples are representative of all people recently charged with drunken driving.

(a) Let p_1 represent the population proportion of all people with recent charges of drunken driving who respond accurately to a face-to-face interview asking if they have been charged with drunken driving during the past 12 months. Let p_2 represent the population proportion of people who respond accurately to the question when it is asked in a telephone interview. Find a 90% confidence interval for $p_1 - p_2$.

(b) ***Interpretation*** Does the interval found in part (a) contain numbers that are all positive? all negative? mixed? Comment on the meaning of the confidence interval in the context of this problem. At the 90% level, do you detect any differences in the proportion of accurate responses to the question from face-to-face interviews as compared with the proportion of accurate responses from telephone interviews?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

19. ***Expand Your Knowledge: Two Confidence Intervals*** What happens if we want several confidence intervals to hold at the same time (concurrently)? Do we still have the same level of confidence we had for *each* individual interval?

(a) Suppose we have two independent random variables x_1 and x_2 with respective population means μ_1 and μ_2. Let us say that we use sample data to construct two 80% confidence intervals.

Confidence Interval	Confidence Level
$A_1 < \mu_1 < B_1$	0.80
$A_2 < \mu_2 < B_2$	0.80

Now, what is the probability that *both* intervals hold at the same time? Use methods of Section 4.2 to show that

$$P(A_1 < \mu_1 < B_1 \quad and \quad A_2 < \mu_2 < B_2) = 0.64$$

Hint: You are combining independent events. If the confidence is 64% that both intervals hold concurrently, explain why the risk that at least one interval does not hold (i.e., fails) must be 36%.

(b) Suppose we want *both* intervals to hold with 90% confidence (i.e., only 10% risk level). How much confidence c should each interval have to achieve this combined level of confidence? (Assume that each interval has the same confidence level c.)

Hint: $P(A_1 < \mu_1 < B_1 \quad and \quad A_2 < \mu_2 < B_2) = 0.90$
$P(A_1 < \mu_1 < B_1) \quad \times \quad P(A_2 < \mu_2 < B_2) = 0.90$
$c \times c = 0.90$

Now solve for c.

(c) If we want *both* intervals to hold at the 90% level of confidence, then the individual intervals must hold at a *higher* level of confidence. Write a brief but detailed explanation of how this could be of importance in a large, complex engineering design such as a rocket booster or a spacecraft.

DATA HIGHLIGHTS: GROUP PROJECTS

Judy Griesedieck/Encyclopedia/Corbis

Digging clams

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

1. Garrison Bay is a small bay in Washington state. A popular recreational activity in the bay is clam digging. For several years, this 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*, Vol. 74, pp. 937–948). The data in Tables 7-4 and 7-5 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?

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

TABLE 7-4 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 7-5 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					

(d) Compute a 95% confidence interval for the population mean width of all Garrison Bay little neck clams.

(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?

(f) The *same* 35 clams were used for measures of length and width. Are the sample measurements length and width independent or dependent? Why?

2. Examine Figure 7-8, "Fall Back."

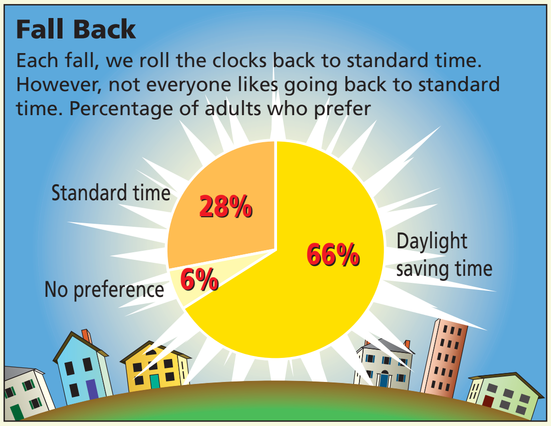
(a) Of the 1024 adults surveyed, 66% were reported to favor daylight saving time. How many people in the sample preferred daylight saving time? Using the statistic $\hat{p} = 0.66$ and sample size $n = 1024$, find a 95% confidence interval for the proportion of people p who favor daylight saving time. How could you report this information in terms of a margin of error?

(b) Look at Figure 7-8 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 7-9,"Coupons: Limited Use."

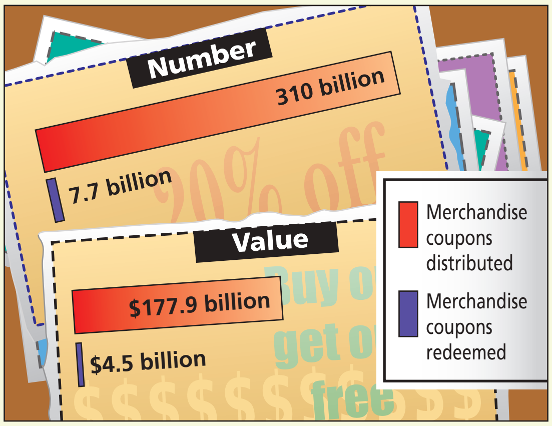
(a) Use Figure 7-9 to estimate the percentage of merchandise coupons that were redeemed. Also estimate the percentage dollar value of the coupons that were redeemed. Are these numbers approximately equal?

FIGURE 7-8



Source: Hilton Time Survey of 1024 adults

FIGURE 7-9 Coupons: Limited Use



Source: NCH Promotional Services

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) Suppose 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 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 the dollar value of a collection of coupons was $10,000. Use the data in Figure 7-9 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 the confidence interval are numerically fixed, 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 7, we have used the normal distribution, the central limit theorem, and the Student's t distribution.

   (a) Give a brief outline describing how confidence intervals for means use the normal distribution 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.

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.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(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 would be accurately represented by a confidence interval based on responses to such a question? Explain your answer.

If the question had been "Considering 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 sentences would be accurately represented by a confidence interval based on responses to such a question? Explain.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

USING TECHNOLOGY

Application 1

Finding a Confidence Interval for a Population Mean μ

Cryptanalysis, the science of breaking codes, makes extensive use of language patterns. The frequency of various letter combinations is an important part of the study. A letter combination consisting of a single letter is a monograph, while combinations consisting of two letters are called digraphs, and those with three letters are called trigraphs. In the English language, the most frequent digraph is the letter combination TH.

The *characteristic rate* of a letter combination is a measurement of its rate of occurrence. To compute the characteristic rate, count the number of occurrences of a given letter combination and divide by the number of letters in the text. For instance, to estimate the characteristic rate of the digraph TH, you could select a newspaper text and pick a random starting place. From that place, mark off 2000 letters and count the number of times that TH occurs. Then divide the number of occurrences by 2000.

The characteristic rate of a digraph can vary slightly depending on the style of the author, so to estimate an overall characteristic frequency, you want to consider several samples of newspaper text by different authors. Suppose you did this with a random sample of 15 articles and found the characteristic rates of the digraph TH in the articles. The results follow.

0.0275	0.0230	0.0300	0.0255
0.0280	0.0295	0.0265	0.0265
0.0240	0.0315	0.0250	0.0265
0.0290	0.0295	0.0275	

(a) Find a 95% confidence interval for the mean characteristic rate of the digraph TH.

(b) Repeat part (a) for a 90% confidence interval.

(c) Repeat part (a) for an 80% confidence interval.

(d) Repeat part (a) for a 70% confidence interval.

(e) Repeat part (a) for a 60% confidence interval.

(f) For each confidence interval in parts (a)–(e), compute the length of the given interval. Do you notice a relation between the confidence level and the length of the interval?

A good reference for cryptanalysis is a book by Sinkov:

Sinkov, Abraham. *Elementary Cryptanalysis*. New York: Random House.

In the book, other common digraphs and trigraphs are given.

Application 2

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/TI-*n*spire

The TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) 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.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 **Zinterval** under the **TESTS** option of the **STAT** key to generate 90% confidence intervals.

Excel 2007

On the **Home** screen, click the **Data** tab. Then in the Analysis Group, click **Data Analysis.** In the resulting dialogue box, select **Random Number Generator**. In that dialogue box, the number of variables refers to the number of samples. The number of random numbers refers to the number of data values in each sample. Select the population distribution (uniform, normal, binomial, etc.). When you click **OK** the data appear in columns on a spreadsheet, with each sample appearing in a separate column. Click on the **Insert function** (f_x). In the dialogue box, select **Statistical** for the category and then select **Confidence**. In the dialogue box for Confidence, **alpha** $= 1 - c$, so for a 90% confidence interval, enter 0.10 for alpha. Then enter the population standard deviation $\boldsymbol{\sigma}$, and the **sample size.** The resulting output gives the value of the maximal margin of error E for the confidence interval for the mean μ. 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 **Insert function** (f_x) with **Statistical** for category in the dialogue box and select **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 and difference of means. Use the menu choices **Analyze ➤ Compare Means** and then **One-Sample T Test** or **Independent-Sample T Tests** for confidence intervals for a single mean or difference of means, respectively. 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 Variable**. In the dialogue box, use Sample1 for the target variable. In the function group select **Random Numbers**. Then select the function **Rv.Normal**. Use 30 for the mean and 4 for the standard deviation. Continue until you have 10 samples. To sample

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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 μ = 30
and σ = 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

Application 3
Bootstrap Demonstration

Bootstrap can be used to construct confidence intervals for μ when traditional methods cannot be used. For example, if the sample size is small and the sample shows extreme outliers or extreme lack of symmetry, use of the Student's t distribution is inappropriate. Bootstrap makes no assumptions about the population.

Consider the following random sample of size 20:

12	15	21	2	6	3	15	51	22	18
37	12	25	19	33	15	14	17	12	27

A stem-and-leaf display shows that the data are skewed with one outlier.

0	2 represents 2
0	236
1	2224555789
2	1257
3	37
4	
5	1

We can use Minitab to model the bootstrap method for constructing confidence intervals for μ. (The Professional edition of Minitab is required because of spreadsheet size and other limitations of the Student edition.) This demonstration uses only 1000 samples. Bootstrap uses many thousands.

Step 1: Create 1000 new samples, each of size 20, by sampling *with replacement* from the original data. To do this in Minitab, we enter the original 20 data values in column C1. Then, in column C2, we place equal probabilities of 0.05 beside each of the original data values. Use the menu choices **Calc ➤ Random Data ➤ Discrete**. In the dialogue box, fill in 1000 as the number of rows, store the data in columns C11–C30, and use column C1 for values and column C2 for probabilities.

Step 2: Find the sample mean of each of the 1000 samples. To do this in Minitab, use the menu choices **Calc ➤ Row Statistics**. In the dialogue box, select **mean**. Use columns C11–C30 as the input variables and store the results in column C31.

Step 3: Order the 1000 means from smallest to largest. In Minitab, use the menu choices **Manip ➤ Sort**. In the dialogue box, indicate C31 as the column to be sorted. Store the results in column C32. Sort by values in column C31.

Step 4: Create a 95% confidence interval by finding the boundaries for the middle 95% of the data. In other words, you need to find the values of the 2.5 percentile ($P_{2.5}$) and the 97.5 percentile ($P_{97.5}$).

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Since there are 1000 data values, the 2.5 percentile is the data value in position 25, while the 97.5 percentile is the data value in position 975. The confidence interval is $P_{2.5} < \mu < P_{97.5}$.

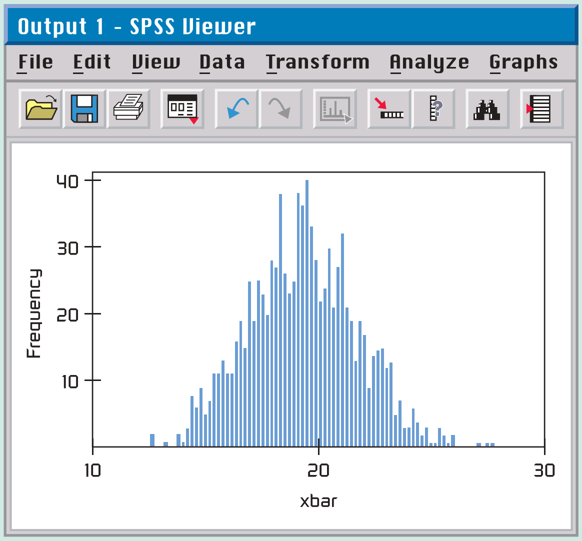
Demonstration Results

Figure 7-10 shows a histogram of the 1000 $\bar{x}$ values from one bootstrap simulation. Three bootstrap simulations produced the following 95% confidence intervals.

13.90 to 23.90
14.00 to 24.15
14.05 to 23.8

Using the t distribution on the sample data, Minitab produced the interval 13.33 to 24.27. The results of the bootstrap simulations and the t distribution method are quite close.

FIGURE 7-10 Bootstrap Simulation, $\bar{x}$ Distribution



Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.