Estimates and Sample Sizes

7-1 Review and Preview
7-2 Estimating a Population Proportion
7-3 Estimating a Population Mean: $\sigma$ Known
7-4 Estimating a Population Mean: $\sigma$ Not Known
7-5 Estimating a Population Variance
Global warming is the increase in the mean temperature of air near the surface of the earth and the increase in the mean temperature of the oceans. Scientists generally agree that global warming is caused by increased amounts of carbon dioxide, methane, ozone, and other gases that result from human activity.

Global warming is believed to be responsible for the retreat of glaciers, the reduction in the Arctic region, and a rise in sea levels. It is feared that continued global warming will result in even higher sea levels, flooding, drought, and more severe weather.

Because global warming appears to have the potential for causing dramatic changes in our environment, it is critical that we recognize that potential. Just how much do we all recognize global warming? In a Pew Research Center poll, respondents were asked “From what you’ve read and heard, is there solid evidence that the average temperature on earth has been increasing over the past few decades, or not?” In response to that question, 70% of 1501 randomly selected adults in the United States answered “yes.” Therefore, among those polled, 70% believe in global warming. Although the subject matter of this poll has great significance, we will focus on the interpretation and analysis of the results. Some important issues that relate to this poll are as follows:

- How can the poll results be used to estimate the percentage of all adults in the United States who believe that the earth is getting warmer?
- How accurate is the result of 70% likely to be?
- Given that only 1501/225,139,000, or 0.0007% of the adult population in the United States were polled, is the sample size too small to be meaningful?
- Does the method of selecting the people to be polled have much of an effect on the results?

We can answer the last question based on the sound sampling methods discussed in Chapter 1. The method of selecting the people to be polled most definitely has an effect on the results. The results are likely to be poor if a convenience sample or some other nonrandom sampling method is used. If the sample is a simple random sample, the results are likely to be good.

Our ability to understand polls and to interpret the results is crucial for our role as citizens. As we consider the topics of this chapter, we will learn more about polls and surveys and how to correctly interpret and present results.
In Chapters 2 and 3 we used “descriptive statistics” when we summarized data using tools such as graphs, and statistics such as the mean and standard deviation. We use “inferential statistics” when we use sample data to make inferences about population parameters. Two major activities of inferential statistics are (1) to use sample data to estimate values of population parameters (such as a population proportion or population mean), and (2) to test hypotheses or claims made about population parameters. In this chapter we begin working with the true core of inferential statistics as we use sample data to estimate values of population parameters. For example, the Chapter Problem refers to a poll of 1501 adults in the United States, and we see that 70% of them believe that the earth is getting warmer. Based on the sample statistic of 70%, we will estimate the percentage of all adults in the United States who believe that the earth is getting warmer. In so doing, we are using the sample results to make an inference about the population.

This chapter focuses on the use of sample data to estimate a population parameter, and Chapter 8 will introduce the basic methods for testing claims (or hypotheses) that have been made about a population parameter.

Because Sections 7-2 and 7-3 use critical values, it is helpful to review this notation introduced in Section 6-2: denotes the z score with an area of to its right. ( is the Greek letter alpha.) See Example 8 in Section 6-2, where it is shown that if , the critical value is . That is, the critical value of has an area of 0.025 to its right.

7-2 Estimating a Population Proportion

Key Concept In this section we present methods for using a sample proportion to estimate a population proportion. There are three main ideas that we should know and understand in this section.

- The sample proportion is the best point estimate of the population proportion.
- We can use a sample proportion to construct a confidence interval to estimate the true value of a population proportion, and we should know how to interpret such confidence intervals.
- We should know how to find the sample size necessary to estimate a population proportion.

The concepts presented in this section are used in the following sections and chapters, so it is important to understand this section quite well.

Proportion, Probability, and Percent Although this section focuses on the population proportion , we can also work with probabilities or percentages. In the Chapter Problem, for example, it was noted that 70% of those polled believe in global warming. The sample statistic of 70% can be expressed in decimal form as 0.70, so the sample proportion is . (Recall from Section 6-4 that represents the population proportion, and is used to denote the sample proportion.)

Point Estimate If we want to estimate a population proportion with a single value, the best estimate is the sample proportion . Because consists of a single value, it is called a point estimate.
The sample proportion \( \hat{p} \) is the best point estimate of the population proportion \( p \).

We use \( \hat{p} \) as the point estimate of \( p \) because it is unbiased and it is the most consistent of the estimators that could be used. It is unbiased in the sense that the distribution of sample proportions tends to center about the value of \( p \); that is, sample proportions \( \hat{p} \) do not systematically tend to underestimate or overestimate \( p \). (See Section 6-4.) The sample proportion \( \hat{p} \) is the most consistent estimator in the sense that the standard deviation of sample proportions tends to be smaller than the standard deviation of any other unbiased estimators.

**Example 1**

**Proportion of Adults Believing in Global Warming** In the Chapter Problem we noted that in a Pew Research Center poll, 70% of 1501 randomly selected adults in the United States believe in global warming, so the sample proportion is \( \hat{p} = 0.70 \). Find the best point estimate of the proportion of all adults in the United States who believe in global warming.

**Solution** Because the sample proportion is the best point estimate of the population proportion, we conclude that the best point estimate of \( p \) is 0.70. When using the sample results to estimate the percentage of all adults in the United States who believe in global warming, the best estimate is 70%.

**Why Do We Need Confidence Intervals?**

In Example 1 we saw that 0.70 was our best point estimate of the population proportion \( p \), but we have no indication of just how good our best estimate is. Because a point estimate has the serious flaw of not revealing anything about how good it is, statisticians have cleverly developed another type of estimate. This estimate, called a confidence interval or interval estimate, consists of a range (or an interval) of values instead of just a single value.

**Definition**

A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

A confidence interval is associated with a confidence level, such as 0.95 (or 95%). The confidence level gives us the success rate of the procedure used to construct the confidence interval. The confidence level is often expressed as the probability or area \( 1 - \alpha \) (lowercase Greek alpha), where \( \alpha \) is the complement of the confidence level. For a 0.95 (or 95%) confidence level, \( \alpha = 0.05 \). For a 0.99 (or 99%) confidence level, \( \alpha = 0.01 \).
Chapter 7

Estimates and Sample Sizes

Curbstoning

The glossary for the Census defines curbstoning as “the practice by which a census enumerator fabricates a questionnaire for a residence without actually visiting it.” Curbstoning occurs when a census enumerator sits on a curbstone (or anywhere else) and fills out survey forms by making up responses. Because data from curbstoning are not real, they can affect the validity of the Census. The extent of curbstoning has been investigated in several studies, and one study showed that about 4% of Census enumerators practiced curbstoning at least some of the time.

The methods of Section 7-2 assume that the sample data have been collected in an appropriate way, so if much of the sample data have been obtained through curbstoning, then the resulting confidence interval estimates might be very flawed.

Confidence Level

The confidence level is the probability $1 - \alpha$ (often expressed as the equivalent percentage value) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called the degree of confidence, or the confidence coefficient.)

The most common choices for the confidence level are 90% (with $\alpha = 0.10$), 95% (with $\alpha = 0.05$), and 99% (with $\alpha = 0.01$). The choice of 95% is most common because it provides a good balance between precision (as reflected in the width of the confidence interval) and reliability (as expressed by the confidence level).

Here’s an example of a confidence interval found later (in Example 3), which is based on the sample data of 1501 adults polled, with 70% of them saying that they believe in global warming:

The 0.95 (or 95%) confidence interval estimate of the population proportion $p$ is $0.677 < p < 0.723$.

It’s common for a media report to include a statement such as this: “Based on a Pew Research Center poll, the proportion of adults believing in global warming is estimated to be 70%, with a margin of error of 2 percentage points.” (We will discuss the margin of error later in this section.) Note that the confidence level is not mentioned. Although the confidence level should be given when reporting information about a poll, the media usually fail to include it.

Interpreting a Confidence Interval

We must be careful to interpret confidence intervals correctly. There is a correct interpretation and many different and creative incorrect interpretations of the confidence interval $0.677 < p < 0.723$.

**Correct:** “We are 95% confident that the interval from 0.677 to 0.723 actually does contain the true value of the population proportion $p$.” This means that if we were to select many different samples of size 1501 and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population proportion $p$. (Note that in this correct interpretation, the level of 95% refers to the success rate of the process being used to estimate the proportion.)

**Incorrect:** “There is a 95% chance that the true value of $p$ will fall between 0.677 and 0.723.” It would also be incorrect to say that “95% of sample proportions fall between 0.677 and 0.723.”

CAUTION

Know the correct interpretation of a confidence interval, as given above.

At any specific point in time, a population has a fixed and constant value $p$, and a confidence interval constructed from a sample either includes $p$ or does not. Similarly, if a baby has just been born and the doctor is about to announce its gender, it’s incorrect to say that there is a probability of 0.5 that the baby is a girl; the baby is a girl or is not, and there’s no probability involved. A population proportion $p$ is like the baby that has been born—the value of $p$ is fixed, so the confidence interval limits either contain $p$ or do not, and that is why it’s incorrect to say that there is a 95% chance that $p$ will fall between values such as 0.677 and 0.723.
A confidence level of 95% tells us that the process we are using will, in the long run, result in confidence interval limits that contain the true population proportion 95% of the time. Suppose that the true proportion of all adults who believe in global warming is \( p = 0.75 \). Then the confidence interval obtained from the Pew Research Center poll does not contain the population proportion, because the true population proportion 0.75 is not between 0.677 and 0.723. This is illustrated in Figure 7-1. Figure 7-1 shows typical confidence intervals resulting from 20 different samples. With 95% confidence, we expect that 19 out of 20 samples should result in confidence intervals that contain the true value of \( p \), and Figure 7-1 illustrates this with 19 of the confidence intervals containing \( p \), while one confidence interval does not contain \( p \).

Confidence intervals can be used informally to compare different data sets, but the overlapping of confidence intervals should not be used for making formal and final conclusions about equality of proportions. (See “On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals,” by Schenker and Gentleman, *American Statistician*, Vol. 55, No. 3.)

**Critical Values**

The methods of this section (and many of the other statistical methods found in the following chapters) include reference to a standard \( z \) score that can be used to distinguish between sample statistics that are likely to occur and those that are unlikely to occur. Such a \( z \) score is called a *critical value*. (Critical values were first presented in Section 6-2, and they are formally defined below.) Critical values are based on the following observations:

1. Under certain conditions, the sampling distribution of sample proportions can be approximated by a normal distribution, as shown in Figure 7-2.
2. A \( z \) score associated with a sample proportion has a probability of \( \alpha/2 \) of falling in the right tail of Figure 7-2.
3. The \( z \) score separating the right-tail region is commonly denoted by \( z_{\alpha/2} \), and is referred to as a *critical value* because it is on the borderline separating \( z \) scores from sample proportions that are likely to occur from those that are unlikely to occur.

**Definition**

A *critical value* is the number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur. The number \( z_{\alpha/2} \) is a critical value that is a \( z \) score with the property that it separates an area of \( \alpha/2 \) in the right tail of the standard normal distribution (as in Figure 7-2).
**Nielsen Ratings for College Students**

The Nielsen ratings are one of the most important measures of television viewing, and they affect billions of dollars in television advertising. In the past, the television viewing habits of college students were ignored, with the result that a large segment of the important young viewing audience was ignored. Nielsen Media Research is now including college students who do not live at home.

Some television shows have large appeal to viewers in the 18-24 age bracket, and the ratings of such shows have increased substantially with the inclusion of college students. For males, NBC’s Sunday Night Football broadcast had an increase of 20% after male college students were included. For females, the TV show Grey’s Anatomy had an increase of 54% after female college students were included. Those increased ratings ultimately translate into greater profits from charges to commercial sponsors. These ratings also give college students recognition that affects the programming they receive.

**Example 2**

**Finding a Critical Value** Find the critical value $z_{a/2}$ corresponding to a 95% confidence level.

**Solution**

A 95% confidence level corresponds to $\alpha = 0.05$. Figure 7-3 shows that the area in each of the red-shaded tails is $\alpha/2 = 0.025$. We find $z_{a/2} = 1.96$ by noting that the cumulative area to its left must be $1 - 0.025$, or 0.975. We can use technology or refer to Table A-2 to find that the area of 0.9750 (found in the body of the table) corresponds to $z = 1.96$. For a 95% confidence level, the critical value is therefore $z_{a/2} = 1.96$. To find the critical $z$ score for a 95% confidence level, look up 0.9750 (not 0.95) in the body of Table A-2.

**Note:** Many technologies can be used to find critical values. STATDISK, Excel, Minitab, and the TI-83/84 Plus calculator all provide critical values for the normal distribution.

Example 2 showed that a 95% confidence level results in a critical value of $z_{a/2} = 1.96$. This is the most common critical value, and it is listed with two other common values in the table that follows.

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>$\alpha$</th>
<th>Critical Value, $z_{a/2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.10</td>
<td>1.645</td>
</tr>
<tr>
<td>95%</td>
<td>0.05</td>
<td>1.96</td>
</tr>
<tr>
<td>99%</td>
<td>0.01</td>
<td>2.575</td>
</tr>
</tbody>
</table>

**Figure 7-3** Finding $z_{a/2}$ for a 95% Confidence Level

**Margin of Error**

When we collect sample data that result in a sample proportion, such as the Pew Research Center poll given in the Chapter Problem, we can calculate the sample proportion $\hat{p}$. Because of random variation in samples, the sample proportion is typically different from the population proportion. The difference between the sample proportion and the population proportion can be thought of as an error. We now define the *margin of error* $E$ as follows.
When data from a simple random sample are used to estimate a population proportion \( p \), the **margin of error**, denoted by \( E \), is the maximum likely difference (with probability \( 1 - \alpha \), such as 0.95) between the observed sample proportion \( \hat{p} \) and the true value of the population proportion \( p \). The margin of error \( E \) is also called the **maximum error of the estimate** and can be found by multiplying the critical value and the standard deviation of sample proportions, as shown in Formula 7-1.

**Formula 7-1**

\[
E = z_{\alpha/2} \sqrt{\frac{pq}{n}}
\]

margin of error for proportions

For a 95% confidence level, \( \alpha = 0.05 \), so there is a probability of 0.05 that the sample proportion will be in error by more than \( E \). This property is generalized in the following box.

**Confidence Interval for Estimating a Population Proportion \( p \)**

**Objective**

Construct a confidence interval used to estimate a population proportion.

**Notation**

\( p \) = population proportion  
\( \hat{p} \) = sample proportion  
\( n \) = number of sample values  
\( E \) = margin of error  
\( z_{\alpha/2} \) = \( z \) score separating an area of \( \alpha/2 \) in the right tail of the standard normal distribution

**Requirements**

1. The sample is a simple random sample. (*Caution:* If the sample data have been obtained in a way that is not appropriate, the estimates of the population proportion may be very wrong.)
2. The conditions for the binomial distribution are satisfied. That is, there is a fixed number of trials, the trials are independent, there are two categories of outcomes, and the probabilities remain constant for each trial. (See Section 5-3.)
3. There are at least 5 successes and at least 5 failures. (With the population proportions \( p \) and \( q \) unknown, we estimate their values using the sample proportion, so this requirement is a way of verifying that \( np \geq 5 \) and \( nq \geq 5 \) are both satisfied, so the normal distribution is a suitable approximation to the binomial distribution. There are procedures for dealing with situations in which the normal distribution is not a suitable approximation, as in Exercise 51.)

**Confidence Interval**

\[
\hat{p} - E < p < \hat{p} + E \quad \text{where} \quad E = z_{\alpha/2} \sqrt{\frac{pq}{n}}
\]

The confidence interval is often expressed in the following equivalent formats:

\[
\hat{p} \pm E
\]

or

\[
(\hat{p} - E, \hat{p} + E)
\]
Chapter 7  Estimates and Sample Sizes

In Chapter 4, when probabilities were given in decimal form, we rounded to three significant digits. We use that same rounding rule here.

**Round-Off Rule for Confidence Interval Estimates of \( p \)**
Round the confidence interval limits for \( p \) to three significant digits.

We now summarize the procedure for constructing a confidence interval estimate of a population proportion \( p \):

**Procedure for Constructing a Confidence Interval for \( p \)**

1. Verify that the requirements are satisfied.
2. Refer to Table A-2 or use technology to find the critical value \( z_{a/2} \) that corresponds to the desired confidence level.
3. Evaluate the margin of error \( E = z_{a/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} \).
4. Using the value of the calculated margin of error \( E \) and the value of the sample proportion \( \hat{p} \), find the values of the confidence interval limits \( \hat{p} - E \) and \( \hat{p} + E \). Substitute those values in the general format for the confidence interval:

   \[
   \hat{p} - E < p < \hat{p} + E
   \]

   or

   \[
   \hat{p} \pm E
   \]

   or

   \[
   (\hat{p} - E, \hat{p} + E)
   \]

5. Round the resulting confidence interval limits to three significant digits.

**Example 3  Constructing a Confidence Interval: Poll Results**

In the Chapter Problem we noted that a Pew Research Center poll of 1501 randomly selected U.S. adults showed that 70% of the respondents believe in global warming. The sample results are \( n = 1501 \), and \( \hat{p} = 0.70 \).

a. Find the margin of error \( E \) that corresponds to a 95% confidence level.

b. Find the 95% confidence interval estimate of the population proportion \( p \).

c. Based on the results, can we safely conclude that the majority of adults believe in global warming?

d. Assuming that you are a newspaper reporter, write a brief statement that accurately describes the results and includes all of the relevant information.

**Solution**

**Requirement Check** We first verify that the necessary requirements are satisfied. (1) The polling methods used by the Pew Research Center result in samples that can be considered to be simple random samples. (2) The conditions for a binomial experiment are satisfied, because there is a fixed number of trials (1501), the trials are independent (because the response from one person doesn’t affect the probability of the response from another person), there are two categories of outcome (subject believes in global warming or does not), and the probability remains constant. Also, with 70% of the respondents believing in global warming,
the number who believe is 1051 (or 70% of 1501) and the number who do not believe is 450, so the number of successes (1051) and the number of failures (450) are both at least 5. The check of requirements has been successfully completed.

a. The margin of error is found by using Formula 7-1 with \( z_{a/2} = 1.96 \) (as found in Example 2), \( \hat{p} = 0.70 \), \( \hat{q} = 0.30 \), and \( n = 1501 \).

\[
E = z_{a/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} = 1.96 \sqrt{\frac{(0.70)(0.30)}{1501}} = 0.023183
\]

b. Constructing the confidence interval is quite easy now that we know the values of \( \hat{p} \) and \( E \). We simply substitute those values to obtain this result:

\[
\hat{p} - E < \mu_p < \hat{p} + E
\]

\[
0.70 - 0.023183 < \mu_p < 0.70 + 0.023183
\]

\[
0.677 < \mu_p < 0.723 \quad \text{(rounded to three significant digits)}
\]

This same result could be expressed in the format of \( 0.70 \pm 0.023 \) or (0.677, 0.723). If we want the 95% confidence interval for the true population percentage, we could express the result as 67.7% < \( \mu_p \) < 72.3%.

c. Based on the confidence interval obtained in part (b), it does appear that the proportion of adults who believe in global warming is greater than 0.5 (or 50%), so we can safely conclude that the majority of adults believe in global warming. Because the limits of 0.677 and 0.723 are likely to contain the true population proportion, it appears that the population proportion is a value greater than 0.5.

d. Here is one statement that summarizes the results: 70% of United States adults believe that the earth is getting warmer. That percentage is based on a Pew Research Center poll of 1501 randomly selected adults in the United States. In theory, in 95% of such polls, the percentage should differ by no more than 2.3 percentage points in either direction from the percentage that would be found by interviewing all adults in the United States.

**Analyzing Polls** Example 3 addresses the poll described in the Chapter Problem. When analyzing results from polls, we should consider the following.

1. The sample should be a simple random sample, not an inappropriate sample (such as a voluntary response sample).

2. The confidence level should be provided. (It is often 95%, but media reports often neglect to identify it.)

3. The sample size should be provided. (It is usually provided by the media, but not always.)

4. Except for relatively rare cases, the quality of the poll results depends on the sampling method and the size of the sample, but the size of the population is usually not a factor.

**CAUTION**

Never follow the common misconception that poll results are unreliable if the sample size is a small percentage of the population size. The population size is usually not a factor in determining the reliability of a poll.
Determining Sample Size

Suppose we want to collect sample data in order to estimate some population proportion. How do we know how many sample items must be obtained? If we solve the formula for the margin of error $E$ (Formula 7-1) for $n$, we get Formula 7-2. Formula 7-2 requires $\hat{p}$ as an estimate of the population proportion $p$, but if no such estimate is known (as is often the case), we replace $\hat{p}$ by 0.5 and replace $\hat{q}$ by 0.5, with the result given in Formula 7-3.

### Finding the Sample Size Required to Estimate a Population Proportion

**Objective**

Determine how large the sample should be in order to estimate the population proportion $p$.

**Notation**

- $p = \text{population proportion}$
- $\hat{p} = \text{sample proportion}$
- $n = \text{number of sample values}$
- $E = \text{desired margin of error}$
- $z_{\alpha/2} = z$ score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

**Requirements**

The sample must be a simple random sample of independent subjects.

When an estimate $\hat{p}$ is known: **Formula 7-2**

$$n = \frac{\left[z_{\alpha/2}\right]^2 \hat{p} \hat{q}}{E^2}$$

When no estimate $\hat{p}$ is known: **Formula 7-3**

$$n = \frac{\left[z_{\alpha/2}\right]^2 0.25}{E^2}$$

If reasonable estimates of $\hat{p}$ can be made by using previous samples, a pilot study, or someone’s expert knowledge, use Formula 7-2. If nothing is known about the value of $\hat{p}$, use Formula 7-3.

Formulas 7-2 and 7-3 are remarkable because they show that the sample size does not depend on the size ($N$) of the population; the sample size depends on the desired confidence level, the desired margin of error, and sometimes the known estimate of $\hat{p}$. (See Exercise 49 for dealing with cases in which a relatively large sample is selected without replacement from a finite population.)

### Round-Off Rule for Determining Sample Size

If the computed sample size $n$ is not a whole number, round the value of $n$ up to the next larger whole number.

#### Example 4

**How Many Adults Use the Internet?** The Internet is affecting us all in many different ways, so there are many reasons for estimating the proportion of adults who use it. Assume that a manager for E-Bay wants to determine the current percentage of U.S. adults who now use the Internet. How many adults must be surveyed in order to be 95% confident that the sample percentage is in error by no more than three percentage points?
7-2 Estimating a Population Proportion

a. Use this result from a Pew Research Center poll: In 2006, 73% of U.S. adults used the Internet.

b. Assume that we have no prior information suggesting a possible value of the proportion.

**Solution**

a. The prior study suggests that \( \hat{p} = 0.73 \), so \( \hat{q} = 0.27 \) (found from \( \hat{q} = 1 - 0.73 \)).

With a 95% confidence level, we have \( \alpha = 0.05 \), so \( z_{\alpha/2} = 1.96 \). Also, the margin of error is \( E = 0.03 \) (the decimal equivalent of “three percentage points”).

Because we have an estimated value of \( \hat{p} \) we use Formula 7-2 as follows:

\[
\begin{align*}
 n &= \frac{[z_{\alpha/2}]^2 \hat{p} \hat{q}}{E^2} \\
 &= \frac{[1.96]^2 (0.73)(0.27)}{0.03^2} \\
 &= 841.3104 = 842 \quad \text{(rounded up)}
\end{align*}
\]

We must obtain a simple random sample that includes at least 842 adults.

b. As in part (a), we again use \( z_{\alpha/2} = 1.96 \) and \( E = 0.03 \), but with no prior knowledge of \( \hat{p} \) (or \( \hat{q} \)), we use Formula 7-3 as follows:

\[
\begin{align*}
 n &= \frac{[z_{\alpha/2}]^2 \cdot 0.25}{E^2} \\
 &= \frac{[1.96]^2 \cdot 0.25}{0.03^2} \\
 &= 1067.1111 = 1068 \quad \text{(rounded up)}
\end{align*}
\]

**Interpretation**

To be 95% confident that our sample percentage is within three percentage points of the true percentage for all adults, we should obtain a simple random sample of 1068 adults. By comparing this result to the sample size of 842 found in part (a), we can see that if we have no knowledge of a prior study, a larger sample is required to achieve the same results as when the value of \( \hat{p} \) can be estimated.

**Caution**

Try to avoid these two common errors when calculating sample size:

1. Don’t make the mistake of using \( E = 3 \) as the margin of error corresponding to “three percentage points.”

2. Be sure to substitute the critical \( z \) score for \( z_{\alpha/2} \). For example, if you are working with 95% confidence, be sure to replace \( z_{\alpha/2} \) with 1.96. Don’t make the mistake of replacing \( z_{\alpha/2} \) with 0.95 or 0.05.

**Finding the Point Estimate and \( E \) from a Confidence Interval**

Sometimes we want to better understand a confidence interval that might have been obtained from a journal article, or generated using computer software or a calculator. If we already know the confidence interval limits, the sample proportion (or the best point estimate) \( \hat{p} \) and the margin of error \( E \) can be found as follows:

Point estimate of \( p \):

\[
\hat{p} = \frac{\text{upper confidence interval limit} + \text{lower confidence interval limit}}{2}
\]
Chapter 7  Estimates and Sample Sizes

Margin of error:
\[ E = \frac{(\text{upper confidence interval limit}) - (\text{lower confidence interval limit})}{2} \]

**Example 5**  The article “High-Dose Nicotine Patch Therapy,” by Dale, Hurt, et al. (Journal of the American Medical Association, Vol. 274, No. 17) includes this statement: “Of the 71 subjects, 70% were abstinent from smoking at 8 weeks (95% confidence interval [CI], 58% to 81%).” Use that statement to find the point estimate \( \hat{p} \) and the margin of error \( E \).

**Solution**  From the given statement, we see that the 95% confidence interval is \( 0.58 < p < 0.81 \). The point estimate \( \hat{p} \) is the value midway between the upper and lower confidence interval limits, so we get

\[ \hat{p} = \frac{(\text{upper confidence limit}) + (\text{lower confidence limit})}{2} = \frac{0.81 + 0.58}{2} = 0.695 \]

The margin of error can be found as follows:

\[ E = \frac{(\text{upper confidence limit}) - (\text{lower confidence limit})}{2} = \frac{0.81 - 0.58}{2} = 0.115 \]

**Better-Performing Confidence Intervals**

*Important note:* The exercises for this section are based on the method for constructing a confidence interval as described above, not the confidence intervals described in the following discussion.

The confidence interval described in this section has the format typically presented in introductory statistics courses, but it does not perform as well as some other confidence intervals. The **adjusted Wald confidence interval** performs better in the sense that its probability of containing the true population proportion \( p \) is closer to the confidence level that is used. The adjusted Wald confidence interval uses this simple procedure: Add 2 to the number of successes \( x \), add 2 to the number of failures (so that the number of trials \( n \) is increased by 4), then find the confidence interval as described in this section. For example, if we use the methods of this section with \( x = 10 \) and \( n = 20 \), we get this 95% confidence interval: \( 0.281 < p < 0.719 \). With \( x = 10 \) and \( n = 20 \) we use the adjusted Wald confidence interval by letting \( x = 12 \) and \( n = 24 \) to get this confidence interval: \( 0.300 < p < 0.700 \). The chance that the confidence interval \( 0.300 < p < 0.700 \) contains \( p \) is closer to 95% than the chance that \( 0.281 < p < 0.719 \) contains \( p \).
Another confidence interval that performs better than the one described in this section and the adjusted Wald confidence interval is the Wilson score confidence interval:

$$\hat{p} + \frac{z^2}{2n} \pm \frac{z^2 \hat{q} + z^2}{4n} \sqrt{\frac{\hat{q}}{n}}$$

(It is easy to see why this approach is not used much in introductory courses.) Using $x = 10$ and $n = 20$, the 95% Wilson score confidence interval is $0.299 < p < 0.701$.

For a discussion of these and other confidence intervals for $p$, see "Approximation Is Better than 'Exact' for Interval Estimation of Binomial Proportions," by Agresti and Coull, American Statistician, Vol. 52, No. 2.

### For Confidence Intervals

**STATDISK** Select Analysis, then Confidence Intervals, then Proportion One Sample, and proceed to enter the requested items. The confidence interval will be displayed.

**MINITAB** Select Stat, Basic Statistics, then 1 Proportion. In the dialog box, click on the button for Summarized Data. Also click on the Options button, enter the desired confidence level (the default is 95%). Instead of using a normal approximation, Minitab's default procedure is to determine the confidence interval limits by using an exact method. To use the normal approximation method presented in this section, click on the Options button and then click on the box with this statement: “Use test and interval based on normal distribution.”

**EXCEL** Use the Data Desk XL add-in that is a supplement to this book. First enter the number of successes in cell A1, then enter the total number of trials in cell B1. Select DDXL, select Confidence Intervals, then select Summ 1 Var Prop Interval (which is an abbreviated form of “confidence interval for a proportion using summary data for one variable”). Click on the pencil icon for “Num successes” and enter A1. Click on the pencil icon for “Num trials” and enter B1. Click OK. In the dialog box, select the level of confidence, then click on Compute Interval.

**TI-83/84 PLUS** Press STAT, select TESTS, then select 1-PropZInt and enter the required items. The accompanying display shows the result for Example 3. Like many technologies, the confidence interval limits are expressed in the format shown on the second line of the display.

### For Sample Size Determination

**STATDISK** Select Analysis, then Sample Size Determination, then Estimate Proportion. Enter the required items in the dialog box.

Sample size determination is not available as a built-in function with Minitab, Excel, or the TI-83/84 Plus calculator.

### 7-2 Basic Skills and Concepts

**Statistical Literacy and Critical Thinking**

1. **Poll Results in the Media** USA Today provided a “snapshot” illustrating poll results from 21,944 subjects. The illustration showed that 43% answered “yes” to this question: “Would you rather have a boring job than no job?” The margin of error was given as ±1 percentage point. What important feature of the poll was omitted?
2. **Margin of Error** For the poll described in Exercise 1, describe what is meant by the statement that “the margin of error is ±1 percentage point.”

3. **Confidence Interval** For the poll described in Exercise 1, we see that 43% of 21,944 people polled answered “yes” to the given question. Given that 43% is the best estimate of the population percentage, why would we need a confidence interval? That is, what additional information does the confidence interval provide?

4. **Sampling** Suppose the poll results from Exercise 1 were obtained by mailing 100,000 questionnaires and receiving 21,944 responses. Is the result of 43% a good estimate of the population percentage of “yes” responses? Why or why not?

**Finding Critical Values. In Exercises 5–8, find the indicated critical z value.**

5. Find the critical value $z_{0.05/2}$ that corresponds to a 99% confidence level.
6. Find the critical value $z_{0.005/2}$ that corresponds to a 99.5% confidence level.
7. Find $z_{0.10}$ for $\alpha = 0.10$.
8. Find $z_{0.05}$ for $\alpha = 0.02$.

**Expressing Confidence Intervals. In Exercises 9–12, express the confidence interval using the indicated format.**

9. Express the confidence interval $0.200 < p < 0.500$ in the form of $\hat{p} \pm E$.
10. Express the confidence interval $0.720 < p < 0.780$ in the form of $\hat{p} \pm E$.
11. Express the confidence interval $(0.437, 0.529)$ in the form of $\hat{p} \pm E$.
12. Express the confidence interval $0.222 \pm 0.044$ in the form of $\hat{p} - E < p < \hat{p} + E$.

**Interpreting Confidence Interval Limits. In Exercises 13–16, use the given confidence interval limits to find the point estimate $\hat{p}$ and the margin of error $E$.**

13. $(0.320, 0.420) \quad 14. 0.772 < p < 0.776$.
15. $0.433 < p < 0.527 \quad 16. 0.102 < p < 0.236$.

**Finding Margin of Error. In Exercises 17–20, assume that a sample is used to estimate a population proportion $p$. Find the margin of error $E$ that corresponds to the given statistics and confidence level.**

17. $n = 1000, x = 400, 95\%$ confidence
18. $n = 500, x = 220, 99\%$ confidence
19. 98\% confidence; the sample size is 1230, of which 40\% are successes.
20. 90\% confidence; the sample size is 1780, of which 35\% are successes.

**Constructing Confidence Intervals. In Exercises 21–24, use the sample data and confidence level to construct the confidence interval estimate of the population proportion $p$.**

21. $n = 200, x = 40, 95\%$ confidence
22. $n = 2000, x = 400, 95\%$ confidence
23. $n = 1236, x = 109, 99\%$ confidence
24. $n = 5200, x = 4821, 99\%$ confidence

**Determining Sample Size. In Exercises 25–28, use the given data to find the minimum sample size required to estimate a population proportion or percentage.**

25. Margin of error: 0.045; confidence level: 95%; $\hat{p}$ and $\hat{q}$ unknown
26. Margin of error: 0.005; confidence level: 99%; $\hat{p}$ and $\hat{q}$ unknown
27. Margin of error: two percentage points; confidence level: 99%; from a prior study, $\hat{p}$ is estimated by the decimal equivalent of 14%.
28. Margin of error: three percentage points; confidence level: 95%; from a prior study, $\hat{p}$ is estimated by the decimal equivalent of 87%.
29. Gender Selection The Genetics and IVF Institute conducted a clinical trial of the XSORT method designed to increase the probability of conceiving a girl. As of this writing, 574 babies were born to parents using the XSORT method, and 525 of them were girls.

a. What is the best point estimate of the population proportion of girls born to parents using the XSORT method?

b. Use the sample data to construct a 95% confidence interval estimate of the percentage of girls born to parents using the XSORT method.

c. Based on the results, does the XSORT method appear to be effective? Why or why not?

30. Gender Selection The Genetics and IVF Institute conducted a clinical trial of the YSORT method designed to increase the probability of conceiving a boy. As of this writing, 152 babies were born to parents using the YSORT method, and 127 of them were boys.

a. What is the best point estimate of the population proportion of boys born to parents using the YSORT method?

b. Use the sample data to construct a 99% confidence interval estimate of the percentage of boys born to parents using the YSORT method.

c. Based on the results, does the YSORT method appear to be effective? Why or why not?

31. Postponing Death An interesting and popular hypothesis is that individuals can temporarily postpone their death to survive a major holiday or important event such as a birthday. In a study of this phenomenon, it was found that in the week before and the week after Thanksgiving, there were 12,000 total deaths, and 6062 of them occurred in the week before Thanksgiving (based on data from "Holidays, Birthdays, and Postponement of Cancer Death," by Young and Hade, Journal of the American Medical Association, Vol. 292, No. 24.)

a. What is the best point estimate of the proportion of deaths in the week before Thanksgiving to the total deaths in the week before and the week after Thanksgiving?

b. Construct a 95% confidence interval estimate of the proportion of deaths in the week before Thanksgiving to the total deaths in the week before and the week after Thanksgiving.

c. Based on the result, does there appear to be any indication that people can temporarily postpone their death to survive the Thanksgiving holiday? Why or why not?

32. Medical Malpractice An important issue facing Americans is the large number of medical malpractice lawsuits and the expenses that they generate. In a study of 1228 randomly selected medical malpractice lawsuits, it is found that 856 of them were later dropped or dismissed (based on data from the Physician Insurers Association of America).

a. What is the best point estimate of the proportion of medical malpractice lawsuits that are dropped or dismissed?

b. Construct a 99% confidence interval estimate of the proportion of medical malpractice lawsuits that are dropped or dismissed.

c. Does it appear that the majority of such suits are dropped or dismissed?

33. Mendelian Genetics When Mendel conducted his famous genetics experiments with peas, one sample of offspring consisted of 428 green peas and 152 yellow peas.

a. Find a 95% confidence interval estimate of the percentage of yellow peas.

b. Based on his theory of genetics, Mendel expected that 25% of the offspring peas would be yellow. Given that the percentage of offspring yellow peas is not 25%, do the results contradict Mendel’s theory? Why or why not?

34. Misleading Survey Responses In a survey of 1002 people, 701 said that they voted in a recent presidential election (based on data from ICR Research Group). Voting records show that 61% of eligible voters actually did vote.

a. Find a 99% confidence interval estimate of the proportion of people who say that they voted.

b. Are the survey results consistent with the actual voter turnout of 61%? Why or why not?
35. **Cell Phones and Cancer** A study of 420,095 Danish cell phone users found that 135 of them developed cancer of the brain or nervous system. Prior to this study of cell phone use, the rate of such cancer was found to be 0.0340% for those not using cell phones. The data are from the *Journal of the National Cancer Institute*.

   a. Use the sample data to construct a 95% confidence interval estimate of the percentage of cell phone users who develop cancer of the brain or nervous system.

   b. Do cell phone users appear to have a rate of cancer of the brain or nervous system that is different from the rate of such cancer among those not using cell phones? Why or why not?

36. **Global Warming Poll** A Pew Research Center poll included 1708 randomly selected adults who were asked whether "global warming is a problem that requires immediate government action." Results showed that 939 of those surveyed indicated that immediate government action is required. A news reporter wants to determine whether these survey results constitute strong evidence that the majority (more than 50%) of people believe that immediate government action is required.

   a. What is the best estimate of the percentage of adults who believe that immediate government action is required?

   b. Construct a 99% confidence interval estimate of the proportion of adults believing that immediate government action is required.

   c. Is there strong evidence supporting the claim that the majority is in favor of immediate government action? Why or why not?

37. **Internet Use** In a Pew Research Center poll, 73% of 3011 adults surveyed said that they use the Internet. Construct a 95% confidence interval estimate of the proportion of all adults who use the Internet. Is it correct for a newspaper reporter to write that "3/4 of all adults use the Internet"? Why or why not?

38. **Job Interview Mistakes** In an Accountemps survey of 150 senior executives, 47% said that the most common job interview mistake is to have little or no knowledge of the company. Construct a 99% confidence interval estimate of the proportion of all senior executives who have that same opinion. Is it possible that exactly half of all senior executives believe that the most common job interview mistake is to have little or no knowledge of the company? Why or why not?

39. **AOL Poll** After 276 passengers on the Queen Elizabeth II cruise ship contracted a norovirus, America Online presented this question on its Internet site: "Would the recent outbreak deter you from taking a cruise?" Among the 34,358 people who responded, 62% answered "yes." Use the sample data to construct a 95% confidence interval estimate of the population of all people who would respond "yes" to that question. Does the confidence interval provide a good estimate of the population proportion? Why or why not?

40. **Touch Therapy** When she was nine years of age, Emily Rosa did a science fair experiment in which she tested professional touch therapists to see if they could sense her energy field. She flipped a coin to select either her right hand or her left hand, then she asked the therapists to identify the selected hand by placing their hand just under Emily's hand without seeing it and without touching it. Among 280 trials, the touch therapists were correct 123 times (based on data in "A Close Look at Therapeutic Touch," *Journal of the American Medical Association*, Vol. 279, No. 13).

   a. Given that Emily used a coin toss to select either her right hand or her left hand, what proportion of correct responses would be expected if the touch therapists made random guesses?

   b. Using Emily's sample results, what is the best point estimate of the therapist's success rate?

   c. Using Emily's sample results, construct a 99% confidence interval estimate of the proportion of correct responses made by touch therapists.

   d. What do the results suggest about the ability of touch therapists to select the correct hand by sensing an energy field?
Determining Sample Size. In Exercises 41–44, find the minimum sample size required to estimate a population proportion or percentage.

41. Internet Use The use of the Internet is constantly growing. How many randomly selected adults must be surveyed to estimate the percentage of adults in the United States who now use the Internet? Assume that we want to be 99% confident that the sample percentage is within two percentage points of the true population percentage.
   a. Assume that nothing is known about the percentage of adults using the Internet.
   b. As of this writing, it was estimated that 73% of adults in the United States use the Internet (based on a Pew Research Center poll).

42. Cell Phones As the newly hired manager of a company that provides cell phone service, you want to determine the percentage of adults in your state who live in a household with cell phones and no land-line phones. How many adults must you survey? Assume that you want to be 90% confident that the sample percentage is within four percentage points of the true population percentage.
   a. Assume that nothing is known about the percentage of adults who live in a household with cell phones and no land-line phones.
   b. Assume that a recent survey suggests that about 8% of adults live in a household with cell phones and no land-line phones (based on data from the National Health Interview Survey).

43. Nitrogen in Tires A campaign was designed to convince car owners that they should fill their tires with nitrogen instead of air. At a cost of about $5 per tire, nitrogen supposedly has the advantage of leaking at a much slower rate than air, so that the ideal tire pressure can be maintained more consistently. Before spending huge sums to advertise the nitrogen, it would be wise to conduct a survey to determine the percentage of car owners who would pay for the nitrogen. How many randomly selected car owners should be surveyed? Assume that we want to be 95% confident that the sample percentage is within three percentage points of the true percentage of all car owners who would be willing to pay for the nitrogen.

44. Name Recognition As this book was being written, former New York City mayor Rudolph Giuliani announced that he was a candidate for the presidency of the United States. If you are a campaign worker and need to determine the percentage of people that recognize his name, how many people must you survey to estimate that percentage? Assume that you want to be 95% confident that the sample percentage is in error by no more than two percentage points, and also assume that a recent survey indicates that Giuliani’s name is recognized by 10% of all adults (based on data from a Gallup poll).

Using Appendix B Data Sets. In Exercises 45–48, use the indicated data set from Appendix B.

45. Green M&M Candies Refer to Data Set 18 in Appendix B and find the sample proportion of M&Ms that are green. Use that result to construct a 95% confidence interval estimate of the population percentage of M&Ms that are green. Is the result consistent with the 16% rate that is reported by the candy maker Mars? Why or why not?

46. Freshman 15 Weight Gain Refer to Data Set 3 in Appendix B.
   a. Based on the sample results, find the best point estimate of the percentage of college students who gain weight in their freshman year.
   b. Construct a 95% confidence interval estimate of the percentage of college students who gain weight in their freshman year.
   c. Assuming that you are a newspaper reporter, write a statement that describes the results. Include all of the relevant information. (Hint: See Example 3 part (d).)

47. Precipitation in Boston Refer to Data Set 14 in Appendix B, and consider days with precipitation values different from 0 to be days with precipitation. Construct a 95% confidence interval estimate of the proportion of Wednesdays with precipitation, and also construct a 95% confidence interval estimate of the proportion of Sundays with precipitation. Compare the results. Does precipitation appear to occur more on either day?
48. Movie Ratings Refer to Data Set 9 in Appendix B and find the proportion of movies with R ratings. Use that proportion to construct a 95% confidence interval estimate of the proportion of all movies with R ratings. Assuming that the listed movies constitute a simple random sample of all movies, can we conclude that most movies have ratings different from R? Why or why not?

7-2 Beyond the Basics

49. Using Finite Population Correction Factor In this section we presented Formulas 7-2 and 7-3, which are used for determining sample size. In both cases we assumed that the population is infinite or very large and that we are sampling with replacement. When we have a relatively small population with size \( N \) and sample without replacement, we modify \( E \) to include the finite population correction factor shown here, and we can solve for \( n \) to obtain the result given here. Use this result to repeat Exercise 43, assuming that we limit our population to the 12,784 car owners living in LaGrange, New York, home of the author. Is the sample size much lower than the sample size required for a population of millions of people?

\[
E = \frac{z_{\alpha/2} \sqrt{pq}}{n} \sqrt{\frac{N-n}{N-1}} \quad n = \frac{Npq (z_{\alpha/2})^2}{\frac{z^2}{4} (z_{\alpha/2})^2 + (N-1)E^2}
\]

50. One-Sided Confidence Interval A one-sided confidence interval for \( p \) can be expressed as \( p < \hat{p} + E \) or \( p > \hat{p} - E \), where the margin of error \( E \) is modified by replacing \( z_{\alpha/2} \) with \( z_{\alpha} \). If Air America wants to report an on-time performance of at least \( x \) percent with 95% confidence, contruct the appropriate one-sided confidence interval and then find the percent in question. Assume that a simple random sample of 750 flights results in 630 that are on time.

51. Confidence Interval from Small Sample Special tables are available for finding confidence intervals for proportions involving small numbers of cases, where the normal distribution approximation cannot be used. For example, given \( x = 3 \) successes among \( n = 8 \) trials, the 95% confidence interval found in Standard Probability and Statistics Tables and Formulae (CRC Press) is \( 0.085 < p < 0.755 \). Find the confidence interval that would result if you were to incorrectly use the normal distribution as an approximation to the binomial distribution. Are the results reasonably close?

52. Interpreting Confidence Interval Limits Assume that a coin is modified so that it favors heads, and 100 tosses result in 95 heads. Find the 99% confidence interval estimate of the proportion of heads that will occur with this coin. What is unusual about the results obtained by the methods of this section? Does common sense suggest a modification of the resulting confidence interval?

53. Rule of Three Suppose \( n \) trials of a binomial experiment result in no successes. According to the Rule of Three, we have 95% confidence that the true population proportion has an upper bound of \( \frac{3}{n} \). (See "A Look at the Rule of Three," by Jovanovic and Levy, American Statistician, Vol. 51, No. 2.)

a. If \( n \) independent trials result in no successes, why can’t we find confidence interval limits by using the methods described in this section?

b. If 20 patients are treated with a drug and there are no adverse reactions, what is the 95% upper bound for \( p \), the proportion of all patients who experience adverse reactions to this drug?

54. Poll Accuracy A New York Times article about poll results states, "In theory, in 19 cases out of 20, the results from such a poll should differ by no more than one percentage point in either direction from what would have been obtained by interviewing all voters in the United States." Find the sample size suggested by this statement.
Key Concept In this section we present methods for estimating a population mean. In addition to knowing the values of the sample data or statistics, we must also know the value of the population standard deviation, $\sigma$. Here are three key concepts that should be learned in this section.

1. We should know that the sample mean $\bar{x}$ is the best point estimate of the population mean $\mu$.

2. We should learn how to use sample data to construct a confidence interval for estimating the value of a population mean, and we should know how to interpret such confidence intervals.

3. We should develop the ability to determine the sample size necessary to estimate a population mean.

Important: The confidence interval described in this section has the requirement that we know the value of the population standard deviation $\sigma$, but that value is rarely known in real circumstances. Section 7-4 describes methods for dealing with realistic cases in which $\sigma$ is not known.

Point Estimate In Section 7-2 we saw that the sample proportion $\hat{p}$ is the best point estimate of the population proportion $p$. The sample mean $\bar{x}$ is an unbiased estimator of the population mean $\mu$, and for many populations, sample means tend to vary less than other measures of center, so the sample mean $\bar{x}$ is usually the best point estimate of the population mean $\mu$.

The sample mean $\bar{x}$ is the best point estimate of the population mean.

Although the sample mean $\bar{x}$ is usually the best point estimate of the population mean $\mu$, it does not give us any indication of just how good our best estimate is. We get more information from a confidence interval (or interval estimate), which consists of a range (or an interval) of values instead of just a single value.

Knowledge of $\sigma$ The listed requirements on the next page include knowledge of the population standard deviation $\sigma$, but Section 7-4 presents methods for estimating a population mean without knowledge of the value of $\sigma$.

Normality Requirement The requirements on the next page include the property that either the population is normally distributed or $n > 30$. If $n \leq 30$, the population need not have a distribution that is exactly normal. The methods of this section are robust against departures from normality, which means that these methods are not strongly affected by departures from normality, provided that those departures are not too extreme. We therefore have a loose normality requirement that can be satisfied if there are no outliers and if a histogram of the sample data is not dramatically different from being bell-shaped. (See Section 6-7.)

Sample Size Requirement The normal distribution is used as the distribution of sample means. If the original population is not itself normally distributed, then we say that means of samples with size $n > 30$ have a distribution that can be approximated by a normal distribution. The condition $n > 30$ is a common guideline, but there is no specific minimum sample size that works for all cases.
The minimum sample size actually depends on how much the population distribution departs from a normal distribution. Sample sizes of 15 to 30 are sufficient if the population has a distribution that is not far from normal, but some other populations have distributions that are extremely far from normal and sample sizes greater than 30 might be necessary. In this book we use the simplified criterion of $n > 30$ as justification for treating the distribution of sample means as a normal distribution.

**Confidence Level** The confidence interval is associated with a confidence level, such as 0.95 (or 95%). The confidence level gives us the success rate of the procedure used to construct the confidence interval. As in Section 7-2, $\alpha$ is the complement of the confidence level. For a 0.95 (or 95%) confidence level, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$.

### Confidence Interval for Estimating a Population Mean (with $\sigma$ Known)

**Objective**
Construct a confidence interval used to estimate a population mean.

**Notation**
- $\mu$ = population mean
- $\sigma$ = population standard deviation
- $\bar{x}$ = sample mean
- $n$ = number of sample values
- $E$ = margin of error
- $z_{\alpha/2}$ = $z$ score separating an area of $\alpha/2$ in the right tail of the standard normal distribution

**Requirements**
1. The sample is a simple random sample.
2. The value of the population standard deviation $\sigma$ is known.
3. Either or both of these conditions is satisfied: The population is normally distributed or $n > 30$.

**Confidence Interval**

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

where

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} \pm E$$

or

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

**Procedure for Constructing a Confidence Interval for $\mu$ (with Known $\sigma$)**

1. Verify that the requirements are satisfied.
2. Refer to Table A-2 or use technology to find the critical value $z_{\alpha/2}$ that corresponds to the desired confidence level. (For example, if the confidence level is 95%, the critical value is $z_{0.05/2} = 1.96$.)
3. Evaluate the margin of error \( E = z_{α/2} \cdot \frac{σ}{\sqrt{n}} \)

4. Using the value of the calculated margin of error \( E \) and the value of the sample mean \( \bar{x} \), find the values of the confidence interval limits: \( \bar{x} - E \) and \( \bar{x} + E \). Substitute those values in the general format for the confidence interval:

\[
\bar{x} - E < \mu < \bar{x} + E
\]

or

\[
\bar{x} \pm E
\]

or

\[
(\bar{x} - E, \bar{x} + E)
\]

5. Round the resulting values by using the following round-off rule.

**Round-Off Rule for Confidence Intervals Used to Estimate \( \mu \)**

1. When using the *original set of data* to construct a confidence interval, round the confidence interval limits to one more decimal place than is used for the original set of data.

2. When the original set of data is unknown and only the *summary statistics* \((n, \bar{x}, s)\) are used, round the confidence interval limits to the same number of decimal places used for the sample mean.

**Interpreting a Confidence Interval** As in Section 7-2, be careful to interpret confidence intervals correctly. After obtaining a confidence interval estimate of the population mean \( \mu \), such as a 95% confidence interval of 164.49 < \( \mu \) < 180.61, there is a correct interpretation and many incorrect interpretations.

**Correct:** “We are 95% confident that the interval from 164.49 to 180.61 actually does contain the true value of \( \mu \).” This means that if we were to select many different samples of the same size and construct the corresponding confidence intervals, in the long run 95% of them would actually contain the value of \( \mu \). (As in Section 7-2, this correct interpretation refers to the success rate of the process being used to estimate the population mean.)

**Incorrect:** Because \( \mu \) is a fixed constant, it would be incorrect to say “there is a 95% chance that \( \mu \) will fall between 164.49 and 180.61.” It would also be incorrect to say that “95% of all data values are between 164.49 and 180.61,” or that “95% of sample means fall between 164.49 and 180.61.” Creative readers can formulate other possible incorrect interpretations.

**Example 1** Weights of Men People have died in boat and aircraft accidents because an obsolete estimate of the mean weight of men was used. In recent decades, the mean weight of men has increased considerably, so we need to update our estimate of that mean so that boats, aircraft, elevators, and other such devices do not become dangerously overloaded. Using the weights of men from Data Set 1 in Appendix B, we obtain these sample statistics for the simple random sample: \( n = 40 \) and \( \bar{x} = 172.55 \) lb. Research from several other sources suggests that the population of weights of men has a standard deviation given by \( \sigma = 26 \) lb.

**continued**
Chapter 7  
Estimates and Sample Sizes

Captured Tank Serial Numbers Reveal Population Size

During World War II, Allied intelligence specialists wanted to determine the number of tanks Germany was producing. Traditional spy techniques provided unreliable results, but statisticians obtained accurate estimates by analyzing serial numbers on captured tanks. As one example, records show that Germany actually produced 271 tanks in June 1941. The estimate based on serial numbers was 244, but traditional intelligence methods resulted in the extreme estimate of 1550. (See “An Empirical Approach to Economic Intelligence in World War II,” by Ruggles and Brodie, *Journal of the American Statistical Association,* Vol. 42.)

a. Find the best point estimate of the mean weight of the population of all men.

b. Construct a 95% confidence interval estimate of the mean weight of all men.

c. What do the results suggest about the mean weight of 166.3 lb that was used to determine the safe passenger capacity of water vessels in 1960 (as given in the National Transportation and Safety Board safety recommendation M-04-04)?

**SOLUTION**

**REQUIREMENT CHECK**  
We must first verify that the requirements are satisfied. (1) The sample is a simple random sample. (2) The value of \( \sigma \) is assumed to be known with \( \sigma = 26 \) lb. (3) With \( n > 30 \), we satisfy the requirement that “the population is normally distributed or \( n > 30 \).” The requirements are therefore satisfied.

a. The sample mean of 172.55 lb is the best point estimate of the mean weight for the population of all men.

b. The 0.95 confidence level implies that \( \alpha = 0.05 \), so \( z_{0.025} = 1.96 \) (as was shown in Example 2 in Section 7-2). The margin of error \( E \) is first calculated as follows. (Extra decimal places are used to minimize rounding errors in the confidence interval.)

\[
E = \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} = 1.96 \cdot \frac{26}{\sqrt{40}} = 8.0574835
\]

With \( \bar{x} = 172.55 \) and \( E = 8.0574835 \), we now construct the confidence interval as follows:

\[
\bar{x} - E < \mu < \bar{x} + E
\]

\[
172.55 - 8.0574835 < \mu < 172.55 + 8.0574835
\]

\[
164.49 < \mu < 180.61 \quad \text{(rounded to two decimal places as in } \bar{x})
\]

c. Based on the confidence interval, it is possible that the mean weight of 166.3 lb used in 1960 could be the mean weight of men today. However, the best point estimate of 172.55 lb suggests that the mean weight of men is now considerably greater than 166.3 lb. Considering that an underestimate of the mean weight of men could result in lives lost through overloaded boats and aircraft, these results strongly suggest that additional data should be collected. (Additional data have been collected, and the assumed mean weight of men has been increased.)

**INTERPRETATION**  
The confidence interval from part (b) could also be expressed as 172.55 ± 8.06 or as (164.49, 180.61). Based on the sample with \( n = 40, \bar{x} = 172.55 \) and \( \sigma \) assumed to be 26, the confidence interval for the population mean \( \mu \) is 164.49 lb < \( \mu < 180.61 \) lb and this interval has a 0.95 confidence level. This means that if we were to select many different simple random samples of 40 men and construct the confidence intervals as we did here, 95% of them would actually contain the value of the population mean \( \mu \).

**Rationale for the Confidence Interval**  
The basic idea underlying the construction of confidence intervals relates to this property of the sampling distribution of sample means: If we collect simple random samples of the same size \( n \), the sample means are (at least approximately) normally distributed with mean \( \mu \) and standard
deviation $\sigma/\sqrt{n}$. In the standard score $z = (\bar{x} - \mu_0)/\sigma_0$, replace $\sigma_0$ with $\sigma/\sqrt{n}$, replace $\mu_0$ with $\mu$, then solve for $\mu$ to get

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

In the above equation, use the positive and negative values of $z$ and replace the right-most term by $E$. The right-hand side of the equation then yields the confidence interval limits of $\bar{x} - E$ and $\bar{x} + E$ that we are given earlier in this section. For a 95% confidence interval, we let $\alpha = 0.05$, so $z_{0.025} = 1.96$, so there is a 0.95 probability that a sample mean will be within 1.96 standard deviations (or $z_{0.025} \cdot \sigma/\sqrt{n}$ or $E$) of $\mu$.

If the sample mean $\bar{x}$ is within $E$ of the population mean, then $\mu$ is between $\bar{x} - E$ and $\bar{x} + E$. That is, $\bar{x} - E \leq \mu \leq \bar{x} + E$.

**Determining Sample Size Required to Estimate $\mu$**

When collecting a simple random sample that will be used to estimate a population mean $\mu$, how many sample values must be obtained? For example, suppose we want to estimate the mean weight of airline passengers (an important value for reasons of safety). How many passengers must be randomly selected and weighed? Determining the size of a simple random sample is a very important issue, because samples that are needlessly large waste time and money, and samples that are too small may lead to poor results.

If we use the expression for the margin of error ($E = \frac{z_{0.025} \cdot \sigma}{\sqrt{n}}$) and solve for the sample size $n$, we get Formula 7-4 shown below.

### Finding the Sample Size Required to Estimate a Population Mean

**Objective**

Determine how large a sample should be in order to estimate the population mean $\mu$.

**Notation**

- $\mu =$ population mean
- $\sigma =$ population standard deviation
- $\bar{x} =$ sample mean
- $E =$ desired margin of error
- $z_{0.025} =$ $z$ score separating an area of $0.025$ in the right tail of the standard normal distribution

**Requirements**

The sample must be a simple random sample.

**Formula 7-4**

$$n = \left[ \frac{z_{0.025} \sigma}{E} \right]^2$$

Formula 7-4 is remarkable because it shows that the sample size does not depend on the size ($N$) of the population; the sample size depends on the desired confidence level, the desired margin of error, and the value of the standard deviation $\sigma$. (See Exercise 38 for dealing with cases in which a relatively large sample is selected without replacement from a finite population.)

The sample size must be a whole number, because it represents the number of sample values that must be found. However, Formula 7-4 usually gives a result that is not a whole number, so we use the following round-off rule. (It is based on the
principle that when rounding is necessary, the required sample size should be rounded upward so that it is at least adequately large as opposed to slightly too small.)

**Round-Off Rule for Sample Size n**

If the computed sample size \( n \) is not a whole number, round the value of \( n \) up to the next larger whole number.

**Dealing with Unknown \( \sigma \) When Finding Sample Size** Formula 7-4 requires that we substitute a known value for the population standard deviation \( \sigma \), but in reality, it is usually unknown. When determining a required sample size (not constructing a confidence interval), here are some ways that we can work around the problem of not knowing the value of \( \sigma \):

1. Use the range rule of thumb (see Section 3-3) to estimate the standard deviation as follows: \( \sigma \approx \text{range}/4 \). (With a sample of 87 or more values randomly selected from a normally distributed population, \( \text{range}/4 \) will yield a value that is greater than or equal to \( \sigma \) at least 95% of the time. (See “Using the Sample Range as a Basis for Calculating Sample Size in Power Calculations,” by Richard Browne, *American Statistician*, Vol. 55, No. 4.)

2. Start the sample collection process without knowing \( \sigma \) and, using the first several values, calculate the sample standard deviation \( s \) and use it in place of \( \sigma \). The estimated value of \( \sigma \) can then be improved as more sample data are obtained, and the sample size can be refined accordingly.

3. Estimate the value of \( \sigma \) by using the results of some other study that was done earlier.

In addition, we can sometimes be creative in our use of other known results. For example, IQ tests are typically designed so that the mean is 100 and the standard deviation is 15. Statistics students have IQ scores with a mean greater than 100 and a standard deviation less than 15 (because they are a more homogeneous group than people randomly selected from the general population). We do not know the specific value of \( \sigma \) for statistics students, but we can play it safe by using \( \sigma = 15 \). Using a value for \( \sigma \) that is larger than the true value will make the sample size larger than necessary, but using a value for \( \sigma \) that is too small would result in a sample size that is inadequate. When calculating the sample size \( n \), any errors should always be conservative in the sense that they make \( n \) too large instead of too small.

**Example 2** IQ Scores of Statistics Students Assume that we want to estimate the mean IQ score for the population of statistics students. How many statistics students must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean?

**Solution** For a 95% confidence interval, we have \( \alpha = 0.05 \), so \( z_{0.025} = 1.96 \). Because we want the sample mean to be within 3 IQ points of \( \mu \), the margin of error is \( E = 3 \). Also, \( \sigma = 15 \) (see the discussion that immediately precedes this example). Using Formula 7-4, we get

\[
 n = \left[ \frac{z_{0.025} \sigma}{E} \right]^2 = \left[ \frac{1.96 \cdot 15}{3} \right]^2 = 96.04 \approx 97 \quad \text{(rounded up)}
\]
Among the thousands of statistics students, we need to obtain a simple random sample of at least 97 students. Then we need to get their IQ scores. With a simple random sample of only 97 statistics students, we will be 95% confident that the sample mean $\bar{x}$ is within 3 IQ points of the true population mean $\mu$.

If we want a more accurate estimate, we can decrease the margin of error. Halving the margin of error quadruples the sample size, so if you want more accurate results, the sample size must be substantially increased. Because large samples generally require more time and money, there is often a need for a tradeoff between the sample size and the margin of error $E$.

Confidence Intervals See the end of Section 7-4 for the confidence interval procedures that apply to the methods of this section as well as those of Section 7-4. STATDISK, Minitab, Excel, and the TI-83/84 Plus calculator can all be used to find confidence intervals when we want to estimate a population mean and the requirements of this section (including a known value of $\sigma$) are all satisfied.

Sample Size Determination Sample size calculations are not included with the TI-83/84 Plus calculator, Minitab, or Excel. The STATDISK procedure for determining the sample size required to estimate a population mean $\mu$ is described below.

### 7-3 Basic Skills and Concepts

#### Statistical Literacy and Critical Thinking

1. **Point Estimate** In general, what is a point estimate of a population parameter? Given a simple random sample of heights from some population, such as the population of all basketball players in the NBA, how would you find the best point estimate of the population mean?

2. **Simple Random Sample** A design engineer for the Ford Motor Company must estimate the mean leg length of all adults. She obtains a list of the 1275 employees at her facility, then obtains a simple random sample of 50 employees. If she uses this sample to construct a 95% confidence interval to estimate the mean leg length for the population of all adults, will her estimate be good? Why or why not?

3. **Confidence Interval** Based on the heights of women listed in Data Set 1 in Appendix B, and assuming that heights of women have a standard deviation of $\sigma = 2.5$ in., this 95% confidence interval is obtained: $62.42$ in. $< \mu < 63.97$ in. Assuming that you are a newspaper reporter, write a statement that correctly interprets that confidence interval and includes all of the relevant information.
Chapter 7  Estimates and Sample Sizes

4. **Unbiased Estimator** One of the features of the sample mean that makes it a good estimator of a population mean \( \mu \) is that the sample mean is an unbiased estimator. What does it mean for a statistic to be an unbiased estimator of a population parameter?

**Finding Critical Values.** In Exercises 5–8, find the indicated critical value \( z_{\frac{\alpha}{2}} \).

5. Find the critical value \( z_{\frac{0.05}{2}} \) that corresponds to a 90% confidence level.
6. Find the critical value \( z_{\frac{0.08}{2}} \) that corresponds to a 98% confidence level.
7. Find \( z_{\frac{\alpha}{2}} \) for \( \alpha = 0.20 \).
8. Find \( z_{\frac{\alpha}{2}} \) for \( \alpha = 0.04 \).

**Verifying Requirements and Finding the Margin of Error.** In Exercises 9–12, find the margin of error and confidence interval if the necessary requirements are satisfied. If the requirements are not all satisfied, state that the margin of error and confidence interval cannot be calculated using the methods of this section.

9. **Credit Rating** FICO (Fair, Isaac, and Company) credit rating scores of a simple random sample of applicants for credit cards: 95% confidence; \( n = 50, \bar{x} = 677 \), and \( \sigma \) is known to be 68.
10. **Braking Distances** The braking distances of a simple random sample of cars: 95% confidence; \( n = 32, \bar{x} = 137 \) ft, and \( \sigma \) is known to be 7 ft.
11. **Rainfall Amounts** The amounts of rainfall for a simple random sample of Saturdays in Boston: 99% confidence; \( n = 12, \bar{x} = 0.133 \) in., \( \sigma \) is known to be 0.212 in., and the population is known to have daily rainfall amounts with a distribution that is far from normal.
12. **Failure Times** The times before failure of integrated circuits used in calculators: 99% confidence; \( n = 25, \bar{x} = 112 \) hours, \( \sigma \) is known to be 18.6 hours, and the distribution of all times before failure is far from normal.

**Finding Sample Size.** In Exercises 13–16, use the given information to find the minimum sample size required to estimate an unknown population mean \( \mu \).

13. **Credit Rating** How many adults must be randomly selected to estimate the mean FICO (credit rating) score of working adults in the United States. We want 95% confidence that the sample mean is within 3 points of the population mean, and the population standard deviation is 68.
14. **Braking Distances** How many cars must be randomly selected and tested in order to estimate the mean braking distance of registered cars in the United States. We want 99% confidence that the sample mean is within 2 ft of the population mean, and the population standard deviation is known to be 7 ft.
15. **Rainfall Amounts** How many daily rainfall amounts in Boston must be randomly selected to estimate the mean daily rainfall amount? We want 99% confidence that the sample mean is within 0.010 in. of the population mean, and the population standard deviation is known to be 0.212 in.
16. **Failure Times** How many integrated circuits must be randomly selected and tested for time to failure in order to estimate the mean time to failure? We want 95% confidence that the sample mean is within 2 hr of the population mean, and the population standard deviation is known to be 18.6 hours.

**Interpreting Results.** In Exercises 17–20, refer to the accompanying TI-83/84 Plus calculator display of a 95% confidence interval. The sample display results from using a simple random sample of the amounts of tar (in milligrams) in cigarettes that are all king size, nonfiltered, nonmenthol, and non-light.

17. Identify the value of the point estimate of the population mean \( \mu \).
18. Express the confidence interval in the format of \( \bar{x} - E < \mu < \bar{x} + E \).
19. Express the confidence interval in the format of \( \bar{x} \pm E \).
20. Write a statement that interprets the 95% confidence interval.

21. Weights of Women Using the simple random sample of weights of women from Data Set 1 in Appendix B, we obtain these sample statistics: \( n = 40 \) and \( \bar{x} = 146.22 \) lb. Research from other sources suggests that the population of weights of women has a standard deviation given by \( \sigma = 30.86 \) lb.
   a. Find the best point estimate of the mean weight of all women.
   b. Find a 95% confidence interval estimate of the mean weight of all women.

22. NCAA Football Coach Salaries A simple random sample of 40 salaries of NCAA football coaches has a mean of $415,953 and a standard deviation of $463,364.
   a. Find the best point estimate of the mean salary of all NCAA football coaches.
   b. Construct a 95% confidence interval estimate of the mean salary of an NCAA football coach.
   c. Does the confidence interval contain the actual population mean of $474,477?

23. Perception of Time Randomly selected statistics students of the author participated in an experiment to test their ability to determine when 1 min (or 60 seconds) has passed. Forty students yielded a sample mean of 58.3 sec. Assume that \( \sigma = 9.5 \) sec.
   a. Find the best point estimate of the mean time for all statistics students.
   b. Construct a 95% confidence interval estimate of the population mean of all statistics students.
   c. Based on the results, is it likely that their estimates have a mean that is reasonably close to 60 sec?

24. Red Blood Cell Count A simple random sample of 50 adults (including males and females) is obtained, and each person's red blood cell count (in cells per microliter) is measured. The sample mean is 4.63. The population standard deviation for red blood cell counts is 0.54.
   a. Find the best point estimate of the mean red blood cell count of adults.
   b. Construct a 99% confidence interval estimate of the mean red blood cell count of adults.
   c. The normal range of red blood cell counts for adults is given by the National Institutes of Health as 4.7 to 6.1 for males and 4.3 to 5.4 for females. What does the confidence interval suggest about these normal ranges?

25. SAT Scores A simple random sample of 125 SAT scores has a mean of 1522. Assume that SAT scores have a standard deviation of 333.
   a. Construct a 95% confidence interval estimate of the mean SAT score.
   b. Construct a 99% confidence interval estimate of the mean SAT score.
   c. Which of the preceding confidence intervals is wider? Why?

26. Birth Weights A simple random sample of birth weights in the United States has a mean of 3433 g. The standard deviation of all birth weights is 495 g.
   a. Using a sample size of 75, construct a 95% confidence interval estimate of the mean birth weight in the United States.
   b. Using a sample size of 75,000, construct a 95% confidence interval estimate of the mean birth weight in the United States.
   c. Which of the preceding confidence intervals is wider? Why?

27. Blood Pressure Levels When 14 different second-year medical students at Bellevue Hospital measured the blood pressure of the same person, they obtained the results listed below. Assuming that the population standard deviation is known to be 10 mmHg, construct a 95% confidence interval estimate of the population mean. Ideally, what should the confidence interval be in this situation?

\[
138 \quad 130 \quad 135 \quad 140 \quad 120 \quad 125 \quad 120 \quad 130 \quad 130 \quad 144 \quad 143 \quad 140 \quad 130 \quad 150
\]
354 Chapter 7 Estimates and Sample Sizes

28. Telephone Digits Polling organizations typically generate the last digits of telephone numbers so that people with unlisted numbers are included. Listed below are digits randomly generated by STATDISK. Such generated digits are from a population with a standard deviation of 2.87.

a. Use the methods of this section to construct a 95% confidence interval estimate of the mean of all such generated digits.

b. Are the requirements for the methods of this section all satisfied? Does the confidence interval from part (a) serve as a good estimate of the population mean? Explain.

1 1 7 0 7 4 5 1 7 6

Large Data Sets from Appendix B. In Exercises 29 and 30, refer to the data set from Appendix B.

29. Movie Gross Amounts Refer to Data Set 9 from Appendix B and construct a 95% confidence interval estimate of the mean gross amount for the population of all movies. Assume that the population standard deviation is known to be 100 million dollars.

30. FICO Credit Rating Scores Refer to Data Set 24 in Appendix B and construct the 99% confidence interval estimate of the mean FICO score for the population. Assume that the population standard deviation is 92.2.

Finding Sample Size. In Exercises 31–36, find the indicated sample size.

31. Sample Size for Mean IQ of NASA Scientists The Wechsler IQ test is designed so that the mean is 100 and the standard deviation is 15 for the population of normal adults. Find the sample size necessary to estimate the mean IQ score of scientists currently employed by NASA. We want to be 95% confident that our sample mean is within five IQ points of the true mean. The mean for this population is clearly greater than 100. The standard deviation for this population is probably less than 15 because it is a group with less variation than a group randomly selected from the general population; therefore, if we use $\sigma = 15$, we are being conservative by using a value that will make the sample size at least as large as necessary. Assume then that $\sigma = 15$ and determine the required sample size.

32. Sample Size for White Blood Cell Count What sample size is needed to estimate the mean white blood cell count (in cells per microliter) for the population of adults in the United States? Assume that you want 99% confidence that the sample mean is within 0.2 of the population mean. The population standard deviation is 2.5.

33. Sample Size for Atkins Weight Loss Program You want to estimate the mean weight loss of people one year after using the Atkins weight loss program. How many people on that program must be surveyed if we want to be 95% confident that the sample mean weight loss is within 0.25 lb of the true population mean? Assume that the population standard deviation is known to be 10.6 lb (based on data from “Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Risk Reduction,” by Dansinger, et al., Journal of the American Medical Association, Vol. 293, No. 1). Is the resulting sample size practical?

34. Grade Point Average A researcher wants to estimate the mean grade point average of all current college students in the United States. She has developed a procedure to standardize scores from colleges using something other than a scale between 0 and 4. How many grade point averages must be obtained so that the sample mean is within 0.1 of the population mean? Assume that a 90% confidence level is desired. Also assume that a pilot study showed that the population standard deviation is estimated to be 0.88.

35. Sample Size Using Range Rule of Thumb You want to estimate the mean amount of annual tuition being paid by current full-time college students in the United States. First use the range rule of thumb to make a rough estimate of the standard deviation of the amounts spent. It is reasonable to assume that tuition amounts range from $0 to about $40,000. Then use that estimated standard deviation to determine the sample size corresponding to 95% confidence and a $100 margin of error.
Estimating a Population Mean: \( \sigma \) Not Known

### 36. Sample Size Using Sample Data
Refer to Data Set 1 in Appendix B and find the maximum and minimum pulse rates for males, then use those values with the range rule of thumb to estimate \( \sigma \). How many adult males must you randomly select and test if you want to be 95% confident that the sample mean pulse rate is within 2 beats (per minute) of the true population mean \( \mu \)? If, instead of using the range rule of thumb, the standard deviation of the male pulse rates in Data Set 1 is used as an estimate of \( \sigma \), is the required sample size very different? Which sample size is likely to be closer to the correct sample size?

### 37. Confidence Interval with Finite Population Correction Factor
The standard error of the mean is \( \frac{\sigma}{\sqrt{n}} \), provided that the population size is infinite or very large or sampling is with replacement. If the population size \( N \) is finite, then the correction factor \( \sqrt{\frac{(N - n)}{(N - 1)}} \) should be used whenever \( n > 0.05N \). The margin of error \( E \) is multiplied by this correction factor as shown below. Repeat part (a) of Exercise 25 assuming that the sample is selected without replacement from a population of size 200. How is the confidence interval affected by the additional information about the population size?

\[
E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \frac{\sqrt{N - n}}{N - 1}
\]

### 38. Sample Size with Finite Population Correction Factor
The methods of this section assume that sampling is from a population that is very large or infinite, and that we are sampling with replacement. If we have a relatively small population and sample without replacement, we should modify \( E \) to include a finite population correction factor, so that the margin of error is as shown in Exercise 37, where \( N \) is the population size. That expression for the margin of error can be solved for \( n \) to yield

\[
n = \frac{N\sigma^2(z_{\alpha/2})^2}{(N - 1)E^2 + \sigma^2(z_{\alpha/2})^2}
\]

Repeat Exercise 32, assuming that a simple random sample is selected without replacement from a population of 500 people. Does the additional information about the population size have much of an effect on the sample size?

### 7-4 Estimating a Population Mean: \( \sigma \) Not Known

**Key Concept** In this section we present methods for estimating a population mean when the population standard deviation \( \sigma \) is unknown. With \( \sigma \) unknown, we use the Student \( t \) distribution (instead of the normal distribution), assuming that the relevant requirements are satisfied. Because \( \sigma \) is typically unknown in real circumstances, the methods of this section are realistic and practical, and they are often used.

As in Section 7-3, the sample mean \( \bar{x} \) is the best point estimate (or single-valued estimate) of the population mean \( \mu \).

The sample mean \( \bar{x} \) is the best point estimate of the population mean \( \mu \).

Here is a major point of this section: If \( \sigma \) is not known, but the relevant requirements are satisfied, we use a Student \( t \) distribution (instead of a normal distribution), as developed by William Gosset (1876–1937). Gosset was a Guinness Brewery employee who needed a distribution that could be used with small samples. The Irish brewery where he worked did not allow the publication of research results, so Gosset published under the pseudonym “Student.” (In the interest of research and better serving his readers, the author visited the Guinness Brewery and sampled some of the product. Such commitment!)
Chapter 7

Estimates and Sample Sizes

**Student t Distribution**

If a population has a normal distribution, then the distribution of

\[ t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \]

is a **Student t distribution** for all samples of size \( n \). A Student \( t \) distribution is often referred to simply as a **\( t \) distribution**.

Because we do not know the value of the population standard deviation \( \sigma \), we estimate it with the value of the sample standard deviation \( s \), but this introduces another source of unreliability, especially with small samples. In order to maintain a desired confidence level, such as 95%, we compensate for this additional unreliability by making the confidence interval wider: We use critical values (from a Student \( t \) distribution) that are larger than the critical values of \( z_{\alpha/2} \) from the normal distribution. A critical value of \( t_{\alpha/2} \) can be found by using technology or Table A-3, but we must first identify the number of **degrees of freedom**.

**Definition**

The number of **degrees of freedom** for a collection of sample data is the number of sample values that can vary after certain restrictions have been imposed on all data values. The number of degrees of freedom is often abbreviated as \( df \).

For example, if 10 students have quiz scores with a mean of 80, we can freely assign values to the first 9 scores, but the 10th score is then determined. The sum of the 10 scores must be 800, so the 10th score must equal 800 minus the sum of the first 9 scores. Because those first 9 scores can be freely selected to be any values, we say that there are 9 degrees of freedom available. For the applications of this section, the number of degrees of freedom is simply the sample size minus 1.

\[ \text{degrees of freedom} = n - 1 \]

**Example 1**

**Finding a Critical \( t \) Value**

A sample of size \( n = 7 \) is a simple random sample selected from a normally distributed population. Find the critical value \( t_{\alpha/2} \) corresponding to a 95% confidence level.

**Solution**

Because \( n = 7 \), the number of degrees of freedom is given by \( n - 1 = 6 \). Using Table A-3, we locate the 6th row by referring to the column at the extreme left. A 95% confidence level corresponds to \( \alpha = 0.05 \), and confidence intervals require that the area \( \alpha \) be divided equally between the left and right tails of the distribution (as in Figure 7-4), so we find the column listing values for an area of \( 0.05 \) in two tails. The value corresponding to the row for 6 degrees of freedom and the column for an area of \( 0.05 \) in two tails is 2.447, so \( t_{\alpha/2} = 2.447 \). (See Figure 7-4.) We could also express this as \( t_{0.025} = 2.447 \). Such critical values \( t_{\alpha/2} \) are used for the margin of error \( E \) and confidence interval as shown below.
7-4 Estimating a Population Mean: \( \sigma \) Not Known

**Objective**
Construct a confidence interval used to estimate a population mean.

**Notation**

- \( \mu \) = population mean
- \( \bar{x} \) = sample mean
- \( s \) = sample standard deviation
- \( n \) = number of sample values
- \( E \) = margin of error
- \( t_{\alpha/2} \) = critical \( t \) value separating an area of \( \alpha/2 \) in the right tail of the \( t \) distribution

**Requirements**

1. The sample is a simple random sample.
2. Either the sample is from a normally distributed population or \( n \geq 30 \).

**Confidence Interval**

\[
\bar{x} - E < \mu < \bar{x} + E \quad \text{where} \quad E = t_{\alpha/2} \frac{s}{\sqrt{n}} \quad (df = n - 1)
\]

or

\[
\bar{x} \pm E
\]

or

\[
(\bar{x} - E, \bar{x} + E)
\]

**Requirements** As in Section 7-3, the requirement of a normally distributed population is not a strict requirement, so we can usually consider the population to be normally distributed after using the sample data to confirm that there are no outliers and the histogram has a shape that is not substantially far from a normal distribution. Also, as in Section 7-3, the requirement that the sample size is \( n \geq 30 \) is commonly used as a guideline, but the minimum sample size actually depends on how much the population distribution departs from a normal distribution. (If a population is known to be normally distributed, the distribution of sample means \( \bar{x} \) is exactly a normal distribution with mean \( \mu \) and standard deviation \( \sigma / \sqrt{n} \); if the population is not
normally distributed, large \((n > 30)\) samples yield sample means with a distribution that is approximately normal with mean \(\mu\) and standard deviation \(\sigma/\sqrt{n}\).)

**Procedure for Constructing a Confidence Interval for \(\mu\) (with \(\sigma\) unknown)**

1. Verify that the requirements are satisfied.
2. Using \(n - 1\) degrees of freedom, refer to Table A-3 or use technology to find the critical value \(t_{\alpha/2}\) that corresponds to the desired confidence level. (For the confidence level, refer to the “Area in Two Tails.”)
3. Evaluate the margin of error \(E = t_{\alpha/2} \cdot s/\sqrt{n}\).
4. Using the value of the calculated margin of error \(E\) and the value of the sample mean \(\bar{x}\), find the values of the confidence interval limits: \(\bar{x} - E\) and \(\bar{x} + E\). Substitute those values in the general format for the confidence interval.
5. Round the resulting confidence interval limits. If using the original set of data, round to one more decimal place than is used for the original set of data. If using summary statistics \((n, \bar{x}, s)\), round the confidence interval limits to the same number of decimal places used for the sample mean.

**EXAMPLE 2**

**Constructing a Confidence Interval: Garlic for Reducing Cholesterol**

A common claim is that garlic lowers cholesterol levels. In a test of the effectiveness of garlic, 49 subjects were treated with doses of raw garlic, and their cholesterol levels were measured before and after the treatment. The changes in their levels of LDL cholesterol (in mg/dL) have a mean of 0.4 and a standard deviation of 21.0 (based on data from “Effect of Raw Garlic vs Commercial Garlic Supplements on Plasma Lipid Concentrations in Adults With Moderate Hypercholesterolemia,” by Gardner et al., *Archives of Internal Medicine*, Vol. 167). Use the sample statistics of \(n, \bar{x}, s\), to construct a 95% confidence interval estimate of the mean net change in LDL cholesterol after the garlic treatment. What does the confidence interval suggest about the effectiveness of garlic in reducing LDL cholesterol?

**REQUIREMENT CHECK**

We must first verify that the requirements are satisfied. (1) The detailed design of the garlic trials justify the assumption that the sample is a simple random sample. (2) The requirement that “the population is normally distributed or \(n > 30\)” is satisfied because \(n = 49\). The requirements are therefore satisfied.

The confidence level of 95% implies that \(\alpha = 0.05\). With \(n = 49\), the number of degrees of freedom is \(n - 1 = 48\). If using Table A-3, we look for the row with 48 degrees of freedom and the column corresponding to \(\alpha = 0.05\) in two tails. Table A-3 does not include 48 degrees of freedom, and the closest number of degrees of freedom is 50, so we can use \(t_{\alpha/2} = 2.009\). (If we use technology, we get the more accurate result of \(t_{\alpha/2} = 2.011\)).

Using \(t_{\alpha/2} = 2.009, s = 21.0,\) and \(n = 49\), we find the margin of error \(E\) as follows:

\[
E = t_{\alpha/2} \frac{s}{\sqrt{n}} = 2.009 \cdot \frac{21.0}{\sqrt{49}} = 6.027
\]
With \( \bar{x} = 0.4 \) and \( E = 6.027 \), we construct the confidence interval as follows:

\[
\bar{x} - E < \mu < \bar{x} + E
\]

\[
0.4 - 6.027 < \mu < 0.4 + 6.027
\]

\[-5.6 < \mu < 6.4 \] (rounded to one decimal place, as in the given sample mean)

**INTERPRETATION**

This result could also be expressed in the format of \( 0.4 \pm 6.0 \) or \((-5.6, 6.4).\) On the basis of the given sample results, we are 95% confident that the limits of \(-5.6\) and 6.4 actually do contain the value of \( \mu, \) the mean of the changes in LDL cholesterol for the population.

Because the confidence interval limits contain the value of 0, it is very possible that the mean of the changes in LDL cholesterol is equal to 0, suggesting that the garlic treatment did not affect the LDL cholesterol levels. It does not appear that the garlic treatment is effective in lowering LDL cholesterol.

We now list the important properties of the Student \( t \) distribution that has been introduced in this section.

**Important Properties of the Student \( t \) Distribution**

1. The Student \( t \) distribution is different for different sample sizes. (See Figure 7-5 for the cases \( n = 3 \) and \( n = 12 \).)
2. The Student \( t \) distribution has the same general symmetric bell shape as the standard normal distribution, but it reflects the greater variability (with wider distributions) that is expected with small samples.
3. The Student \( t \) distribution has a mean of \( t = 0 \) (just as the standard normal distribution has a mean of \( z = 0 \)).

**Figure 7-5**

*Student \( t \) Distributions for \( n = 3 \) and \( n = 12 \)*

The Student \( t \) distribution has the same general shape and symmetry as the standard normal distribution, but it reflects the greater variability that is expected with small samples.
4. The standard deviation of the Student \( t \) distribution varies with the sample size, but it is greater than 1 (unlike the standard normal distribution, which has \( \sigma = 1 \)).

5. As the sample size \( n \) gets larger, the Student \( t \) distribution gets closer to the standard normal distribution.

**Choosing the Appropriate Distribution**

It is sometimes difficult to decide whether to use the standard normal \( z \) distribution or the Student \( t \) distribution. The flowchart in Figure 7-6 and the accompanying Table 7-1 both summarize the key points to consider when constructing confidence intervals for estimating the population mean. In Figure 7-6 or Table 7-1, note that if we have a small sample (\( n \leq 30 \)) drawn from a distribution that differs dramatically from a normal distribution, we can’t use the methods described in this chapter. One alternative is to use nonparametric methods (see Chapter 13), and another alternative is to use the computer bootstrap method. In both of those approaches, no assumptions are made about the original population. The bootstrap method is described in the Technology Project at the end of this chapter.

*Important:* Figure 7-6 and Table 7-1 assume that the sample is a simple random sample. If the sample data have been collected using some inappropriate method, such as a convenience sample or a voluntary response sample, it is very possible that no methods of statistics can be used to find a useful estimate of a population mean.

---

**Figure 7-6**  Choosing Between \( z \) and \( t \)
Notes: 1. **Criteria for deciding whether the population is normally distributed:** The population need not be exactly normal, but it should appear to be somewhat symmetric with one mode and no outliers.
2. **Sample size** \( n > 30 \): This is a common guideline, but sample sizes of 15 to 30 are adequate if the population appears to have a distribution that is not far from being normal and there are no outliers. For some population distributions that are extremely far from normal, the sample size might need to be much larger than 30.

### Table 7-1 Choosing Between \( z \) and \( t \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use normal (( z )) distribution.</td>
<td>( \sigma ) known and normally distributed population or ( \sigma ) known and ( n &gt; 30 )</td>
</tr>
<tr>
<td>Use ( t ) distribution.</td>
<td>( \sigma ) not known and normally distributed population or ( \sigma ) not known and ( n &gt; 30 )</td>
</tr>
<tr>
<td>Use a nonparametric method or bootstrapping.</td>
<td>Population is not normally distributed and ( n \leq 30 ).</td>
</tr>
</tbody>
</table>

**Estimating Crowd Size**

There are sophisticated methods of analyzing the size of a crowd. Aerial photographs and measures of people density can be used with reasonably good accuracy. However, reported crowd size estimates are often simple guesses. After the Boston Red Sox won the World Series for the first time in 86 years, Boston city officials estimated that the celebration parade was attended by 3.2 million fans. Boston police provided an estimate of around 1 million, but it was admittedly based on guesses by police commanders. A photo analysis led to an estimate of around 150,000. Boston University Professor Farouk El-Baz used images from the U.S. Geological Survey to develop an estimate of at most 400,000. MIT physicist Bill Donnelly said that “it’s a serious thing if people are just putting out any number. It means other things aren’t being vetted that carefully.”

**Example 3**

**Choosing Distributions** You plan to construct a confidence interval for the population mean \( \mu \). Use the given data to determine whether the margin of error \( E \) should be calculated using a critical value of \( Z_{\alpha/2} \) (from the normal distribution), a critical value of \( t_{\alpha/2} \) (from a \( t \) distribution), or neither (so that the methods of Section 7-3 and this section cannot be used).

a. \( n = 9, \bar{x} = 75, s = 15 \), and the population has a normal distribution.

b. \( n = 5, \bar{x} = 20, s = 2 \), and the population has a skewed distribution.

c. \( n = 12, \bar{x} = 98.6, \sigma = 0.6 \), and the population has a normal distribution.
   (In reality, \( \sigma \) is rarely known.)

d. \( n = 75, \bar{x} = 98.6, \sigma = 0.6 \), and the population has a skewed distribution.
   (In reality, \( \sigma \) is rarely known.)

e. \( n = 75, \bar{x} = 98.6, s = 0.6 \), and the population has a skewed distribution.

**Solution**

Refer to Figure 7-6 or Table 7-1.

a. Because the population standard deviation \( \sigma \) is not known and the population is normally distributed, the margin of error is calculated using \( t_{\alpha/2} \).

b. Because the sample is small (\( n \leq 30 \)) and the population does not have a normal distribution, the margin of error \( E \) should not be calculated using a critical value of \( Z_{\alpha/2} \) or \( t_{\alpha/2} \). The methods of Section 7-3 and this section do not apply.
Estimates and Sample Sizes

Chapter 7

c. Because $\sigma$ is known and the population has a normal distribution, the margin of error is calculated using $z_{\alpha/2}$.

d. Because the sample is large ($n > 30$) and $\sigma$ is known, the margin of error is calculated using $z_{\alpha/2}$.

e. Because the sample is large ($n > 30$) and $\sigma$ is not known, the margin of error is calculated using $t_{\alpha/2}$.

**Example 4**

**Confidence Interval for Alcohol in Video Games**

Twelve different video games showing substance use were observed. The duration times (in seconds) of alcohol use were recorded, with the times listed below (based on data from “Content and Ratings of Teen-Rated Video Games,” by Haninger and Thompson, *Journal of the American Medical Association*, Vol. 291, No. 7). The design of the study justifies the assumption that the sample can be treated as a simple random sample. Use the sample data to construct a 95% confidence interval estimate of the mean duration time that the video showed the use of alcohol.

**MINITAB**

```
84 14 583 50 0 57 207 43 178 0 2 57
```

**Solution**

**Requirement Check**

We must first verify that the requirements are satisfied. (1) We can consider the sample to be a simple random sample. (2) When checking the requirement that “the population is normally distributed or $n > 30$,” we see that the sample size is $n = 12$, so we must determine whether the data appear to be from a population with a normal distribution. Shown below are a Minitab-generated histogram and a STATDISK-generated normal quantile plot. The histogram does not appear to be bell-shaped, and the points in the normal quantile plot are not reasonably close to a straight-line pattern, so it appears that the times are not from a population having a normal distribution. The requirements are not satisfied. If we were to proceed with the construction of the confidence interval, we would get $1.8 \text{ sec} < \mu < 210.7 \text{ sec}$, but this result is questionable because it assumes incorrectly that the requirements are satisfied.

**Interpretation**

Because the requirement that “the population is normally distributed or $n > 30$” is not satisfied, we do not have 95% confidence that the limits of 1.8 sec and 210.7 sec actually do contain the value of the population mean. We should use some other approach for finding the confidence interval limits. For example, the author used bootstrap resampling as described in the Technology Project at the end of this section. The confidence interval of $35.3 \text{ sec} < \mu < 205.6 \text{ sec}$ was obtained.
Finding Point Estimate and $E$ from a Confidence Interval

Later in this section we will describe how computer software and calculators can be used to find a confidence interval. A typical use requires that you enter a confidence level and sample statistics, and the display shows the confidence interval limits. The sample mean $\bar{x}$ is the value midway between those limits, and the margin of error $E$ is one-half the difference between those limits (because the upper limit is $\bar{x} + E$ and the lower limit is $\bar{x} - E$, the distance separating them is $2E$).

Point estimate of $\mu$: $\bar{x} = \frac{(\text{upper confidence limit}) + (\text{lower confidence limit})}{2}$

Margin of error: $E = \frac{(\text{upper confidence limit}) - (\text{lower confidence limit})}{2}$

**Example 5** Weights of Garbage Data Set 22 in Appendix B lists the weights of discarded garbage from a sample of 62 households. The accompanying TI-83/84 Plus calculator screen displays results from using the 62 amounts of total weights (in pounds) to construct a 95% confidence interval estimate of the mean weight of garbage discarded by the population of all households. Use the displayed confidence interval to find the values of the best point estimate $\bar{x}$ and the margin of error $E$.

In the following calculations, results are rounded to three decimal places, which is one additional decimal place beyond the two decimal places used for the original list of weights.

$\bar{x} = \frac{(\text{upper confidence limit}) + (\text{lower confidence limit})}{2}$

$E = \frac{(\text{upper confidence limit}) - (\text{lower confidence limit})}{2}$

**TI-83/84 PLUS**

TI Interval: $(24.28, 30.607)$

$\bar{x} = 27.443 \text{ lb}$

$n = 62$
Chapter 7      Estimates and Sample Sizes

Using Confidence Intervals to Describe, Explore, or Compare Data

In some cases, we might use a confidence interval to achieve an ultimate goal of estimating the value of a population parameter. In other cases, confidence intervals might be among the different tools used to describe, explore, or compare data sets. Figure 7-7 shows graphs of confidence intervals for the body mass indexes (BMI) of a sample of females and a separate sample of males. (Both samples are listed in Data Set 1 in Appendix B.) Because the confidence intervals in Figure 7-7 overlap, it is possible that females and males have the same mean BMI index, so there does not appear to be a significant difference between the mean BMI index of females and males.

CAUTION

As in Sections 7-2 and 7-3, confidence intervals can be used informally to compare different data sets, but the overlapping of confidence intervals should not be used for making formal and final conclusions about equality of means.

Determining Sample Size  Section 7-2 included a subsection describing methods for determining the size of a sample needed to estimate a population proportion, and Section 7-3 included a subsection with methods for determining the size of a sample needed to estimate a population mean. This section does not include a similar subsection. When determining the sample size needed to estimate a population mean, use the procedure described in Section 7-3, which requires an estimated or known value of the population standard deviation.

The following procedures apply to confidence intervals for estimating a mean \( \mu \), and they include the confidence intervals described in Section 7-3 as well as the confidence intervals presented in this section. Before using computer software or a calculator to generate a confidence interval, be sure to first check that the relevant requirements are satisfied. See the requirements listed near the beginning of this section and Section 7-3.

**STATDISK**  You must first find the sample size \( n \), the sample mean \( \bar{x} \), and the sample standard deviation \( s \). (See the STATDISK procedure described in Section 3-3.) Select **Analysis** from the main menu bar, select **Confidence Intervals**, then select **Population Mean**. Enter the items in the dialog box, then click the **Evaluate** button. The confidence interval will be displayed. STATDISK will automatically choose between the normal and \( t \) distributions, depending on whether a value for the population standard deviation is entered.

**MINITAB**  Minitab allows you to use either the summary statistics \( n, \bar{x}, \) and \( s \) or a list of the original sample values. Select **Stat** and **Basic Statistics**. If \( \sigma \) is not known, select **1-sample \( t \)** and enter the summary statistics or enter \( C1 \) in the box located at the top right. (If \( \sigma \) is known, select **1-sample \( Z \)** and enter the summary statistics or enter \( C1 \) in the box located at the top right. Also enter the value of \( \sigma \) in the “Standard Deviation” or “Sigma” box.) Use the **Options** button to enter the confidence level, such as 95.0.

**EXCEL**  Use the Data Desk XL add-in that is a supplement to this book. Select **DDXL** and select **Confidence Intervals**. Under the Function Type options, select **1 Var \( t \) Interval** if \( \sigma \) is not known. (If \( \sigma \) is known, select **1 Var \( Z \) Interval**.) Click on the pencil icon and enter the range of data, such as A1:A12 if you have 12 values listed in column A. Click **OK**. In the dialog box, select the level of confidence. (If using 1 Var \( Z \) Interval, also enter the value of \( \sigma \).) Click on **Compute Interval** and the confidence interval will be displayed. (The use of Excel’s **CONFIDENCE** tool is not recommended, for a variety of reasons.)

continued
The TI-83/84 Plus calculator can be used to generate confidence intervals for original sample values stored in a list, or you can use the summary statistics \( n, \bar{x}, \) and \( \sigma. \) Either enter the data in list L1 or have the summary statistics available, then press the STAT key. Now select TESTS and choose TInterval if \( \sigma \) is not known. (Choose ZInterval if \( \sigma \) is known.)

After making the required entries, the calculator display will include the confidence interval in the format of \((\bar{x} - E, \bar{x} + E).\) For example, see the TI-83/84 Plus display that accompanies Example 5 in this section.

### Basic Skills and Concepts

#### Statistical Literacy and Critical Thinking

1. **What’s Wrong?** A “snapshot” in *USA Today* noted that “Consumers will spend an estimated average of $483 on merchandise” for back-to-school spending. It was reported that the value is based on a survey of 8453 consumers, and the margin of error is “\( \pm \)1 percentage point.” What’s wrong with this information?

2. **Robust** What does it mean when we say that the methods for constructing confidence intervals in this section are robust against departures from normality? Are the methods for constructing confidence intervals in this section robust against poor sampling methods?

3. **Sampling** A national polling organization has been hired to estimate the mean amount of cash carried by adults in the United States. The original sampling plan involved telephone calls placed to 2500 different telephone numbers throughout the United States, but a manager decides to save long-distance telephone expenses by using a simple random sample of 2500 telephone numbers that are all within the state of California. If this sample is used to construct a 95% confidence interval to estimate the population mean, will the estimate be good? Why or why not?

4. **Degrees of Freedom** A simple random sample of size \( n = 5 \) is obtained from the population of drivers living in New York City, and the braking reaction time of each driver is measured. The results are to be used for constructing a 95% confidence interval. What is the number of degrees of freedom that should be used for finding the critical value \( t_{0.05} \)? Give a brief explanation of the number of degrees of freedom.

#### Using Correct Distribution

In Exercises 5–12, assume that we want to construct a confidence interval using the given confidence level. Do one of the following, as appropriate: (a) Find the critical value \( z_{0.025} \) (b) find the critical value \( t_{0.025,n} \) (c) state that neither the normal nor the t distribution applies.

5. 95%; \( n = 23; \) \( \sigma \) is unknown; population appears to be normally distributed.
6. 99%; \( n = 25; \) \( \sigma \) is known; population appears to be normally distributed.
7. 99%; \( n = 6; \) \( \sigma \) is unknown; population appears to be very skewed.
8. 95%; \( n = 40; \) \( \sigma \) is unknown; population appears to be skewed.
9. 90%; \( n = 200; \) \( \sigma = 15.0; \) population appears to be skewed.
10. 95%; \( n = 9; \) \( \sigma \) is unknown; population appears to be very skewed.
11. 99%; \( n = 12; \) \( \sigma \) is unknown; population appears to be normally distributed.
12. 95%; \( n = 38; \) \( \sigma \) is unknown; population appears to be skewed.
Chapter 7
Estimates and Sample Sizes

Finding Confidence Intervals. In Exercises 13 and 14, use the given confidence level and sample data to find (a) the margin of error and (b) the confidence interval for the population mean \( \mu \). Assume that the sample is a simple random sample and the population has a normal distribution.

13. Hospital Costs 95% confidence; \( n = 20, \bar{x} = \$9004, s = \$569 \) (based on data from hospital costs for car crash victims who wore seat belts, from the U.S. Department of Transportation)

14. Car Pollution 99% confidence; \( n = 7, \bar{x} = 0.12, s = 0.04 \) (original values are nitrogen-oxide emissions in grams/mile, from the Environmental Protection Agency)

Interpreting Display. In Exercises 15 and 16, use the given data and the corresponding display to express the confidence interval in the format of \( \bar{x} - E < \mu < \bar{x} + E \). Also write a statement that interprets the confidence interval.

15. Weights of Dollar Coins 95% confidence; \( n = 20, \bar{x} = 8.0710 \text{ g}, s = 0.0411 \text{ g} \) (based on measurements made by the author). See the following SPSS display.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0516</td>
</tr>
<tr>
<td>95% Confidence Interval for Mean</td>
<td>0.0903</td>
</tr>
</tbody>
</table>

16. Weights of Plastic Discarded by Households 99% confidence; \( n = 62, \bar{x} = 1.911 \text{ lb}, s = 1.065 \text{ lb} \) (based on data from the Garbage Project, University of Arizona). See the TI-83/84 Plus calculator display in the margin.

Constructing Confidence Intervals. In Exercises 17–30, construct the confidence interval.

17. Garlic for Reducing Cholesterol In a test of the effectiveness of garlic for lowering cholesterol, 47 subjects were treated with Garlicin, which is garlic in a processed tablet form. Cholesterol levels were measured before and after the treatment. The changes in their levels of LDL cholesterol (in mg/dL) have a mean of 3.2 and a standard deviation of 18.6 (based on data from “Effect of Raw Garlic vs Commercial Garlic Supplements on Plasma Lipid Concentrations in Adults With Moderate Hypercholesterolemia,” by Gardner et al., Archives of Internal Medicine, Vol. 167).

a. What is the best point estimate of the population mean net change in LDL cholesterol after the Garlicin treatment?

b. Construct a 95% confidence interval estimate of the mean net change in LDL cholesterol after the Garlicin treatment. What does the confidence interval suggest about the effectiveness of Garlicin in reducing LDL cholesterol?

18. Birth Weights A random sample of the birth weights of 186 babies has a mean of 3103 g and a standard deviation of 696 g (based on data from “Cognitive Outcomes of Preschool Children with Prenatal Cocaine Exposure,” by Singer et al., Journal of the American Medical Association, Vol. 291, No. 20). These babies were born to mothers who did not use cocaine during their pregnancies.

a. What is the best point estimate of the mean weight of babies born to mothers who did not use cocaine during their pregnancies?

b. Construct a 95% confidence interval estimate of the mean birth weight for all such babies.

c. Compare the confidence interval from part (b) to this confidence interval obtained from birth weights of babies born to mothers who used cocaine during pregnancy: \( 2608 \text{ g} < \mu < 2792 \text{ g} \). Does cocaine use appear to affect the birth weight of a baby?
19. **Mean Body Temperature** Data Set 2 in Appendix B includes 106 body temperatures for which \( \bar{x} = 98.20^\circ F \) and \( s = 0.62^\circ F \).

a. What is the best point estimate of the mean body temperature of all healthy humans?

b. Using the sample statistics, construct a 99% confidence interval estimate of the mean body temperature of all healthy humans. Do the confidence interval limits contain 98.6°F? What does the sample suggest about the use of 98.6°F as the mean body temperature?

20. **Atkins Weight Loss Program** In a test of the Atkins weight loss program, 40 individuals participated in a randomized trial with overweight adults. After 12 months, the mean weight loss was found to be 2.1 lb, with a standard deviation of 4.8 lb.

a. What is the best point estimate of the mean weight loss of all overweight adults who follow the Atkins program?

b. Construct a 99% confidence interval estimate of the mean weight loss for all such subjects.

c. Does the Atkins program appear to be effective? Is it practical?

21. **Echinacea Treatment** In a study designed to test the effectiveness of echinacea for treating upper respiratory tract infections in children, 337 children were treated with echinacea and 370 other children were given a placebo. The numbers of days of peak severity of symptoms for the echinacea treatment group had a mean of 6.0 days and a standard deviation of 2.3 days. The numbers of days of peak severity of symptoms for the placebo group had a mean of 6.1 days and a standard deviation of 2.4 days (based on data from "Efficacy and Safety of Echinacea in Treating Upper Respiratory Tract Infections in Children," by Taylor et al., Journal of the American Medical Association, Vol. 290, No. 21).

a. Construct the 95% confidence interval for the mean number of days of peak severity of symptoms for those who receive echinacea treatment.

b. Construct the 95% confidence interval for the mean number of days of peak severity of symptoms for those who are given a placebo.

c. Compare the two confidence intervals. What do the results suggest about the effectiveness of echinacea?

22. **Acupuncture for Migraines** In a study designed to test the effectiveness of acupuncture for treating migraine, 142 subjects were treated with acupuncture and 80 subjects were given a sham treatment. The numbers of migraine attacks for the acupuncture treatment group had a mean of 1.8 and a standard deviation of 1.4. The numbers of migraine attacks for the sham treatment group had a mean of 1.6 and a standard deviation of 1.2.

a. Construct the 95% confidence interval estimate of the mean number of migraine attacks for those treated with acupuncture.

b. Construct the 95% confidence interval estimate of the mean number of migraine attacks for those given a sham treatment.

c. Compare the two confidence intervals. What do the results suggest about the effectiveness of acupuncture?

23. **Magnets for Treating Back Pain** In a study designed to test the effectiveness of magnets for treating back pain, 20 patients were given a treatment with magnets and also a sham treatment without magnets. Pain was measured using a standard Visual Analog Scale (VAS). After given the magnet treatments, the 20 patients had VAS scores with a mean of 5.0 and a standard deviation of 2.4. After being given the sham treatments, the 20 patients had VAS scores with a mean of 4.7 and a standard deviation of 2.9.

a. Construct the 95% confidence interval estimate of the mean VAS score for patients given the magnet treatment.

b. Construct the 95% confidence interval estimate of the mean VAS score for patients given a sham treatment.

c. Compare the results. Does the treatment with magnets appear to be effective?
24. Ages of Oscar Winning Actresses and Actors

The ages of the 79 actresses at the time that they won Oscars for the Best Actress category have a mean of 35.8 years and a standard deviation of 11.3 years. The ages of the 79 actors at the time that they won Oscars for the category of Best Actor have a mean of 43.8 years and a standard deviation of 8.9 years. Assume that the samples are simple random samples.

a. Construct the 99% confidence interval estimate of the mean age of actresses at the time that they win Oscars for the Best Actress category.

b. Construct the 99% confidence interval estimate of the mean age of actors at the time that they win Oscars for the Best Actor category.

c. Compare the results.

25. Monitoring Lead in Air

Listed below are measured amounts of lead (in micrograms per cubic meter, or \( \mu g/m^3 \)) in the air. The Environmental Protection Agency (EPA) has established an air quality standard for lead of \( 1.5 \mu g/m^3 \). The measurements shown below were recorded at Building 5 of the World Trade Center site on different days immediately following the destruction caused by the terrorist attacks of September 11, 2001. After the collapse of the two World Trade Center buildings, there was considerable concern about the quality of the air. Use the given values to construct a 95% confidence interval estimate of the mean amount of lead in the air. Is there anything about this data set suggesting that the confidence interval might not be very good? Explain.

\[
\begin{align*}
5.40 & \quad 1.10 & \quad 0.42 & \quad 0.73 & \quad 0.48 & \quad 1.10 \\
5.40 & \quad 1.10 & \quad 0.42 & \quad 0.73 & \quad 0.48 & \quad 1.10
\end{align*}
\]

26. Estimating Car Pollution

In a sample of seven cars, each car was tested for nitrogen-oxide emissions (in grams per mile) and the following results were obtained: 0.06, 0.11, 0.16, 0.15, 0.14, 0.08, 0.15 (based on data from the EPA). Assuming that this sample is representative of the cars in use, construct a 98% confidence interval estimate of the mean amount of nitrogen-oxide emissions for all cars. If the EPA requires that nitrogen-oxide emissions be less than 0.165 g/mi, can we safely conclude that this requirement is being met?

27. TV Salaries

Listed below are the top 10 salaries (in millions of dollars) of television personalities in a recent year (listed in order for Letterman, Cowell, Sheindlin, Leno, Couric, Lauer, Sawyer, Viera, Sutherland, and Sheen, based on data from OK! magazine).

a. Use the sample data to construct the 95% confidence interval for the population mean.

b. Do the sample data represent a simple random sample of TV salaries?

c. What is the assumed population? Is the sample representative of the population?

d. Does the confidence interval make sense?

\[
\begin{align*}
38 & \quad 36 & \quad 35 & \quad 27 & \quad 15 & \quad 13 & \quad 12 & \quad 10 & \quad 9.6 & \quad 8.4
\end{align*}
\]

28. Movie Lengths

Listed below are 12 lengths (in minutes) of randomly selected movies from Data Set 9 in Appendix B.

a. Construct a 99% confidence interval estimate of the mean length of all movies.

b. Assuming that it takes 30 min to empty a theater after a movie, clean it, allow time for the next audience to enter, and show previews, what is the minimum time that a theater manager should plan between start times of movies, assuming that this time will be sufficient for typical movies?

\[
\begin{align*}
110 & \quad 96 & \quad 125 & \quad 94 & \quad 132 & \quad 120 & \quad 136 & \quad 154 & \quad 149 & \quad 94 & \quad 119 & \quad 132
\end{align*}
\]

29. Video Games

Twelve different video games showing substance use were observed and the duration times of game play (in seconds) are listed below (based on data from "Content and Ratings of Teen-Rated Video Games," by Haninger and Thompson, Journal of the American Medical Association, Vol. 291, No. 7). The design of the study justifies the assumption that the sample can be treated as a simple random sample. Use the sample data to construct a 95% confidence interval estimate of \( \mu \), the mean duration of game play.

\[
\begin{align*}
4049 & \quad 3884 & \quad 3859 & \quad 4027 & \quad 4318 & \quad 4813 & \quad 4657 & \quad 4033 & \quad 5004 & \quad 4823 & \quad 4334 & \quad 4317
\end{align*}
\]
30. **Ages of Presidents** Listed below are the ages of the Presidents of the United States at the times of their inaugurations. Construct a 99% confidence interval estimate of the mean age of presidents at the times of their inaugurations. What is the population? Does the confidence interval provide a good estimate of the population mean? Why or why not?

42 43 46 46 47 48 49 49 50 51 51 51 51 51 52 52 54 54 54 54 54 55
55 55 55 56 56 56 57 57 57 57 58 60 61 61 62 64 64 65 68 69

**Appendix B Data Sets. In Exercises 31 and 32, use the data sets from Appendix B.**

31. **Nicotine in Cigarettes** Refer to Data Set 4 in Appendix B and assume that the samples are simple random samples obtained from normally distributed populations.

a. Construct a 95% confidence interval estimate of the mean amount of nicotine in cigarettes that are king size, nonfiltered, nonmenthol, and non-light.

b. Construct a 95% confidence interval estimate of the mean amount of nicotine in cigarettes that are 100 mm, filtered, nonmenthol, and non-light.

c. Compare the results. Do filters on cigarettes appear to be effective?

32. **Pulse Rates** A physician wants to develop criteria for determining whether a patient’s pulse rate is atypical, and she wants to determine whether there are significant differences between males and females. Use the sample pulse rates in Data Set 1 from Appendix B.

a. Construct a 95% confidence interval estimate of the mean pulse rate for males.

b. Construct a 95% confidence interval estimate of the mean pulse rate for females.

c. Compare the preceding results. Can we conclude that the population means for males and females are different? Why or why not?

### 7-4 Beyond the Basics

33. **Effect of an Outlier** Use the sample data from Exercise 30 to find a 99% confidence interval estimate of the population mean, after changing the first age from 42 years to 422 years. This value is not realistic, but such an error can easily occur during a data entry process. Does the confidence interval change much when 42 years is changed to 422 years? Are confidence interval limits sensitive to outliers? How should you handle outliers when they are found in sample data sets that will be used for the construction of confidence intervals?

34. **Alternative Method** Figure 7-6 and Table 7-1 summarize the decisions made when choosing between the normal and t distributions. An alternative method included in some textbooks (but almost never used by professional statisticians and almost never included in professional journals) is based on this criterion: Substitute the sample standard deviation $s$ for $\sigma$ whenever $n > 30$, then proceed as if $\sigma$ is known. Using this alternative method, repeat Exercise 30. Compare the results to those found in Exercise 30, and comment on the implications of the change in the width of the confidence interval.

35. **Finite Population Correction Factor** If a simple random sample of size $n$ is selected without replacement from a finite population of size $N$, and the sample size is more than 5% of the population size ($n > 0.05N$), better results can be obtained by using the finite population correction factor, which involves multiplying the margin of error $E$ by $\sqrt{(N - n)/(N - 1)}$. For the sample of 100 weights of M&M candies in Data Set 18 from Appendix B, we get $\bar{x} = 0.8565$ g and $s = 0.0518$ g. First construct a 95% confidence interval estimate of $\mu$ assuming that the population is large, then construct a 95% confidence interval estimate of the mean weight of M&Ms in the full bag from which the sample was taken. The full bag has 465 M&Ms. Compare the results.
36. Confidence Interval for Sample of Size \( n = 1 \) When a manned NASA spacecraft lands on Mars, the astronauts encounter a single adult Martian, who is found to be 12.0 ft tall. It is reasonable to assume that the heights of all Martians are normally distributed.

a. The methods of this chapter require information about the variation of a variable. If only one sample value is available, can it give us any information about the variation of the variable?

b. Based on the article “An Effective Confidence Interval for the Mean with Samples of Size One and Two,” by Wall, Boen, and Tweedie (American Statistician, Vol. 55, No. 2), a 95% confidence interval for \( \mu \) can be found (using methods not discussed in this book) for a sample of size \( n = 1 \) randomly selected from a normally distributed population, and it can be expressed as \( x \pm 9.68\sigma \). Use this result to construct a 95% confidence interval using the single sample value of 12.0 ft, and express it in the format of \( x - E < \mu < x + E \). Based on the result, is it likely that some other randomly selected Martian might be 50 ft tall?

7-5 Estimating a Population Variance

Key Concept In this section we introduce the chi-square probability distribution so that we can construct confidence interval estimates of a population standard deviation or variance. We also present a method for determining the sample size required to estimate a population standard deviation or variance.

When we considered estimates of proportions and means, we used the normal and Student \( t \) distributions. When developing estimates of variances or standard deviations, we use another distribution, referred to as the chi-square distribution. We will examine important features of that distribution before proceeding with the development of confidence intervals.

Chi-Square Distribution

In a normally distributed population with variance \( \sigma^2 \), assume that we randomly select independent samples of size \( n \) and, for each sample, compute the sample variance \( s^2 \) (which is the square of the sample standard deviation \( s \)). The sample statistic \( \chi^2 = (n - 1)s^2/\sigma^2 \) has a sampling distribution called the chi-square distribution.

**Chi-Square Distribution**

**Formula 7-5**

\[
\chi^2 = \frac{(n - 1)s^2}{\sigma^2}
\]

Where

- \( n \) = number of sample values
- \( s^2 \) = sample variance
- \( \sigma^2 \) = population variance

We denote chi-square by \( \chi^2 \), pronounced “kigh square.” To find critical values of the chi-square distribution, refer to Table A-4. The chi-square distribution is determined by the number of degrees of freedom, and in this chapter we use \( n - 1 \) degrees of freedom.

\[
\text{degrees of freedom} = n - 1
\]
In later chapters we will encounter situations in which the degrees of freedom are not \( n - 1 \), so we should not make the incorrect generalization that the number of degrees of freedom is always \( n - 1 \).

**Properties of the Chi-Square Distribution**

1. The chi-square distribution is not symmetric, unlike the normal and Student \( t \) distributions (see Figure 7-8). (As the number of degrees of freedom increases, the distribution becomes more symmetric, as Figure 7-9 illustrates.)

2. The values of chi-square can be zero or positive, but they cannot be negative (see Figure 7-8).

3. The chi-square distribution is different for each number of degrees of freedom (see Figure 7-9), and the number of degrees of freedom is given by \( df = n - 1 \). As the number of degrees of freedom increases, the chi-square distribution approaches a normal distribution.

**Push Polling**

"Push polling" is the practice of political campaigning under the guise of a poll. Its name is derived from its objective of pushing voters away from opposition candidates by asking loaded questions designed to discredit them. Here’s an example of one such question that was used: “Please tell me if you would be more likely or less likely to vote for Roy Romer if you knew that Gov. Romer appoints a parole board which has granted early release to an average of four convicted felons per day every day since Romer took office.” The National Council on Public Polls characterizes push polls as unethical, but some professional pollsters do not condemn the practice as long as the questions do not include outright lies.
Table A-2 for the standard normal distribution provides cumulative areas from the left, but Table A-4 for the chi-square distribution provides cumulative areas from the right.

**Example 1** Finding Critical Values of $\chi^2$ A simple random sample of ten voltage levels is obtained. Construction of a confidence interval for the population standard deviation $\sigma$ requires the left and right critical values of $\chi^2$ corresponding to a confidence level of 95% and a sample size of $n = 10$. Find the critical value of $\chi^2$ separating an area of 0.025 in the left tail, and find the critical value of $\chi^2$ separating an area of 0.025 in the right tail.

**Solution** With a sample size of $n = 10$, the number of degrees of freedom is $df = n - 1 = 9$. See Figure 7-10.

![Figure 7-10 Critical Values of the Chi-Square Distribution](image)

If using Table A-4, the critical value to the right ($\chi^2_R = 19.023$) is obtained in a straightforward manner by locating 9 in the degrees-of-freedom column at the left and 0.025 across the top row. The critical value of $\chi^2_L = 2.700$ to the left once again corresponds to 9 in the degrees-of-freedom column, but we must locate 0.975 (found by subtracting 0.025 from 1) across the top row because the values in the top row are always areas to the right of the critical value. Refer to Figure 7-10 and see that the total area to the right of $\chi^2_L = 2.700$ is 0.975. Figure 7-10 shows that, for a sample of 10 values taken from a normally distributed population, the chi-square statistic $(n - 1)s^2/\sigma^2$ has a 0.95 probability of falling between the chi-square critical values of 2.700 and 19.023.

Instead of using Table A-4, technology (such as STATDISK, Excel, and Minitab) can be used to find critical values of $\chi^2$. A major advantage of technology is that it can be used for any number of degrees of freedom and any confidence level, not just the limited choices included in Table A-4.
When obtaining critical values of $\chi^2$ from Table A-4, note that the numbers of degrees of freedom are consecutive integers from 1 to 30, followed by 40, 50, 60, 70, 80, 90, and 100. When a number of degrees of freedom (such as 52) is not found in the table, you can usually use the closest critical value. For example, if the number of degrees of freedom is 52, refer to Table A-4 and use 50 degrees of freedom. (If the number of degrees of freedom is exactly midway between table values, such as 55, simply find the mean of the two $\chi^2$ values.) For numbers of degrees of freedom greater than 100, use the equation given in Exercise 27, or a more extensive table, or use technology.

Estimators of $\sigma^2$

In Section 6-4 we showed that sample variances $s^2$ tend to target (or center on) the value of the population variance $\sigma^2$, so we say that $s^2$ is an unbiased estimator of $\sigma^2$. That is, sample variances $s^2$ do not systematically tend to overestimate the value of $\sigma^2$, nor do they systematically tend to underestimate $\sigma^2$. Instead, they tend to target the value of $\sigma^2$ itself. Also, the values of $s^2$ tend to produce smaller errors by being closer to $\sigma^2$ than do other unbiased measures of variation. For these reasons, $s^2$ is generally used to estimate $\sigma^2$. (However, there are other estimators of $\sigma^2$ that could be considered better than $s^2$. For example, even though $(n - 1)s^2/(n + 1)$ is a biased estimator of $\sigma^2$, it has the desirable property of minimizing the mean of the squares of the errors and therefore has a better chance of being closer to $\sigma^2$. See Exercise 28.)

The sample variance $s^2$ is the best point estimate of the population variance $\sigma^2$.

Because $s^2$ is an unbiased estimator of $\sigma^2$, we might expect that $s$ would be an unbiased estimator of $\sigma$, but this is not the case. (See Section 6-4.) If the sample size is large, however, the bias is small, so that we can use $s$ as a reasonably good estimate of $\sigma$. Even though it is a biased estimate, $s$ is often used as a point estimate of $\sigma$.

The sample standard deviation $s$ is commonly used as a point estimate of $\sigma$ (even though it is a biased estimate).

Although $s^2$ is the best point estimate of $\sigma^2$, there is no indication of how good it actually is. To compensate for that deficiency, we develop an interval estimate (or confidence interval) that gives us a range of values associated with a confidence level.

### Confidence Interval for Estimating a Population Standard Deviation or Variance

**Objective**

Construct a confidence interval used to estimate a population standard deviation or variance.

**Notation**

- $\sigma$ = population standard deviation
- $s$ = sample standard deviation
- $n$ = number of sample values
- $s^2$ = sample variance
- $\sigma^2$ = population variance
- $E$ = margin of error
- $\chi^2_L$ = left-tailed critical value of $\chi^2$
- $\chi^2_R$ = right-tailed critical value of $\chi^2$

**continued**
Requirements

1. The sample is a simple random sample.

2. The population must have normally distributed values (even if the sample is large).

Confidence Interval for the Population Variance $\sigma^2$

\[
\frac{(n - 1)s^2}{\chi^2_{R}} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_{L}}
\]

Confidence Interval for the Population Standard Deviation $\sigma$

\[
\frac{(n - 1)s^2}{\chi^2_{R}} < \sigma < \frac{(n - 1)s^2}{\chi^2_{L}}
\]

Requirements

For the methods of this section, departures from normal distributions can lead to gross errors. Consequently, the requirement of a normal distribution is much stricter here than in earlier sections, and we should check the distribution of data by constructing histograms and normal quantile plots, as described in Section 6-7.

Procedure for Constructing a Confidence Interval for $\sigma$ or $\sigma^2$

1. Verify that the requirements are satisfied.

2. Using $n - 1$ degrees of freedom, refer to Table A-4 or use technology to find the critical values $\chi^2_{R}$ and $\chi^2_{L}$ that correspond to the desired confidence level.

3. Evaluate the upper and lower confidence interval limits using this format of the confidence interval:

\[
\frac{(n - 1)s^2}{\chi^2_{R}} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_{L}}
\]

4. If a confidence interval estimate of $\sigma$ is desired, take the square root of the upper and lower confidence interval limits and change $\sigma^2$ to $\sigma$.

5. Round the resulting confidence interval limits. If using the original set of data, round to one more decimal place than is used for the original set of data. If using the sample standard deviation or variance, round the confidence interval limits to the same number of decimal places.

CAUTION

Confidence intervals can be used informally to compare the variation in different data sets, but the overlapping of confidence intervals should not be used for making formal and final conclusions about equality of variances or standard deviations.

Example 2: Confidence Interval for Home Voltage

The proper operation of typical home appliances requires voltage levels that do not vary much. Listed below are ten voltage levels (in volts) recorded in the author’s home on ten different days. (The voltages are from Data Set 13 in Appendix B.) These ten values have a standard deviation of $s = 0.15$ volt. Use the sample data to construct a 95% confidence interval estimate of the standard deviation of all voltage levels.

123.3 123.5 123.7 123.4 123.6 123.5 123.5 123.4 123.6 123.8
**SOLUTION**

**REQUIREMENT CHECK** We first verify that the requirements are satisfied. (1) The sample can be treated as a simple random sample. (2) The following display shows a Minitab-generated histogram, and the shape of the histogram is very close to the bell shape of a normal distribution, so the requirement of normality is satisfied. (This check of requirements is Step 1 in the process of finding a confidence interval of \( \sigma \), so we proceed next with Step 2.)

![Minitab Histogram](image)

**MINITAB**

**Step 2:** The sample size is \( n = 10 \), so the number of degrees of freedom is given by \( df = 10 - 1 = 9 \). If we use Table A-4, we refer to the row corresponding to 9 degrees of freedom, and we refer to the columns with areas of 0.975 and 0.025. (For a 95% confidence level, we divide \( \alpha = 0.05 \) equally between the two tails of the chi-square distribution, and we refer to the values of 0.975 and 0.025 across the top row of Table A-4.) The critical values are \( \chi^2_L = 2.700 \) and \( \chi^2_R = 19.023 \). (See Example 1.)

**Step 3:** Using the critical values of 2.700 and 19.023, the sample standard deviation of \( s = 0.15 \), and the sample size of \( n = 10 \), we construct the 95% confidence interval by evaluating the following:

\[
\frac{(n - 1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_L}
\]

\[
\frac{(10 - 1)(0.15)^2}{19.023} < \sigma^2 < \frac{(10 - 1)(0.15)^2}{2.700}
\]

**Step 4:** Evaluating the above expression results in \( 0.010645 < \sigma^2 < 0.075000 \). Finding the square root of each part (before rounding), then rounding to two decimal places, yields this 95% confidence interval estimate of the population standard deviation: \( 0.10 \) volt < \( \sigma < 0.27 \) volt.

**INTERPRETATION** Based on this result, we have 95% confidence that the limits of 0.10 volt and 0.27 volt contain the true value of \( \sigma \). The confidence interval can also be expressed as \( (0.10, 0.27) \), but the format of \( s \pm E \) cannot be used because the confidence interval does not have \( s \) at its center.

**Rationale for the Confidence Interval** We now explain why the confidence intervals for \( \sigma \) and \( \sigma^2 \) have the forms just given. If we obtain simple random samples
Chapter 7  Estimates and Sample Sizes

of size \( n \) from a population with variance \( \sigma^2 \), there is a probability of \( 1 - \alpha \) that the statistic \( (n - 1)s^2/\sigma^2 \) will fall between the critical values of \( \chi^2_R \) and \( \chi^2_L \). In other words (and symbols), there is a \( 1 - \alpha \) probability that both of the following are true:

\[
\frac{(n - 1)s^2}{\sigma^2} < \chi^2_R \quad \text{and} \quad \frac{(n - 1)s^2}{\sigma^2} > \chi^2_L
\]

If we multiply both of the preceding inequalities by \( \sigma^2 \) and divide each inequality by the appropriate critical value of \( \chi^2 \), we see that the two inequalities can be expressed in the equivalent forms:

\[
\frac{(n - 1)s^2}{\chi^2_R} < \sigma^2 \quad \text{and} \quad \frac{(n - 1)s^2}{\chi^2_L} > \sigma^2
\]

These last two inequalities can be combined into one inequality:

\[
\frac{(n - 1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_L}
\]

**Determining Sample Size**  The procedures for finding the sample size necessary to estimate \( \sigma^2 \) are much more complex than the procedures given earlier for means and proportions. Instead of using very complicated procedures, we will use Table 7-2. STATDISK also provides sample sizes. With STATDISK, select **Analysis, Sample Size Determination**, and then **Estimate St Dev**. Minitab, Excel, and the TI-83/84 Plus calculator do not provide such sample sizes.

<table>
<thead>
<tr>
<th>Sample Size for ( \sigma^2 )</th>
<th>Sample Size for ( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>To be 95% confident that ( s^2 ) is within of the value of ( \sigma^2 ), the sample size ( n ) should be at least</td>
<td>To be 95% confident that ( s ) is within of the value of ( \sigma ), the sample size ( n ) should be at least</td>
</tr>
<tr>
<td>1%</td>
<td>77,208</td>
</tr>
<tr>
<td>5%</td>
<td>3,149</td>
</tr>
<tr>
<td>10%</td>
<td>806</td>
</tr>
<tr>
<td>20%</td>
<td>211</td>
</tr>
<tr>
<td>30%</td>
<td>98</td>
</tr>
<tr>
<td>40%</td>
<td>57</td>
</tr>
<tr>
<td>50%</td>
<td>38</td>
</tr>
<tr>
<td>To be 99% confident that ( s^2 ) is within of the value of ( \sigma^2 ), the sample size ( n ) should be at least</td>
<td>To be 99% confident that ( s ) is within of the value of ( \sigma ), the sample size ( n ) should be at least</td>
</tr>
<tr>
<td>1%</td>
<td>133,449</td>
</tr>
<tr>
<td>5%</td>
<td>5,458</td>
</tr>
<tr>
<td>10%</td>
<td>1,402</td>
</tr>
<tr>
<td>20%</td>
<td>369</td>
</tr>
<tr>
<td>30%</td>
<td>172</td>
</tr>
<tr>
<td>40%</td>
<td>101</td>
</tr>
<tr>
<td>50%</td>
<td>68</td>
</tr>
</tbody>
</table>
Example 3: Finding Sample Size for Estimating \( \sigma \)

We want to estimate the standard deviation \( \sigma \) of all voltage levels in a home. We want to be 95% confident that our estimate is within 20% of the true value of \( \sigma \). How large should the sample be? Assume that the population is normally distributed.

Solution:

From Table 7-2, we can see that 95% confidence and an error of 20% for \( \sigma \) correspond to a sample of size 48. We should obtain a simple random sample of 48 voltage levels from the population of voltage levels.

For Confidence Intervals

**STATDISK** First obtain the descriptive statistics and verify that the distribution is normal by using a histogram or normal quantile plot. Next, select Analysis from the main menu, then select Confidence Intervals, and Population StDev. Enter the required data.

**MINITAB** Click on Stat, click on Basic Statistics, and select 1 Variance. Enter the column containing the list of sample data or enter the indicated summary statistics. Click on Options button and enter the confidence level, such as 95.0. Click OK twice. The results will include a standard confidence interval for the standard deviation and variance.

**EXCEL** Use DDXL. Select Confidence Intervals, then select the function type of Chi-square Confidence Ints for SD. Click on the pencil icon, and enter the range of cells with the sample data, such as A1:A10. Select a confidence level, then click OK.

**TI-83/84 PLUS** The TI-83/84 Plus calculator does not provide confidence intervals for \( \sigma \) or \( \sigma^2 \) directly, but the program S2INT can be used. That program was written by Michael Lloyd of Henderson State University, and it can be downloaded from www.aw.com/triola. The program S2INT uses the program ZZINEWT, so that program must also be installed. After storing the programs on the calculator, press the PRGM key, select S2INT, and enter the sample variance \( s^2 \), the sample size \( n \), and the confidence level (such as 0.95). Press the ENTER key, and wait a while for the display of the confidence interval limits for \( \sigma^2 \). Find the square root of the confidence interval limits if an estimate of \( \sigma \) is desired.

7-5 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Interpreting a Confidence Interval Using the weights of the M&M candies listed in Data Set 18 from Appendix B, we use the standard deviation of the sample \((s = 0.05179 \text{ g})\) to obtain the following 95% confidence interval estimate of the standard deviation of the weights of all M&Ms: \( 0.0455 < \sigma < 0.0602 \text{ g} \). Write a statement that correctly interprets that confidence interval.

2. Expressing Confidence Intervals Is the confidence interval given in Exercise 1 equivalent to the expression \((0.0455, 0.0602)\)? Is the confidence interval given in Exercise 1 equivalent to the expression \(0.05285 \pm 0.00735\text{ g}\)? Why or why not?

3. Valid Confidence Interval? A pollster for the Gallup Organization randomly generates the last two digits of telephone numbers to be called, so the numbers from 00 to 99 are all equally likely. Can the methods of this section be used to construct a confidence interval estimate of the standard deviation of the population of all outcomes? Why or why not?
4. Unbiased Estimators
What is an unbiased estimator? Is the sample variance an unbiased estimator of the population variance? Is the sample standard deviation an unbiased estimator of the population standard deviation?

Finding Critical Values. In Exercises 5–8, find the critical values \( \chi^2_L \) and \( \chi^2_R \) that correspond to the given confidence level and sample size.

5. 95%; \( n = 9 \)
6. 95%; \( n = 20 \)
7. 99%; \( n = 81 \)
8. 90%; \( n = 51 \)

Finding Confidence Intervals. In Exercises 9–12, use the given confidence level and sample data to find a confidence interval for the population standard deviation \( \sigma \). In each case, assume that a simple random sample has been selected from a population that has a normal distribution.

9. SAT Scores of College Students 95% confidence; \( n = 30, \bar{x} = 1533, s = 333 \)
10. Speeds of Drivers Ticketed in a 65 mi/h Zone on the Massachusetts Turnpike 95% confidence; \( n = 25, \bar{x} = 81.0 \text{ mi/h}, s = 2.3 \text{ mi/h} \)
11. White Blood Cell Count (in Cells per Microliter) 99% confidence; \( n = 7, \bar{x} = 7.106, s = 2.019 \)
12. Reaction Times of NASCAR Drivers 99% confidence; \( n = 8, \bar{x} = 1.24 \text{ sec}, s = 0.12 \text{ sec} \)

Determining Sample Size. In Exercises 13–16, assume that each sample is a simple random sample obtained from a normally distributed population. Use Table 7-2 on page 376 to find the indicated sample size.

13. Find the minimum sample size needed to be 95% confident that the sample standard deviation \( s \) is within 1% of \( \sigma \). Is this sample size practical in most applications?
14. Find the minimum sample size needed to be 95% confident that the sample standard deviation \( s \) is within 30% of \( \sigma \). Is this sample size practical in most applications?
15. Find the minimum sample size needed to be 99% confident that the sample variance is within 40% of the population variance. Is such a sample size practical in most cases?
16. Find the minimum sample size needed to be 95% confident that the sample variance is within 20% of the population variance.

Finding Confidence Intervals. In Exercises 17–24, assume that each sample is a simple random sample obtained from a population with a normal distribution.

17. Birth Weights In a study of the effects of prenatal cocaine use on infants, the following sample data were obtained for weights at birth: \( n = 190, \bar{x} = 2700 \text{ g}, s = 645 \text{ g} \) (based on data from “Cognitive Outcomes of Preschool Children with Prenatal Cocaine Exposure,” by Singer et al., Journal of the American Medical Association, Vol. 291, No. 20). Use the sample data to construct a 95% confidence interval estimate of the standard deviation of all birth weights of infants born to mothers who used cocaine during pregnancy. (Because Table A-4 has a maximum of 100 degrees of freedom while we require 189 degrees of freedom, use these critical values obtained from STATDISK: \( \chi^2_{0.025} = 152.8222 \) and \( \chi^2_{0.975} = 228.9638 \).) Based on the result, does the standard deviation appear to be different from the standard deviation of 696 g for birth weights of babies born to mothers who did not use cocaine during pregnancy?
18. Weights of M&Ms Data Set 18 in Appendix B lists 100 weights (in grams) of M&M candies. The minimum weight is 0.696 g and the maximum weight is 1.015 g.
   a. Use the range rule of thumb to estimate \( \sigma \), the standard deviation of weights of all such M&Ms.
   b. The 100 weights have a standard deviation of 0.0518 g. Construct a 95% confidence interval estimate of the standard deviation of weights of all M&Ms.
   c. Does the confidence interval from part (b) contain the estimated value of \( \sigma \) from part (a)? What do the results suggest about the estimate from part (a)?
19. **Movie Lengths** Data Set 9 in Appendix B includes 23 movies with ratings of PG or PG-13, and those movies have lengths (in minutes) with a mean of 120.8 min and a standard deviation of 22.9 min. That same data set also includes 12 movies with R ratings, and those movies have lengths with a mean of 118.1 min and a standard deviation of 20.8 min.

**a.** Construct a 95% confidence interval estimate of the standard deviation of the lengths of all movies with ratings of PG or PG-13.

**b.** Construct a 95% confidence interval estimate of the standard deviation of the lengths of all movies with ratings of R.

**c.** Compare the variation of the lengths of movies with ratings of PG or PG-13 to the variation of the lengths of movies with ratings of R. Does there appear to be a difference?

20. **Pulse Rates of Men and Women** Data Set 1 in Appendix B includes 40 pulse rates of men, and those pulse rates have a mean of 69.4 beats per minute and a standard deviation of 11.3 beats per minute. That data set also includes 40 pulse rates of women, and those pulse rates have a mean of 76.3 beats per minute and a standard deviation of 12.5 beats per minute.

**a.** Construct a 99% confidence interval estimate of the standard deviation of the pulse rates of men.

**b.** Construct a 99% confidence interval estimate of the standard deviation of the pulse rates of women.

**c.** Compare the variation of the pulse rates of men and women. Does there appear to be a difference?

21. **Video Games** Twelve different video games showing substance use were observed and the duration times of game play (in seconds) are listed below (based on data from “Content and Ratings of Teen-Rated Video Games,” by Haninger and Thompson, *Journal of the American Medical Association*, Vol. 291, No. 7). The design of the study justifies the assumption that the sample can be treated as a simple random sample. Use the sample data to construct a 99% confidence interval estimate of , the standard deviation of the duration times of game play.

   4049 3884 3859 4027 4318 4813 4657 4033 5004 4823 4334 4317

   849 807 821 859 864 877 772 848 802 807 887 815

22. **Designing Theater Seats** In the course of designing theater seats, the sitting heights (in mm) of a simple random sample of adult women is obtained, and the results are listed below (based on anthropometric survey data from Gordon, Churchill, et al.). Use the sample data to construct a 95% confidence interval estimate of , the standard deviation of sitting heights of all women. Does the confidence interval contain the value of 35 mm, which is believed to be the standard deviation of sitting heights of women?

   5.40 1.10 0.42 0.73 0.48 1.10

23. **Monitoring Lead in Air** Listed below are measured amounts of lead (in micrograms per cubic meter, or µg/m³) in the air. The EPA has established an air quality standard for lead of 1.5 µg/m³. The measurements shown below were recorded at Building 5 of the World Trade Center site on different days immediately following the destruction caused by the terrorist attacks of September 11, 2001. Use the given values to construct a 95% confidence interval estimate of the standard deviation of the amounts of lead in the air. Is there anything about this data set suggesting that the confidence interval might not be very good? Explain.

   5.40 1.10 0.42 0.73 0.48 1.10

24. **a. Comparing Waiting Lines** The listed values are waiting times (in minutes) of customers at the Jefferson Valley Bank, where customers enter a single waiting line that feeds three teller windows. Construct a 95% confidence interval for the population standard deviation .

   6.5 6.6 6.7 6.8 7.1 7.3 7.4 7.7 7.7 7.7
Chapter 7  Estimates and Sample Sizes

27. Finding Critical Values

In constructing confidence intervals for $\sigma$ or $\sigma^2$, we use Table A-4 to find the critical values $\chi^2_\alpha$ and $\chi^2_{1-\alpha}$, but that table applies only to cases in which $n \leq 101$, so the number of degrees of freedom is 100 or smaller. For larger numbers of degrees of freedom, we can approximate $\chi^2_{1-\alpha}$ by using

$$\chi^2 = \frac{1}{2} \left[ \pm z_{\alpha/2} + \sqrt{2k - 1} \right]^2$$

where $k$ is the number of degrees of freedom and $z_{\alpha/2}$ is the critical $z$ score described in Section 7-2. STATDISK was used to find critical values for 189 degrees of freedom with a confidence level of 95%, and those critical values are given in Exercise 17. Use the approximation shown here to find the critical values and compare the results to those found from STATDISK.

28. Finding the Best Estimator

Values of $\hat{s}^2$ tend to produce smaller errors by being closer to $\sigma^2$ than do other unbiased measures of variation. Let's now consider the biased estimator of $(n - 1)s^2/(n + 1)$. Given the population of values [2, 3, 7], use the value of $\sigma^2$, and use the nine different possible samples of size $n = 2$ (for sampling with replacement) for the following.

a. Find $s^2$ for each of the nine samples, then find the error $s^2 - \sigma^2$ for each sample. Then square those errors. Then find the mean of those squares. The result is the value of the mean square error.

b. Find the value of $(n - 1)s^2/(n + 1)$ for each of the nine samples. Then find the error $(n - 1)s^2/(n + 1) - \sigma^2$ for each sample. Square those errors, then find the mean of those squares. The result is the mean square error.

c. The mean square error can be used to measure how close an estimator comes to the population parameter. Which estimator does a better job by producing the smaller mean square error? Is that estimator biased or unbiased?

b. The listed values are waiting times (in minutes) of customers at the Bank of Providence, where customers may enter any one of three different lines that have formed at three teller windows. Construct a 95% confidence interval for the population standard deviation $\sigma$.

4.2  5.4  5.8  6.2  6.7  7.7  7.7  8.5  9.3  10.0

c. Interpret the results found in parts (a) and (b). Do the confidence intervals suggest a difference in the variation among waiting times? Which arrangement seems better: the single-line system or the multiple-line system?

Using Large Data Sets from Appendix B. In Exercises 25 and 26, use the data set from Appendix B. Assume that each sample is a simple random sample obtained from a population with a normal distribution.

25. FICO Credit Rating Scores

Refer to Data Set 24 in Appendix B and use the credit rating scores to construct a 95% confidence interval estimate of the standard deviation of all credit rating scores.

26. Home Energy Consumption

Refer to Data Set 12 in Appendix B and use the sample amounts of home energy consumption (in kWh) to construct a 99% confidence interval estimate of the standard deviation of all energy consumption amounts.
Review

In this chapter we introduced basic methods for finding estimates of population proportions, means, and variances. This chapter included procedures for finding each of the following:

- point estimate
- confidence interval
- required sample size

We discussed the point estimate (or single-valued estimate) and formed these conclusions:

- Proportion: The best point estimate of $p$ is $\hat{p}$.
- Mean: The best point estimate of $\mu$ is $\bar{x}$.
- Variation: The value of $s$ is commonly used as a point estimate of $\sigma$, even though it is a biased estimate. Also, $s^2$ is the best point estimate of $\sigma^2$.

Because the above point estimates consist of single values, they have the serious shortcoming of not revealing how close to the population parameter that they are likely to be, so confidence intervals (or interval estimates) are commonly used as more informative and useful estimates. We also considered ways of determining the sample sizes necessary to estimate parameters to within given margins of error. This chapter also introduced the Student $t$ and chi-square distributions. We must be careful to use the correct probability distribution for each set of circumstances. This chapter used the following criteria for selecting the appropriate distribution:

Confidence interval for proportion $p$: Use the normal distribution (assuming that the required conditions are satisfied and there are at least 5 successes and at least 5 failures so that the normal distribution can be used to approximate the binomial distribution).

Confidence interval for $\mu$: See Figure 7-6 (page 360) or Table 7-1 (page 361) to choose between the normal or $t$ distributions (or conclude that neither applies).

Confidence interval for $\sigma$ or $\sigma^2$: Use the chi-square distribution (assuming that the required conditions are satisfied).

For the confidence interval and sample size procedures in this chapter, it is very important to verify that the requirements are satisfied. If they are not, then we cannot use the methods of this chapter and we may need to use other methods, such as the bootstrap method described in the Technology Project at the end of this chapter, or nonparametric methods, such as those discussed in Chapter 13.

Statistical Literacy and Critical Thinking

1. Estimating Population Parameters Quest Diagnostics is a provider of drug testing for job applicants, and its managers want to estimate the proportion of job applicants who test positive for drugs. In this context, what is a point estimate of that proportion? What is a confidence interval? What is a major advantage of the confidence interval estimate over the point estimate?

2. Interpreting a Confidence Interval Here is a 95% confidence interval estimate of the proportion of all job applicants who test positive when they are tested for drug use: $0.0262 < p < 0.0499$ (based on data from Quest Diagnostics). Write a statement that correctly interprets this confidence interval.

3. Confidence Level What is the confidence level of the confidence interval given in Exercise 2? What is a confidence level in general?
Chapter 7  Estimates and Sample Sizes

1. **Reporting Income**  In a Pew Research Center poll of 745 randomly selected adults, 589 said that it is morally wrong to not report all income on tax returns. Construct a 95% confidence interval estimate of the percentage of all adults who have that belief, and then write a statement interpreting the confidence interval.

2. **Determining Sample Size**  See the survey described in Exercise 1. Assume that you must conduct a new poll to determine the percentage of adults who believe that it is morally wrong to not report all income on tax returns. How many randomly selected adults must you survey if you want 99% confidence that the margin of error is two percentage points? Assume that nothing is known about the percentage that you are trying to estimate.

3. **Determining Sample Size**  See the survey described in Exercise 1. Assume that you must conduct a survey to determine the mean income reported on tax returns, and you have access to actual tax returns. How many randomly selected tax returns must you survey if you want to

---

**Chapter Quick Quiz**

1. The following 95% confidence interval estimate is obtained for a population mean: $10.0 < \mu < 20.0$. Interpret that confidence interval.

2. With a Democrat and a Republican candidate running for office, a newspaper conducts a poll to determine the proportion of voters who favor the Republican candidate. Based on the poll results, this 95% confidence interval estimate of that proportion is obtained: $0.492 < p < 0.588$. Which of the following statements better describes the results: (1) The Republican is favored by a majority of the voters. (2) The election is too close to call.

3. Find the critical value of $t_{n/2}$ for $n = 20$ and $\alpha = 0.05$.

4. Find the critical value of $z_{\alpha/2}$ for $n = 20$ and $\alpha = 0.10$.

5. Find the sample size required to estimate the percentage of college students who use loans to help fund their tuition. Assume that we want 95% confidence that the proportion from the sample is within two percentage points of the true population percentage.

6. In a poll of 600 randomly selected subjects, 240 answered “yes” when asked if they planned to vote in a state election. What is the best point estimate of the population proportion of all who plan to vote in that election.

7. In a poll of 600 randomly selected subjects, 240 answered “yes” when asked if they planned to vote in a state election. Construct a 95% confidence interval estimate of the proportion of all who plan to vote in that election.

8. In a survey of randomly selected subjects, the mean age of the 36 respondents is 40.0 years and the standard deviation of the ages is 10.0 years. Use these sample results to construct a 95% confidence interval estimate of the mean age of the population from which the sample was selected.

9. Repeat Exercise 8 assuming that the population standard deviation is known to be 10.0 years.

10. Find the sample size required to estimate the mean age of registered drivers in the United States. Assume that we want 95% confidence that the sample mean is within 1/2 year of the true mean age of the population. Also assume that the standard deviation of the population is known to be 12 years.

---

**Review Exercises**

1. **Reporting Income**  In a Pew Research Center poll of 745 randomly selected adults, 589 said that it is morally wrong to not report all income on tax returns. Construct a 95% confidence interval estimate of the percentage of all adults who have that belief, and then write a statement interpreting the confidence interval.

2. **Determining Sample Size**  See the survey described in Exercise 1. Assume that you must conduct a new poll to determine the percentage of adults who believe that it is morally wrong to not report all income on tax returns. How many randomly selected adults must you survey if you want 99% confidence that the margin of error is two percentage points? Assume that nothing is known about the percentage that you are trying to estimate.

3. **Determining Sample Size**  See the survey described in Exercise 1. Assume that you must conduct a survey to determine the mean income reported on tax returns, and you have access to actual tax returns. How many randomly selected tax returns must you survey if you want to
be 99% confident that the mean of the sample is within $500 of the true population mean? Assume that reported incomes have a standard deviation of $28,785 (based on data from the U.S. Census Bureau). Is the sample size practical?

4. Penny Weights A simple random sample of 37 weights of pennies made after 1983 has a mean of 2.4991 g and a standard deviation of 0.0165 g (based on Data Set 20 in Appendix B). Construct a 99% confidence interval estimate of the mean weight of all such pennies. Design specifications require a population mean of 2.5 g. What does the confidence interval suggest about the manufacturing process?

5. Crash Test Results The National Transportation Safety Administration conducted crash test experiments on five subcompact cars. The head injury data (in hic) recorded from crash test dummies in the driver's seat are as follows: 681, 428, 917, 898, 420. Use these sample results to construct a 95% confidence interval for the mean of head injury measurements from all subcompact cars.

6. Confidence Interval for \( \sigma \) New car design specifications are being considered to control the variation of the head injury measurements. Use the same sample data from Exercise 5 to construct a 95% confidence interval estimate of \( \sigma \).

7. Cloning Survey A Gallup poll consisted of 1012 randomly selected adults who were asked whether “cloning of humans should or should not be allowed.” Results showed that 901 adults surveyed indicated that cloning should not be allowed.

a. Find the best point estimate of the proportion of adults believing that cloning of humans should not be allowed.

b. Construct a 95% confidence interval estimate of the proportion of adults believing that cloning of humans should not be allowed.

c. A news reporter wants to determine whether these survey results constitute strong evidence that the majority (more than 50%) of people are opposed to such cloning. Based on the results, is there strong evidence supporting the claim that the majority is opposed to such cloning? Why or why not?

8. Sample Size You have been hired by a consortium of local car dealers to conduct a survey about the purchases of new and used cars.

a. If you want to estimate the percentage of car owners in your state who purchased new cars (not used), how many adults must you survey if you want 95% confidence that your sample percentage is in error by no more than four percentage points?

b. If you want to estimate the mean amount of money spent by car owners on their last car purchase, how many car owners must you survey if you want 95% confidence that your sample mean is in error by no more than $750? (Based on results from a pilot study, assume that the standard deviation of amounts spent on car purchases is $14,227.)

c. If you plan to obtain the estimates described in parts (a) and (b) with a single survey having several questions, how many people must be surveyed?

9. Discarded Glass Listed below are weights (in pounds) of glass discarded in one week by randomly selected households (based on data from the Garbage Project at the University of Arizona).

a. What is the best point estimate of the weight of glass discarded by all households in one week?

b. Construct a 95% confidence interval estimate of the mean weight of glass discarded by all households.

c. Repeat part (b) assuming that the population is normally distributed with a standard deviation known to be 3.108 lb.

\[
\begin{align*}
3.52 & \quad 8.87 & \quad 3.99 & \quad 3.61 & \quad 2.33 & \quad 3.21 & \quad 0.25 & \quad 4.94 \\
\end{align*}
\]

10. Confidence Intervals for \( \sigma \) and \( \sigma^2 \)

a. Use the sample data from Exercise 9 to construct a 95% confidence interval estimate of the population standard deviation.

b. Use the sample data from Exercise 9 to construct a 95% confidence interval estimate of the population variance.
Cumulative Review Exercises

Weights of Supermodels. Supermodels are sometimes criticized on the grounds that their low weights encourage unhealthy eating habits among young women. In Exercises 1–4, use the following weights (in pounds) of randomly selected supermodels.

125 (Taylor) 119 (Auermann) 128 (Schiffer) 125 (Bundchen) 119 (Turlington)
127 (Hall) 105 (Moss) 123 (Mazza) 110 (Reilly) 103 (Barton)

1. Find the mean, median, and standard deviation.
2. What is the level of measurement of these data (nominal, ordinal, interval, ratio)?
3. Construct a 95% confidence interval for the population mean.
4. Find the sample size necessary to estimate the mean weight of all supermodels so that there is 95% confidence that the sample mean is in error by no more than 2 lb. Assume that a pilot study suggests that the weights of all supermodels have a standard deviation of 7.5 lb.

5. Employment Drug Test If a randomly selected job applicant is given a drug test, there is a 0.038 probability that the applicant will test positive for drug use (based on data from Quest Diagnostics).
   a. If a job applicant is randomly selected and given a drug test, what is the probability that the applicant does not test positive for drug use?
   b. Find the probability that when two different job applicants are randomly selected and given drug tests, they both test positive for drugs.
   c. If 500 job applicants are randomly selected and they are all given drug tests, find the probability that at least 20 of them test positive for drugs.

6. ACT Scores Scores on the ACT test are normally distributed with a mean of 21.1 and a standard deviation of 4.8.
   a. If one ACT score is randomly selected, find the probability that it is greater than 20.0.
   b. If 25 ACT scores are randomly selected, find the probability that they have a mean greater than 20.
   c. Find the ACT score that is the 90th percentile.

7. Sampling What is a simple random sample? What is a voluntary response sample?

8. Range Rule of Thumb Use the range rule of thumb to estimate the standard deviation of grade point averages at a college with a grading system designed so that the lowest and highest possible grade point averages are 0 and 4.

9. Rare Event Rule Find the probability of making random guesses to 12 true/false questions and getting 12 correct answers. If someone did get 12 correct answers, is it possible that they made random guesses? Is it likely that they made random guesses?

10. Sampling Method If you conduct a poll by surveying all of your friends that you see during the next week, which of the following terms best describes the type of sampling used: random, systematic, cluster, convenience, voluntary response? Is the sample likely to be representative of the population?

Technology Project

Bootstrap Resampling The bootstrap resampling method can be used to construct confidence intervals for situations in which traditional methods cannot (or should not) be used. Example 4 in Section 7-4 included the following sample of times that different video games showed the use of alcohol (based on data from "Content and Ratings of Teen-Rated
Example 4 in Section 7-4 showed the histogram and normal quantile plot, and they both suggest that the times are not from a normally distributed population, so methods requiring a normal distribution should not be used.

If we want to use the above sample data for the construction of a confidence interval estimate of the population mean $\mu$, one approach is to use the bootstrap resampling method, which has no requirements about the distribution of the population. This method typically requires a computer to build a bootstrap population by replicating (duplicating) a sample many times. We draw from the sample with replacement, thereby creating an approximation of the original population. In this way, we pull the sample up "by its own bootstraps" to simulate the original population. Using the sample data given above, construct a 95% confidence interval estimate of the population mean $\mu$ by using the bootstrap method.

Various technologies can be used for this procedure. The STATDISK statistical software program that is on the CD included with this book is very easy to use. Enter the listed sample values in column 1 of the Data Window, then select the main menu item of Analysis, and select the menu item of Bootstrap Resampling.

a. Create 500 new samples, each of size 12, by selecting 12 values with replacement from the 12 sample values given above. In STATDISK, enter 500 for the number of resamplings and click on Resample.

b. Find the means of the 500 bootstrap samples generated in part (a). In STATDISK, the means will be listed in the second column of the Data Window.

c. Sort the 500 means (arrange them in order). In STATDISK, click on the Data Tools button and sort the means in column 2.

d. Find the percentiles $P_{2.5}$ and $P_{97.5}$ for the sorted means that result from the preceding step. ($P_{2.5}$ is the mean of the 12th and 13th scores in the sorted list of means; $P_{97.5}$ is the mean of the 487th and 488th scores in the sorted list of means.) Identify the resulting confidence interval by substituting the values for $P_{2.5}$ and $P_{97.5}$ in $P_{2.5} < \mu < P_{97.5}$.

There is a special software package designed specifically for bootstrap resampling methods: Resampling Stats, available from Resampling Stats, Inc., 612 N. Jackson St., Arlington, VA, 22201; telephone number: (703) 522-2713.

---

**Internet Project**

Go to: [http://www.aw.com/triola](http://www.aw.com/triola)

The confidence intervals in this chapter illustrate an important point in the science of statistical estimation. Namely, estimations based on sample data are made with certain degrees of confidence. In the Internet Project for this chapter, you will use confidence intervals to make a statement about the temperature where you live.

After going to this book’s Web site, locate the project for this chapter. There you will find instructions on how to use the Internet to find temperature data collected by the weather station nearest your home. With this data in hand, you will construct confidence intervals for temperatures during different time periods and attempt to draw conclusions about temperature change in your area. In addition, you will learn more about the relationship between confidence and probability.
Cooperative Group Activities

1. **Out-of-class activity** Collect sample data, and use the methods of this chapter to construct confidence interval estimates of population parameters. Here are some suggestions for parameters:

   • Proportion of students at your college who can raise one eyebrow without raising the other eyebrow.

   • Mean age of cars driven by statistics students and/or the mean age of cars driven by faculty.
Cooperative Group Activities

- Mean length of words in *New York Times* editorials and mean length of words in editorials found in your local newspaper.
- Mean lengths of words in *Time* magazine, *Newsweek* magazine, and *People* magazine.
- Proportion of students at your college who can correctly identify the president, vice president, and secretary of state.
- Proportion of students at your college who are over the age of 18 and are registered to vote.
- Mean age of full-time students at your college.
- Proportion of motor vehicles in your region that are cars.

2. **In-class activity** Without using any measuring device, each student should draw a line believed to be 3 in. long and another line believed to be 3 cm long. Then use rulers to measure and record the lengths of the lines drawn. Find the means and standard deviations of the two sets of lengths. Use the sample data to construct a confidence interval for the length of the line estimated to be 3 in., then do the same for the length of the line estimated to be 3 cm. Do the confidence interval limits actually contain the correct length? Compare the results. Do the estimates of the 3-in. line appear to be more accurate than those for the 3-cm line?

3. **In-class activity** Assume that a method of gender selection can affect the probability of a baby being a girl, so that the probability becomes 1/4. Each student should simulate 20 births by drawing 20 cards from a shuffled deck. Replace each card after it has been drawn, then reshuffle. Consider the hearts to be girls and consider all other cards to be boys. After making 20 selections and recording the “genders” of the babies, construct a confidence interval estimate of the proportion of girls. Does the result appear to be effective in identifying the true value of the population proportion? (If decks of cards are not available, use some other way to simulate the births, such as using the random number generator on a calculator or using digits from phone numbers or social security numbers.)

4. **Out-of-class activity** Groups of three or four students should go to the library and collect a sample consisting of the ages of books (based on copyright dates). Plan and describe the sampling procedure, execute the sampling procedure, then use the results to construct a confidence interval estimate of the mean age of all books in the library.

5. **In-class activity** Each student should write an estimate of the age of the current President of the United States. All estimates should be collected and the sample mean and standard deviation should be calculated. Then use the sample results to construct a confidence interval. Do the confidence interval limits contain the correct age of the President?

6. **In-class activity** A class project should be designed to conduct a test in which each student is given a taste of Coke and a taste of Pepsi. The student is then asked to identify which sample is Coke. After all of the results are collected, analyze the claim that the success rate is better than the rate that would be expected with random guesses.

7. **In-class activity** Each student should estimate the length of the classroom. The values should be based on visual estimates, with no actual measurements being taken. After the estimates have been collected, construct a confidence interval, then measure the length of the room. Does the confidence interval contain the actual length of the classroom? Is there a “collective wisdom,” whereby the class mean is approximately equal to the actual room length?

8. **In-class activity** Divide into groups of three or four. Examine a current magazine such as *Time* or *Newsweek*, and find the proportion of pages that include advertising. Based on the results, construct a 95% confidence interval estimate of the percentage of all such pages that have advertising. Compare results with other groups.

9. **In-class activity** Divide into groups of two. First find the sample size required to estimate the proportion of times that a coin turns up heads when tossed, assuming that you want 80% confidence that the sample proportion is within 0.08 of the true population proportion. Then toss a coin the required number of times and record your results. What percentage of such confidence intervals should actually contain the true value of the population proportion,
which we know is $p = 0.5$? Verify this last result by comparing your confidence interval with the confidence intervals found in other groups.

10. **Out-of-class activity** Identify a topic of general interest and coordinate with all members of the class to conduct a survey. Instead of conducting a "scientific" survey using sound principles of random selection, use a convenience sample consisting of respondents that are readily available, such as friends, relatives, and other students. Analyze and interpret the results. Identify the population. Identify the shortcomings of using a convenience sample, and try to identify how a sample of subjects randomly selected from the population might be different.

11. **Out-of-class activity** Each student should find an article in a professional journal that includes a confidence interval of the type discussed in this chapter. Write a brief report describing the confidence interval and its role in the context of the article.

12. **Out-of-class activity** Obtain a sample and use it to estimate the mean number of hours per week that students at your college devote to studying.
Artem Boytsov is a Senior Software Engineer at Google, Inc. Founded in 1998 by Stanford University students Larry Page and Sergey Brin, Google has become the most widely used Internet search engine. It has become so popular that the word google is now used as a verb in everyday language. To “google” is to use the Google search engine for finding information on the Internet.

Q: What concepts of statistics do you use in your work for Google?
A: The statistics tools I use include sampling, normal distribution, Zipf/Pareto distributions, standard deviations, standard errors, correlations, confidence intervals, and conditional probabilities.

Q: How do you use statistics at Google?
A: Statistics and probability theory lie in the core of Google Trends. We use statistics to analyze our users’ search behavior. We measure the popularity of terms on the Web to identify factors such as seasonal patterns and correlations between queries.

Q: Please describe at least one specific example illustrating how the use of statistics was successful in improving a product or service.
A: Knowledge of normal distributions, confidence intervals, and error concepts are crucial in analyzing the data. Performance and data quality trade-offs are calculated and made in our product. Further performance improvements were implemented by dismissing insignificant data and exploiting its properties (for example, that query frequencies follow the Zipf distribution).

Q: Is your use of probability and statistics increasing, decreasing, or remaining stable?
A: It is increasing with every feature or internal application of our product.

Q: Please describe how you try to ensure objectivity.
A: We use uniformly selected samples of Google’s query log to calculate the probability that a given keyword is present in a random user query. We use conditional probabilities to factor out such factors as Google’s overall traffic growth, differences in traffic volume between countries, etc.

Q: How critical do you find your knowledge of statistics for performing your responsibilities?
A: Knowledge of probability theory and statistics is crucial to my job as a Google Trends engineer. Many different projects in Google use statistics, probability, and information theories.

Q: Please cite an example of how your data are used.
A: A merchant may use Google Trends to plan for fluctuations in seasonal demand. General trend (upward/downward) in popularity helps predict and prepare for such demand changes.

Q: In terms of statistics, what would you recommend for prospective employees?
A: I recommend an introductory course in statistics to anyone. To those pursuing degrees in engineering or finance, I would recommend courses in statistics, probability theory, and information theory. There’s a world of information out there, and statistics is your first step to understanding and utilizing this information.

Q: Do you recommend statistics for today’s college students?
A: Yes. The ability to analyze data is crucial in the era of information. I face misinterpretation and misuse of data by “statistically illiterate” people quite often, and it’s very sad.