CHAPTER 7

Confidence Intervals and Sample Size

Objectives

After completing this chapter, you should be able to

1. Find the confidence interval for the mean when σ is known.
2. Determine the minimum sample size for finding a confidence interval for the mean.
3. Find the confidence interval for the mean when σ is unknown.
4. Find the confidence interval for a proportion.
5. Determine the minimum sample size for finding a confidence interval for a proportion.
6. Find a confidence interval for a variance and a standard deviation.

Outline

Statistics Today

Would You Change the Channel?

A survey by the Roper Organization found that 45% of the people who were offended by a television program would change the channel, while 15% would turn off their television sets. The survey further stated that the margin of error is 3 percentage points, and 4000 adults were interviewed.

Several questions arise:

1. How do these estimates compare with the true population percentages?
2. What is meant by a margin of error of 3 percentage points?
3. Is the sample of 4000 large enough to represent the population of all adults who watch television in the United States?

See Statistics Today—Revisited at the end of the chapter for the answers.

After reading this chapter, you will be able to answer these questions, since this chapter explains how statisticians can use statistics to make estimates of parameters.

Source: The Associated Press.

Introduction

One aspect of inferential statistics is **estimation,** which is the process of estimating the value of a parameter from information obtained from a sample. For example, *The Book of Odds,* by Michael D. Shook and Robert L. Shook (New York: Penguin Putnam, Inc.), contains the following statements:

"One out of 4 Americans is currently dieting." (Calorie Control Council)
"Seventy-two percent of Americans have flown on commercial airlines." ("The Bristol Meyers Report: Medicine in the Next Century")
"The average kindergarten student has seen more than 5000 hours of television." (U.S. Department of Education)
"The average school nurse makes $32,786 a year." (National Association of School Nurses)
"The average amount of life insurance is $108,000 per household with life insurance." (American Council of Life Insurance)

Since the populations from which these values were obtained are large, these values are only *estimates* of the true parameters and are derived from data collected from samples.

The statistical procedures for estimating the population mean, proportion, variance, and standard deviation will be explained in this chapter.

An important question in estimation is that of sample size. How large should the sample be in order to make an accurate estimate? This question is not easy to answer since the size of the sample depends on several factors, such as the accuracy desired and the probability of making a correct estimate. The question of sample size will be explained in this chapter also.

Inferential statistical techniques have various **assumptions** that must be met before valid conclusions can be obtained. One common assumption is that the samples must be randomly selected. Chapter 1 explains how to obtain a random sample. The other common assumption is that either the sample size must be greater than or equal to 30 or the population must be normally or approximately normally distributed if the sample size is less than 30.

To check this assumption, you can use the methods explained in Chapter 6. Just for review, the methods are to check the histogram to see if it is approximately bell-shaped, check for outliers, and if possible, generate a normal quartile plot and see if the points fall close to a straight line. (*Note:* An area of statistics called nonparametric statistics does not require the variable to be normally distributed.)

Some statistical techniques are called **robust.** This means that the distribution of the variable can depart somewhat from normality, and valid conclusions can still be obtained.

7–1

Confidence Intervals for the Mean When σ Is Known

Objective 1

Find the confidence interval for the mean when σ is known.

Suppose a college president wishes to estimate the average age of students attending classes this semester. The president could select a random sample of 100 students and find the average age of these students, say, 22.3 years. From the sample mean, the president could infer that the average age of all the students is 22.3 years. This type of estimate is called a *point estimate.*

A **point estimate** is a specific numerical value estimate of a parameter. The best point estimate of the population mean μ is the sample mean $\overline{X}$.

You might ask why other measures of central tendency, such as the median and mode, are not used to estimate the population mean. The reason is that the means of samples vary less than other statistics (such as medians and modes) when many samples are selected from the same population. Therefore, the sample mean is the best estimate of the population mean.

Sample measures (i.e., statistics) are used to estimate population measures (i.e., parameters). These statistics are called **estimators.** As previously stated, the sample mean is a better estimator of the population mean than the sample median or sample mode.

A good estimator should satisfy the three properties described now.

Three Properties of a Good Estimator

1. The estimator should be an **unbiased estimator.** That is, the expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.
2. The estimator should be consistent. For a **consistent estimator,** as sample size increases, the value of the estimator approaches the value of the parameter estimated.
3. The estimator should be a **relatively efficient estimator.** That is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance.

Confidence Intervals

As stated in Chapter 6, the sample mean will be, for the most part, somewhat different from the population mean due to sampling error. Therefore, you might ask a second question: How good is a point estimate? The answer is that there is no way of knowing how close a particular point estimate is to the population mean.

This answer places some doubt on the accuracy of point estimates. For this reason, statisticians prefer another type of estimate, called an *interval estimate.*

An **interval estimate** of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.

In an interval estimate, the parameter is specified as being between two values. For example, an interval estimate for the average age of all students might be $21.9 < \mu < 22.7$, or 22.3 ± 0.4 years.

Either the interval contains the parameter or it does not. A degree of confidence (usually a percent) can be assigned before an interval estimate is made. For instance, you may wish to be 95% confident that the interval contains the true population mean. Another question then arises. Why 95%? Why not 99 or 99.5%?

If you desire to be more confident, such as 99 or 99.5% confident, then you must make the interval larger. For example, a 99% confidence interval for the mean age of college students might be $21.7 < \mu < 22.9$, or 22.3 ± 0.6. Hence, a tradeoff occurs. To be more confident that the interval contains the true population mean, you must make the interval wider.

Historical Notes

Point and interval estimates were known as long ago as the late 1700s. However, it wasn't until 1937 that a mathematician, J. Neyman, formulated practical applications for them.

The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.

A **confidence interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.

Intervals constructed in this way are called *confidence intervals.* Three common confidence intervals are used: the 90, the 95, and the 99% confidence intervals.

The algebraic derivation of the formula for determining a confidence interval for a mean will be shown later. A brief intuitive explanation will be given first.

The central limit theorem states that when the sample size is large, approximately 95% of the sample means taken from a population and same sample size will fall within ± 1.96 standard errors of the population mean, that is,

$$\mu \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Now, if a specific sample mean is selected, say, $\overline{X}$, there is a 95% probability that the interval $\mu \pm 1.96(\sigma/\sqrt{n})$ contains $\overline{X}$. Likewise, there is a 95% probability that the interval specified by

$$\overline{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

will contain μ, as will be shown later. Stated another way,

$$\overline{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Hence, you can be 95% confident that the population mean is contained within that interval when the values of the variable are normally distributed in the population.

Interesting Fact

A postal worker who delivers mail walks on average 5.2 miles per day.

The value used for the 95% confidence interval, 1.96, is obtained from Table E in Appendix C. For a 99% confidence interval, the value 2.58 is used instead of 1.96 in the formula. This value is also obtained from Table E and is based on the standard normal distribution. Since other confidence intervals are used in statistics, the symbol $z_{\alpha/2}$ (read "zee sub alpha over two") is used in the general formula for confidence intervals. The Greek letter α (alpha) represents the total area in both tails of the standard normal distribution curve, and $\alpha/2$ represents the area in each one of the tails. More will be said after Examples 7–1 and 7–2 about finding other values for $z_{\alpha/2}$.

The relationship between α and the confidence level is that the stated confidence level is the percentage equivalent to the decimal value of $1 - \alpha$, and vice versa. When the 95% confidence interval is to be found, $\alpha = 0.05$, since $1 - 0.05 = 0.95$, or 95%. When $\alpha = 0.01$, then $1 - \alpha = 1 - 0.01 = 0.99$, and the 99% confidence interval is being calculated.

Formula for the Confidence Interval of the Mean for a Specific α When σ is Known

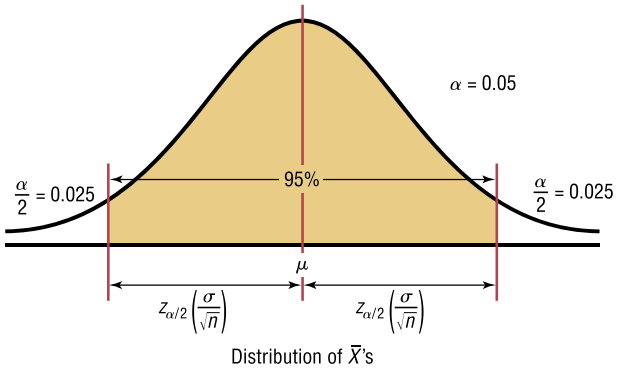
$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval, $z_{\alpha/2} = 1.65$; for a 95% confidence interval, $z_{\alpha/2} = 1.96$; and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.

The term $z_{\alpha/2}(\sigma/\sqrt{n})$ is called the *margin of error* (also called the *maximum error of the estimate*). For a specific value, say, $\alpha = 0.05$, 95% of the sample means will fall within this error value on either side of the population mean, as previously explained. See Figure 7–1.

Figure 7–1

95% Confidence Interval



When $n \geq 30$, s can be substituted for σ, but a different distribution is used.

The **margin of error** also called the maximum error of the estimate is the maximum likely difference between the point estimate of a parameter and the actual value of the parameter.

A more detailed explanation of the margin of error follows Examples 7–1 and 7–2, which illustrate the computation of confidence intervals.

Assumptions for Finding a Confidence Interval for a Mean When σ Is Known

1. The sample is a random sample.
2. Either $n \geq 30$ or the population is normally distributed if $n < 30$.

Rounding Rule for a Confidence Interval for a Mean When you are computing a confidence interval for a population mean by using *raw data,* round off to one more decimal place than the number of decimal places in the original data. When you are computing a confidence interval for a population mean by using a sample mean and a standard deviation, round off to the same number of decimal places as given for the mean.

Example 7–1

Days It Takes to Sell an Aveo

A researcher wishes to estimate the number of days it takes an automobile dealer to sell a Chevrolet Aveo. A sample of 50 cars had a mean time on the dealer's lot of 54 days. Assume the population standard deviation to be 6.0 days. Find the best point estimate of the population mean and the 95% confidence interval of the population mean.

Source: Based on information obtained from Power Information Network.

Solution

The best point estimate of the mean is 54 days. For the 95% confidence interval use $z = 1.96$.

$$\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$54 - 1.96\left(\frac{6.0}{\sqrt{50}}\right) < \mu < 54 + 1.96\left(\frac{6.0}{\sqrt{50}}\right)$$

$$54 - 1.7 < \mu < 54 + 1.7$$

$$52.3 < \mu < 55.7 \text{ or } 54 \pm 1.7$$

Hence one can say with 95% confidence that the interval between 52.3 and 55.7 days does contain the population mean, based on a sample of 50 automobiles.

Example 7–2

Waiting Times in Emergency Rooms

A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes. Assuming that the population standard deviation is 46.5 minutes, find the best point estimate of the population mean and the 99% confidence of the population mean.

Source: Based on information from Press Ganey Associates Inc.

Solution

The best point estimate is 174.3 minutes. The 99% confidence is interval is

$$\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$174.3 - 2.58\left(\frac{46.5}{\sqrt{30}}\right) < \mu < \overline{X} + 2.58\left(\frac{46.5}{\sqrt{30}}\right)$$

$$174.3 - 21.9 < \mu < 174.3 + 21.9$$

$$152.4 < \mu < 196.2$$

Hence, one can be 99% confident that the mean waiting time for emergency room treatment is between 152.4 and 196.2 minutes.

Another way of looking at a confidence interval is shown in Figure 7–2. According to the central limit theorem, approximately 95% of the sample means fall within 1.96 standard deviations of the population mean if the sample size is 30 or more, or if σ is known when n

Figure 7–2
95% Confidence Interval for Sample Means

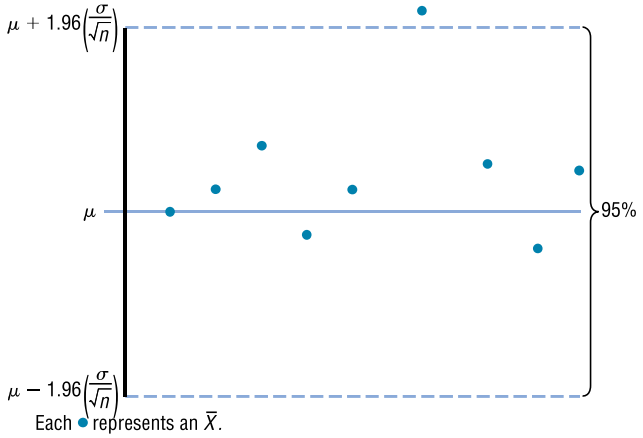
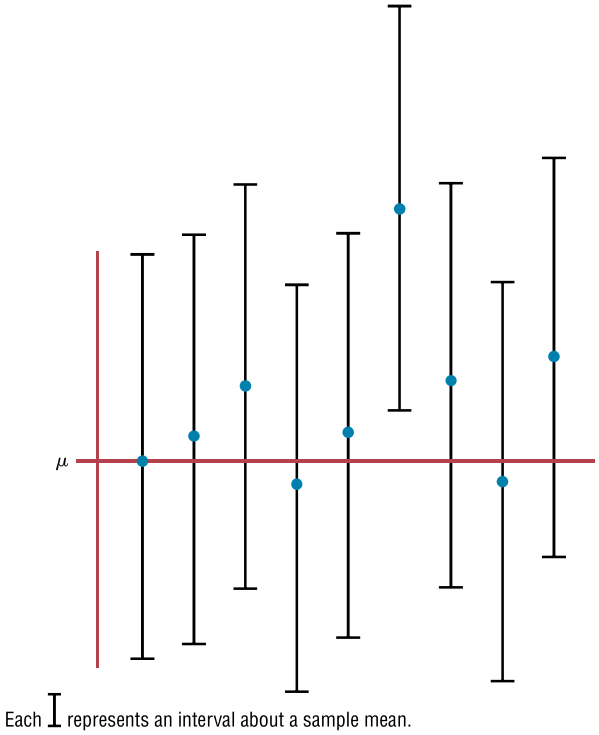


Figure 7–3
95% Confidence Intervals for Each Sample Mean



is less than 30 and the population is normally distributed. If it were possible to build a confidence interval about each sample mean, as was done in Examples 7–1 and 7–2 for μ, 95% of these intervals would contain the population mean, as shown in Figure 7–3. Hence, you can be 95% confident that an interval built around a specific sample mean would contain the population mean. If you desire to be 99% confident, you must enlarge the confidence intervals so that 99 out of every 100 intervals contain the population mean.

Since other confidence intervals (besides 90, 95, and 99%) are sometimes used in statistics, an explanation of how to find the values for $z_{\alpha/2}$ is necessary. As stated previously, the Greek letter α represents the total of the areas in both tails of the normal distribution. The value for α is found by subtracting the decimal equivalent for the desired confidence level from 1. For example, if you wanted to find the 98% confidence interval, you would change 98% to 0.98 and find $\alpha = 1 - 0.98$, or 0.02. Then $\alpha/2$ is obtained by dividing α by 2. So $\alpha/2$ is 0.02/2, or 0.01. Finally, $z_{0.01}$ is the z value that will give an area of 0.01 in the right tail of the standard normal distribution curve. See Figure 7–4.

Figure 7–4
Finding $\alpha/2$ for a 98% Confidence Interval

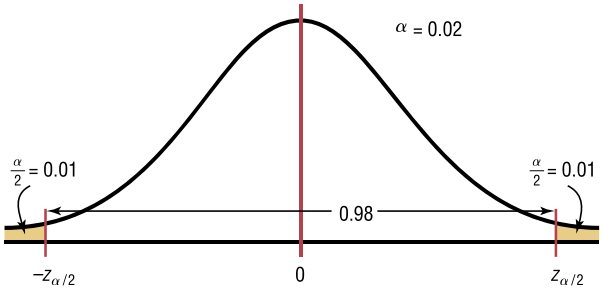
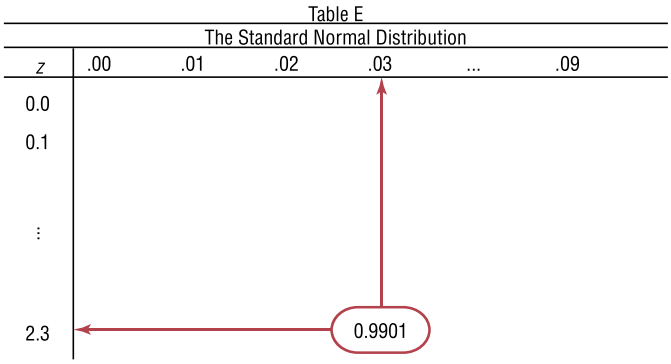


Figure 7–5
Finding $z_{\alpha/2}$ for a 98% Confidence Interval



Once $\alpha/2$ is determined, the corresponding $z_{\alpha/2}$ value can be found by using the procedure shown in Chapter 6, which is reviewed here. To get the $z_{\alpha/2}$ value for a 98% confidence interval, subtract 0.01 from 1.0000 to get 0.9900. Next, locate the area that is closest to 0.9900 (in this case, 0.9901) in Table E, and then find the corresponding z value. In this example, it is 2.33. See Figure 7–5.

For confidence intervals, only the positive z value is used in the formula.

When the original variable is normally distributed and σ is known, the standard normal distribution can be used to find confidence intervals regardless of the size of the sample. When $n \geq 30$, the distribution of means will be approximately normal even if the original distribution of the variable departs from normality.

When σ is unknown, s can be used as an estimate of σ, but a different distribution is used for the critical values. This method is explained in Section 7–2.

Example 7–3 Credit Union Assets

The following data represent a sample of the assets (in millions of dollars) of 30 credit unions in southwestern Pennsylvania. Find the 90% confidence interval of the mean.

12.23	16.56	4.39
2.89	1.24	2.17
13.19	9.16	1.42
73.25	1.91	14.64
11.59	6.69	1.06
8.74	3.17	18.13
7.92	4.78	16.85
40.22	2.42	21.58
5.01	1.47	12.24
2.27	12.77	2.76

Source: *Pittsburgh Post Gazette*.

Solution

Step 1 Find the mean and standard deviation for the data. Use the formulas shown in Chapter 3 or your calculator. The mean $\bar{X} = 11.091$. Assume the standard deviation of the population is 14.405.

Step 2 Find $\alpha/2$. Since the 90% confidence interval is to be used, $\alpha = 1 - 0.90 = 0.10$, and

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

Step 3 Find $z_{\alpha/2}$. Subtract 0.05 from 1.000 to get 0.9500. The corresponding z value obtained from Table E is 1.65. (*Note:* This value is found by using the z value for an area between 0.9495 and 0.9505. A more precise z value obtained mathematically is 1.645 and is sometimes used; however, 1.65 will be used in this textbook.)

Step 4 Substitute in the formula

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$11.091 - 1.65\left(\frac{14.405}{\sqrt{30}}\right) < \mu < 11.091 + 1.65\left(\frac{14.405}{\sqrt{30}}\right)$$

$$11.091 - 4.339 < \mu < 11.091 + 4.339$$

$$6.752 < \mu < 15.430$$

Hence, one can be 90% confident that the population mean of the assets of all credit unions is between \$6.752 million and \$15.430 million, based on a sample of 30 credit unions.

Comment to Computer and Statistical Calculator Users

This chapter and subsequent chapters include examples using raw data. If you are using computer or calculator programs to find the solutions, the answers you get may vary somewhat from the ones given in the textbook. This is so because computers and calculators do not round the answers in the intermediate steps and can use 12 or more decimal places for computation. Also, they use more-exact critical values than those given in the tables in the back of this book. These small discrepancies are part and parcel of statistics.

Objective 2

Determine the minimum sample size for finding a confidence interval for the mean.

Sample Size

Sample size determination is closely related to statistical estimation. Quite often you ask, How large a sample is necessary to make an accurate estimate? The answer is not simple, since it depends on three things: the margin of error, the population standard deviation, and the degree of confidence. For example, how close to the true mean do you want to be (2 units, 5 units, etc.), and how confident do you wish to be (90, 95, 99%, etc.)? For the purpose of this chapter, it will be assumed that the population standard deviation of the variable is known or has been estimated from a previous study.

The formula for sample size is derived from the margin of error formula

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

and this formula is solved for n as follows:

$$E\sqrt{n} = z_{\alpha/2}(\sigma)$$

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

Hence, $$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Formula for the Minimum Sample Size Needed for an Interval Estimate of the Population Mean

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

where E is the margin of error. If necessary, round the answer up to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size n.

Example 7–4

Depth of a River

A scientist wishes to estimate the average depth of a river. He wants to be 99% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.33 feet.

Solution

Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\alpha/2} = 2.58$ and $E = 2$. Substituting in the formula,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left[\frac{(2.58)(4.33)}{2}\right]^2 = 31.2$$

Round the value 31.2 up to 32. Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs at least a sample of 32 measurements.

In most cases in statistics, we round off. However, when determining sample size, we always round up to the next whole number.

Interesting Fact

It has been estimated that the amount of pizza consumed every day in the United States would cover a farm consisting of 75 acres.

Notice that when you are finding the sample size, the size of the population is irrelevant when the population is large or infinite or when sampling is done with replacement. In other cases, an adjustment is made in the formula for computing sample size. This adjustment is beyond the scope of this book.

The formula for determining sample size requires the use of the population standard deviation. What happens when σ is unknown? In this case, an attempt is made to estimate σ. One such way is to use the standard deviation s obtained from a sample taken previously as an estimate for σ. The standard deviation can also be estimated by dividing the range by 4.

Sometimes, interval estimates rather than point estimates are reported. For instance, you may read a statement: "On the basis of a sample of 200 families, the survey estimates that an American family of two spends an average of $84 per week for groceries. One

can be 95% confident that this estimate is accurate within \$3 of the true mean." This statement means that the 95% confidence interval of the true mean is

$$\$84 - \$3 < \mu < \$84 + \$3$$
$$\$81 < \mu < \$87$$

The algebraic derivation of the formula for a confidence interval is shown next. As explained in Chapter 6, the sampling distribution of the mean is approximately normal when large samples ($n \geq 30$) are taken from a population. Also,

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

Furthermore, there is a probability of $1 - \alpha$ that a z will have a value between $-z_{\alpha/2}$ and $+z_{\alpha/2}$. Hence,

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

By using algebra, the formula can be rewritten as

$$-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Subtracting $\bar{X}$ from both sides and from the middle gives

$$-\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Multiplying by -1 gives

$$\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Reversing the inequality yields the formula for the confidence interval:

$$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Applying the Concepts 7–1

Making Decisions with Confidence Intervals

Assume you work for Kimberly Clark Corporation, the makers of Kleenex. The job you are presently working on requires you to decide how many Kleenexes are to be put in the new automobile glove compartment boxes. Complete the following.

1. How will you decide on a reasonable number of Kleenexes to put in the boxes?
2. When do people usually need Kleenexes?
3. What type of data collection technique would you use?
4. Assume you found out that from your sample of 85 people, on average about 57 Kleenexes are used throughout the duration of a cold, with a population standard deviation of 15. Use a confidence interval to help you decide how many Kleenexes will go in the boxes.
5. Explain how you decided how many Kleenexes will go in the boxes.

See page 398 for the answers.