Objective 6
Use the central limit theorem to solve problems involving sample means for large samples.

6–5 The Central Limit Theorem

In addition to knowing how individual data values vary about the mean for a population, statisticians are interested in knowing how the means of samples of the same size taken from the same population vary about the population mean.

Distribution of Sample Means
Suppose a researcher selects a sample of 30 adult males and finds the mean of the measure of the triglyceride levels for the sample subjects to be 187 milligrams/deciliter. Then suppose a second sample is selected, and the mean of that sample is found to be 192 milligrams/deciliter. Continue the process for 100 samples. What happens then is that the mean becomes a random variable, and the sample means 187, 192, 184, . . . , 196 constitute a sampling distribution of sample means.

A sampling distribution of sample means is a distribution using the means computed from all possible random samples of a specific size taken from a population.

If the samples are randomly selected with replacement, the sample means, for the most part, will be somewhat different from the population mean $\mu$. These differences are caused by sampling error.

Sampling error is the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population.

When all possible samples of a specific size are selected with replacement from a population, the distribution of the sample means for a variable has two important properties, which are explained next.

Properties of the Distribution of Sample Means

1. The mean of the sample means will be the same as the population mean.
2. The standard deviation of the sample means will be smaller than the standard deviation of the population, and it will be equal to the population standard deviation divided by the square root of the sample size.

The following example illustrates these two properties. Suppose a professor gave an 8-point quiz to a small class of four students. The results of the quiz were 2, 6, 4, and 8. For the sake of discussion, assume that the four students constitute the population. The mean of the population is

$$\mu = \frac{2 + 6 + 4 + 8}{4} = 5$$

The standard deviation of the population is

$$\sigma = \sqrt{\frac{(2 - 5)^2 + (6 - 5)^2 + (4 - 5)^2 + (8 - 5)^2}{4}} = 2.236$$

The graph of the original distribution is shown in Figure 6–40. This is called a uniform distribution.
Section 6–5  The Central Limit Theorem

Now, if all samples of size 2 are taken with replacement and the mean of each sample is found, the distribution is as shown.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>Sample</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2, 2</td>
<td>2</td>
<td>6, 2</td>
<td>4</td>
</tr>
<tr>
<td>2, 4</td>
<td>3</td>
<td>6, 4</td>
<td>5</td>
</tr>
<tr>
<td>2, 6</td>
<td>4</td>
<td>6, 6</td>
<td>6</td>
</tr>
<tr>
<td>2, 8</td>
<td>5</td>
<td>6, 8</td>
<td>7</td>
</tr>
<tr>
<td>4, 2</td>
<td>3</td>
<td>8, 2</td>
<td>5</td>
</tr>
<tr>
<td>4, 4</td>
<td>4</td>
<td>8, 4</td>
<td>6</td>
</tr>
<tr>
<td>4, 6</td>
<td>5</td>
<td>8, 6</td>
<td>7</td>
</tr>
<tr>
<td>4, 8</td>
<td>6</td>
<td>8, 8</td>
<td>8</td>
</tr>
</tbody>
</table>

A frequency distribution of sample means is as follows.

<table>
<thead>
<tr>
<th>$\bar{x}$</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

For the data from the example just discussed, Figure 6–41 shows the graph of the sample means. The histogram appears to be approximately normal.

The mean of the sample means, denoted by $\mu_{\bar{x}}$, is

$$\mu_{\bar{x}} = \frac{2 + 3 + \ldots + 8}{16} = \frac{80}{16} = 5$$

**Historical Notes**

Two mathematicians who contributed to the development of the central limit theorem were Abraham DeMoivre (1667–1754) and Pierre Simon Laplace (1749–1827). DeMoivre was once jailed for his religious beliefs. After his release, DeMoivre made a living by consulting on the mathematics of gambling and insurance. He wrote two books, *Annuities Upon Lives* and *The Doctrine of Chance*.

Laplace held a government position under Napoleon and later under Louis XVIII. He once computed the probability of the sun rising to be $18,226,214/18,226,215$.  

**Figure 6–40**  
Distribution of Quiz Scores

**Figure 6–41**  
Distribution of Sample Means
which is the same as the population mean. Hence,

\[ \mu_{\bar{x}} = \mu \]

The standard deviation of sample means, denoted by \( \sigma_{\bar{x}} \), is

\[ \sigma_{\bar{x}} = \sqrt{\frac{(2 - 5)^2 + (3 - 5)^2 + \cdots + (8 - 5)^2}{16}} = 1.581 \]

which is the same as the population standard deviation, divided by \( \sqrt{2} \):

\[ \sigma_{\bar{x}} = \frac{2.236}{\sqrt{2}} = 1.581 \]

(Note: Rounding rules were not used here in order to show that the answers coincide.)

In summary, if all possible samples of size \( n \) are taken with replacement from the same population, the mean of the sample means, denoted by \( \mu_{\bar{x}} \), equals the population mean \( \mu \); and the standard deviation of the sample means, denoted by \( \sigma_{\bar{x}} \), equals \( \sigma/\sqrt{n} \). The standard deviation of the sample means is called the standard error of the mean. Hence,

\[ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \]

A third property of the sampling distribution of sample means pertains to the shape of the distribution and is explained by the central limit theorem.

### The Central Limit Theorem

As the sample size \( n \) increases without limit, the shape of the distribution of the sample means taken with replacement from a population with mean \( \mu \) and standard deviation \( \sigma \) will approach a normal distribution. As previously shown, this distribution will have a mean \( \mu \) and a standard deviation \( \sigma/\sqrt{n} \).

If the sample size is sufficiently large, the central limit theorem can be used to answer questions about sample means in the same manner that a normal distribution can be used to answer questions about individual values. The only difference is that a new formula must be used for the \( z \) values. It is

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

Notice that \( \bar{X} \) is the sample mean, and the denominator must be adjusted since means are being used instead of individual data values. The denominator is the standard deviation of the sample means.

If a large number of samples of a given size are selected from a normally distributed population, or if a large number of samples of a given size that is greater than or equal to 30 are selected from a population that is not normally distributed, and the sample means are computed, then the distribution of sample means will look like the one shown in Figure 6–42. Their percentages indicate the areas of the regions.

It's important to remember two things when you use the central limit theorem:

1. When the original variable is normally distributed, the distribution of the sample means will be normally distributed, for any sample size \( n \).
2. When the distribution of the original variable might not be normal, a sample size of 30 or more is needed to use a normal distribution to approximate the distribution of the sample means. The larger the sample, the better the approximation will be.
Examples 6–21 through 6–23 show how the standard normal distribution can be used to answer questions about sample means.

**Example 6–21**

A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours.


**Solution**

Since the variable is approximately normally distributed, the distribution of sample means will be approximately normal, with a mean of 25. The standard deviation of the sample means is

\[
\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{20}} = 0.671
\]

The distribution of the means is shown in Figure 6–43, with the appropriate area shaded.

The z value is

\[
z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = \frac{1.3}{0.671} = 1.94
\]

The area between 0 and 1.94 is 0.4738. Since the desired area is in the tail, subtract 0.4738 from 0.5000. Hence, 0.5000 − 0.4738 = 0.0262, or 2.62%.

One can conclude that the probability of obtaining a sample mean larger than 26.3 hours is 2.62% [i.e., \(P(\bar{X} > 26.3) = 2.62\%\)].
Example 6–22

The average age of a vehicle registered in the United States is 8 years, or 96 months. Assume the standard deviation is 16 months. If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.

*Source: Harper’s Index.*

**Solution**

Since the sample is 30 or larger, the normality assumption is not necessary. The desired area is shown in Figure 6–44.

The two $z$ values are

$$z_1 = \frac{90 - 96}{16/\sqrt{36}} = -2.25$$

$$z_2 = \frac{100 - 96}{16/\sqrt{36}} = 1.50$$

The two areas corresponding to the $z$ values of $-2.25$ and $1.50$, respectively, are $0.4878$ and $0.4332$. Since the $z$ values are on opposite sides of the mean, find the probability by adding the areas: $0.4878 + 0.4332 = 0.921$, or $92.1\%$.

Hence, the probability of obtaining a sample mean between 90 and 100 months is $92.1\%$; that is, $P(90 < \bar{X} < 100) = 92.1\%$.

Students sometimes have difficulty deciding whether to use

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

The formula

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

should be used to gain information about a sample mean, as shown in this section. The formula

$$z = \frac{X - \mu}{\sigma}$$

is used to gain information about an individual data value obtained from the population. Notice that the first formula contains $\bar{X}$, the symbol for the sample mean, while the second formula contains $X$, the symbol for an individual data value. Example 6–23 illustrates the uses of the two formulas.
Example 6–23

The average number of pounds of meat that a person consumes a year is 218.4 pounds. Assume that the standard deviation is 25 pounds and the distribution is approximately normal.


\[ a. \] Find the probability that a person selected at random consumes less than 224 pounds per year.

\[ b. \] If a sample of 40 individuals is selected, find the probability that the mean of the sample will be less than 224 pounds per year.

Solution

\[ a. \] Since the question asks about an individual person, the formula \( z = \frac{X - \mu}{\sigma} \) is used. The distribution is shown in Figure 6–45.

\[ \begin{align*}
& \text{The } z \text{ value is} \\
& z = \frac{X - \mu}{\sigma} = \frac{224 - 218.4}{25} = 0.22
\end{align*} \]

The area between 0 and 0.22 is 0.0871; this area must be added to 0.5000 to get the total area to the left of \( z = 0.22 \).

\[ 0.0871 + 0.5000 = 0.5871 \]

Hence, the probability of selecting an individual who consumes less than 224 pounds per year is 0.5871, or 58.71% [i.e., \( P(X < 224) = 0.5871 \)].

\[ b. \] Since the question concerns the mean of a sample with a size of 40, the formula \( z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \) is used. The area is shown in Figure 6–46.

\[ \begin{align*}
& \text{The } z \text{ value is} \\
& z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{224 - 218.4}{25/\sqrt{40}} = 1.42
\end{align*} \]
The area between $z = 0$ and $z = 1.42$ is 0.4222; this value must be added to 0.5000 to get the total area.

$$0.4222 + 0.5000 = 0.9222$$

Hence, the probability that the mean of a sample of 40 individuals is less than 224 pounds per year is 0.9222, or 92.22%. That is, $P(\bar{X} < 224) = 0.9222$.

Comparing the two probabilities, one can see that the probability of selecting an individual who consumes less than 224 pounds of meat per year is 58.71%, but the probability of selecting a sample of 40 people with a mean consumption of meat that is less than 224 pounds per year is 92.22%. This rather large difference is due to the fact that the distribution of sample means is much less variable than the distribution of individual data values. (Note: An individual person is the equivalent of saying $n = 1$.)

### Finite Population Correction Factor (Optional)

The formula for the standard error of the mean $\sigma/\sqrt{n}$ is accurate when the samples are drawn with replacement or are drawn without replacement from a very large or infinite population. Since sampling with replacement is for the most part unrealistic, a correction factor is necessary for computing the standard error of the mean for samples drawn without replacement from a finite population. Compute the correction factor by using the expression

$$\frac{\sqrt{N - n}}{N - 1}$$

where $N$ is the population size and $n$ is the sample size.

This correction factor is necessary if relatively large samples are taken from a small population, because the sample mean will then more accurately estimate the population mean and there will be less error in the estimation. Therefore, the standard error of the mean must be multiplied by the correction factor to adjust for large samples taken from a small population. That is,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$$

Finally, the formula for the $z$ value becomes

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n} \cdot \sqrt{\frac{N - n}{N - 1}}}$$

When the population is large and the sample is small, the correction factor is generally not used, since it will be very close to 1.00.

The formulas and their uses are summarized in Table 6–1.

### Table 6–1 Summary of Formulas and Their Uses

<table>
<thead>
<tr>
<th>Formula</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $z = \frac{\bar{X} - \mu}{\sigma}$</td>
<td>Used to gain information about an individual data value when the variable is normally distributed.</td>
</tr>
<tr>
<td>2. $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$</td>
<td>Used to gain information when applying the central limit theorem about a sample mean when the variable is normally distributed or when the sample size is 30 or more.</td>
</tr>
</tbody>
</table>
Applying the Concepts 6–5

Central Limit Theorem

Twenty students from a statistics class each collected a random sample of times on how long it took students to get to class from their homes. All the sample sizes were 30. The resulting means are listed.

<table>
<thead>
<tr>
<th>Student</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Student</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>3.7</td>
<td>11</td>
<td>27</td>
<td>1.4</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>4.6</td>
<td>12</td>
<td>24</td>
<td>2.2</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>2.4</td>
<td>13</td>
<td>14</td>
<td>3.1</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
<td>1.9</td>
<td>14</td>
<td>29</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>3.0</td>
<td>15</td>
<td>37</td>
<td>2.8</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>2.8</td>
<td>16</td>
<td>23</td>
<td>2.7</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>1.9</td>
<td>17</td>
<td>26</td>
<td>1.8</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>4.2</td>
<td>18</td>
<td>21</td>
<td>2.0</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>2.6</td>
<td>19</td>
<td>30</td>
<td>2.2</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>2.1</td>
<td>20</td>
<td>29</td>
<td>2.8</td>
</tr>
</tbody>
</table>

1. The students noticed that everyone had different answers. If you randomly sample over and over from any population, with the same sample size, will the results ever be the same?

2. The students wondered whose results were right. How can they find out what the population mean and standard deviation are?

3. Input the means into the computer and check to see if the distribution is normal.

4. Check the mean and standard deviation of the means. How do these values compare to the students’ individual scores?

5. Is the distribution of the means a sampling distribution?

6. Check the sampling error for students 3, 7, and 14.

7. Compare the standard deviation of the sample of the 20 means. Is that equal to the standard deviation from student 3 divided by the square of the sample size? How about for student 7, or 14?

See page 344 for the answers.

Exercises 6–5

1. If samples of a specific size are selected from a population and the means are computed, what is this distribution of means called?

2. Why do most of the sample means differ somewhat from the population mean? What is this difference called?

3. What is the mean of the sample means?

4. What is the standard deviation of the sample means called? What is the formula for this standard deviation?

5. What does the central limit theorem say about the shape of the distribution of sample means?

6. What formula is used to gain information about an individual data value when the variable is normally distributed?

7. What formula is used to gain information about a sample mean when the variable is normally distributed or when the sample size is 30 or more?

For Exercises 8 through 25, assume that the sample is taken from a large population and the correction factor can be ignored.

8. A survey found that the American family generates an average of 17.2 pounds of glass garbage each year. Assume the standard deviation of the distribution is 2.5 pounds. Find the probability that the mean of a sample of 55 families will be between 17 and 18 pounds.


9. The average yearly cost per household of owning a dog is $186.80. Suppose that we randomly select 50 households that own a dog. What is the probability that
10. The average teacher’s salary in New Jersey (ranked first among states) is $52,174. Suppose that the distribution is normal with standard deviation equal to $7500.
   \( a. \) What is the probability that a randomly selected teacher makes less than $50,000 a year?
   \( b. \) If we sample 100 teachers’ salaries, what is the probability that the sample mean is less than $50,000?

Source: N.Y. Times Almanac.

11. The mean weight of 15-year-old males is 142 pounds, and the standard deviation is 12.3 pounds. If a sample of thirty-six 15-year-old males is selected, find the probability that the mean of the sample will be greater than 144.5 pounds. Assume the variable is normally distributed. Based on your answer, would you consider the group overweight?

12. The average teacher’s salary in North Dakota is $29,863. Assume a normal distribution with \( \sigma = $1000. \)
   \( a. \) What is the probability that a randomly selected teacher’s salary is greater than $40,000?
   \( b. \) What is the probability that the mean for a sample of 80 teachers’ salaries is greater than $30,000?

Source: N.Y. Times Almanac.

13. The average price of a pound of sliced bacon is $2.02. Assume the standard deviation is $0.08. If a random sample of 40 one-pound packages is selected, find the probability that the mean of the sample will be less than $2.00.

Source: Statistical Abstract of the United States.

14. The national average SAT score is 1019. Suppose that nothing is known about the shape of the distribution and that the standard deviation is 100. If a random sample of 200 scores were selected and the sample mean were calculated to be 1050, would you be surprised? Explain.

Source: N.Y. Times Almanac.

15. The average number of milligrams (mg) of sodium in a certain brand of low-salt microwave frozen dinners is 660 mg, and the standard deviation is 35 mg. Assume the variable is normally distributed.
   \( a. \) If a single dinner is selected, find the probability that the sodium content will be more than 670 mg.
   \( b. \) If a sample of 10 dinners is selected, find the probability that the mean of the sample will be larger than 670 mg.
   \( c. \) Why is the probability for part \( a \) greater than that for part \( b \)?

16. The average age of chemical engineers is 37 years with a standard deviation of 4 years. If an engineering firm employs 25 chemical engineers, find the probability that the average age of the group is greater than 38.2 years old. If this is the case, would it be safe to assume that the engineers in this group are generally much older than average?

17. The Old Farmer’s Almanac reports that the average person uses 123 gallons of water daily. If the standard deviation is 21 gallons, find the probability that the mean of a randomly selected sample of 15 people will be between 120 and 126 gallons. Assume the variable is normally distributed.

18. The average public elementary school has 458 students. Assume the standard deviation is 97. If a random sample of 36 public elementary schools is selected, find the probability that the number of students enrolled is between 450 and 465.

19. Procter & Gamble reported that an American family of four washes an average of 1 ton (2000 pounds) of clothes each year. If the standard deviation of the distribution is 187.5 pounds, find the probability that the mean of a randomly selected sample of 50 families of four will be between 1980 and 1990 pounds.

Source: The Harper’s Index Book.

20. The average annual salary in Pennsylvania was $24,393 in 1992. Assume that salaries were normally distributed for a certain group of wage earners, and the standard deviation of this group was $4362.
   \( a. \) Find the probability that a randomly selected individual earned less than $26,000.
   \( b. \) Find the probability that, for a randomly selected sample of 25 individuals, the mean salary was less than $26,000.
   \( c. \) Why is the probability for part \( b \) higher than the probability for part \( a \)?

Source: Associated Press.

21. The average time it takes a group of adults to complete a certain achievement test is 46.2 minutes. The standard deviation is 8 minutes. Assume the variable is normally distributed.
   \( a. \) Find the probability that a randomly selected adult will complete the test in less than 43 minutes.
   \( b. \) Find the probability that, if 50 randomly selected adults take the test, the mean time it takes the group to complete the test will be less than 43 minutes.
   \( c. \) Does it seem reasonable that an adult would finish the test in less than 43 minutes? Explain.
   \( d. \) Does it seem reasonable that the mean of the 50 adults could be less than 43 minutes?
22. Assume that the mean systolic blood pressure of normal adults is 120 millimeters of mercury (mm Hg) and the standard deviation is 5.6. Assume the variable is normally distributed.
   a. If an individual is selected, find the probability that the individual’s pressure will be between 120 and 121.8 mm Hg.
   b. If a sample of 30 adults is randomly selected, find the probability that the sample mean will be between 120 and 121.8 mm Hg.
   c. Why is the answer to part a so much smaller than the answer to part b?

23. The average cholesterol content of a certain brand of eggs is 215 milligrams, and the standard deviation is 15 milligrams. Assume the variable is normally distributed.
   a. If a single egg is selected, find the probability that the cholesterol content will be greater than 220 milligrams.
   b. If a sample of 25 eggs is selected, find the probability that the mean of the sample will be larger than 220 milligrams.
   Source: Living Fit.

24. At a large publishing company, the mean age of proofreaders is 36.2 years, and the standard deviation is 3.7 years. Assume the variable is normally distributed.
   a. If a proofreader from the company is randomly selected, find the probability that his or her age will be between 36 and 37.5 years.
   b. If a random sample of 15 proofreaders is selected, find the probability that the mean age of the proofreaders in the sample will be between 36 and 37.5 years.

25. In the United States, one farmworker supplied agricultural products for an average of 106 people. Assume the standard deviation is 16.1. If 35 farmworkers are selected, find the probability that the mean number of people supplied is between 100 and 110.

---

**Extending the Concepts**

For Exercises 26 and 27, check to see whether the correction factor should be used. If so, be sure to include it in the calculations.

26. In a study of the life expectancy of 500 people in a certain geographic region, the mean age at death was 72.0 years, and the standard deviation was 5.3 years. If a sample of 50 people from this region is selected, find the probability that the mean life expectancy will be less than 70 years.

27. A study of 800 homeowners in a certain area showed that the average value of the homes was $82,000, and the standard deviation was $5000. If 50 homes are for sale, find the probability that the mean of the values of these homes is greater than $83,500.

28. The average breaking strength of a certain brand of steel cable is 2000 pounds, with a standard deviation of 100 pounds. A sample of 20 cables is selected and tested. Find the sample mean that will cut off the upper 95% of all samples of size 20 taken from the population. Assume the variable is normally distributed.

29. The standard deviation of a variable is 15. If a sample of 100 individuals is selected, compute the standard error of the mean. What size sample is necessary to double the standard error of the mean?

30. In Exercise 29, what size sample is needed to cut the standard error of the mean in half?

---

**6–6 The Normal Approximation to the Binomial Distribution**

A normal distribution is often used to solve problems that involve the binomial distribution since, when $n$ is large (say, 100), the calculations are too difficult to do by hand using the binomial distribution. Recall from Chapter 5 that a binomial distribution has the following characteristics:

1. There must be a fixed number of trials.
2. The outcome of each trial must be independent.