6–3

Objective 6

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

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 μ. 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

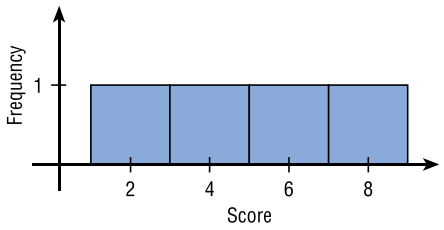
$$\mu = \frac{2 + 6 + 4 + 8}{4} = 5$$

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 6–29. This is called a *uniform distribution.*

Figure 6–29
Distribution of Quiz Scores



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.

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

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

A frequency distribution of sample means is as follows.

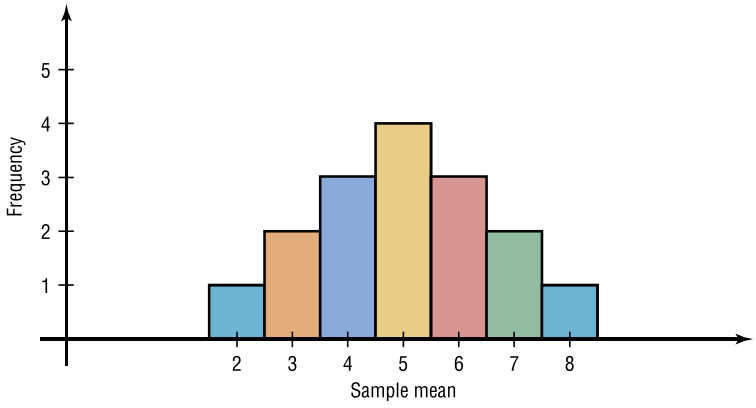
$\overline{X}$	f
2	1
3	2
4	3
5	4
6	3
7	2
8	1

For the data from the example just discussed, Figure 6–30 shows the graph of the sample means. The histogram appears to be approximately normal.

The mean of the sample means, denoted by $\mu_{\overline{X}}$, is

$$\mu_{\overline{X}} = \frac{2 + 3 + \cdots + 8}{16} = \frac{80}{16} = 5$$

Figure 6–30
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 $\sqrt{2}$:

$$\sigma_{\bar{X}} = \frac{2.236}{\sqrt{2}} = 1.581$$

Unusual Stats

Each year a person living in the United States consumes on average 1400 pounds of food.

(*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 μ; 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 μ and standard deviation σ will approach a normal distribution. As previously shown, this distribution will have a mean μ 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

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

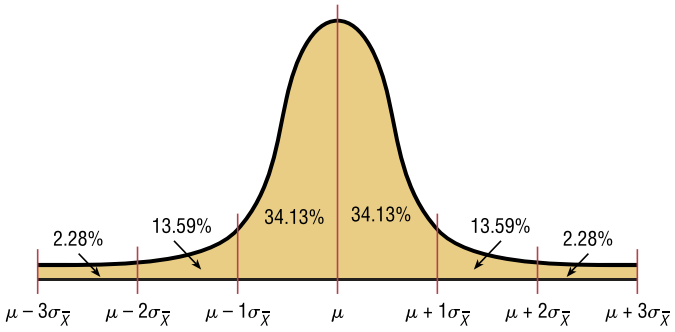
Notice that $\bar{X}$ is the sample mean, and the denominator must be adjusted since means are being used instead of individual data values. The denominator is the standard deviation of the sample means.

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 6–31. 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.

Figure 6–31
Distribution of Sample Means for a Large Number of Samples



Examples 6–13 through 6–15 show how the standard normal distribution can be used to answer questions about sample means.

Example 6–13

Hours That Children Watch Television

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.

Source: Michael D. Shook and Robert L. Shook, *The Book of Odds*.

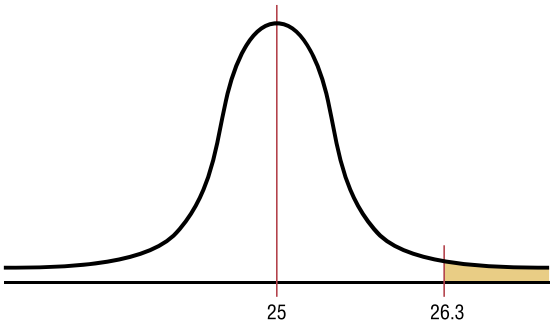
Solution

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 6–32, with the appropriate area shaded.

Figure 6–32
Distribution of the Means for Example 6–13



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 to the right of 1.94 is $1.000 - 0.9738 = 0.0262$, 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 6–14

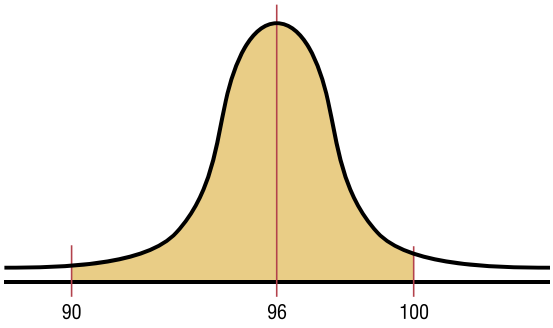
The average age of a vehicle registered in the United States is 8 years, or 96 months. Assume the standard deviation is 16 months. If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.

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 6–33.

Figure 6–33
Area Under a Normal Curve for Example 6–14



The two z values are

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

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

To find the area between the two z values of -2.25 and 1.50, look up the corresponding area in Table E and subtract one from the other. The area for $z = -2.25$ is 0.0122, and the area for $z = 1.50$ is 0.9332. Hence the area between the two values is $0.9332 - 0.0122 = 0.9210$, or 92.1%.

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

Students sometimes have difficulty deciding whether to use

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

The formula

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

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 $\overline{X}$, the symbol for the sample mean, while the second formula contains X, the symbol for an individual data value. Example 6–15 illustrates the uses of the two formulas.

Example 6–15

Meat Consumption

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

Source: Michael D. Shook and Robert L. Shook, *The Book of Odds*.

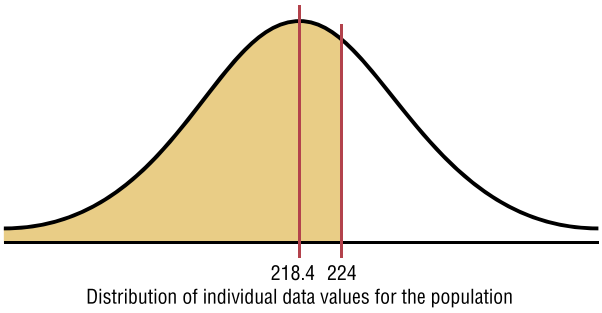
a. Find the probability that a person selected at random consumes less than 224 pounds per year.

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 $z = (X - \mu)/\sigma$ is used. The distribution is shown in Figure 6–34.

Figure 6–34
Area Under a Normal Curve for Part *a* of Example 6–15



Distribution of individual data values for the population

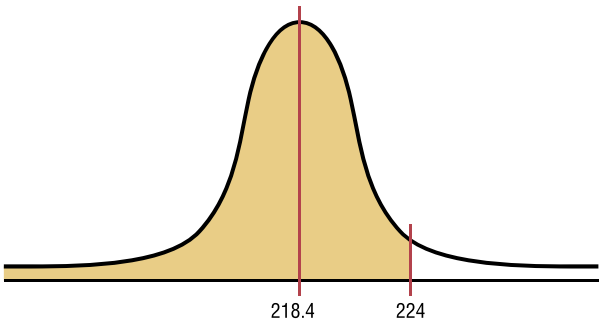
The z value is

$$z = \frac{X - \mu}{\sigma} = \frac{224 - 218.4}{25} = 0.22$$

The area to the left of $z = 0.22$ is 0.5871. Hence, the probability of selecting an individual who consumes less than 224 pounds of meat 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 $z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ is used. The area is shown in Figure 6–35.

Figure 6–35
Area Under a Normal Curve for Part *b* of Example 6–15



Distribution of means for all samples of size 40 taken from the population

The z value is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{224 - 218.4}{25/\sqrt{40}} = 1.42$$

The area to the left of $z = 1.42$ is 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, $P(\overline{X} < 224) = 0.9222$.

Comparing the two probabilities, you 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

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

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

Interesting Fact

The bubonic plague killed more than 25 million people in Europe between 1347 and 1351.

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,

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

Finally, the formula for the z value becomes

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

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

Formula	Use
1. $z = \frac{X - \mu}{\sigma}$	Used to gain information about an individual data value when the variable is normally distributed.
2. $z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$	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.