characteristics of the sampling distribution of $\bar{x}$. In Section 7.6 we discuss the characteristics of the sampling distribution of $\bar{p}$.

7.5 Sampling Distribution of $\bar{x}$

In the previous section we said that the sample mean $\bar{x}$ is a random variable and its probability distribution is called the sampling distribution of $\bar{x}$.

SAMPLING DISTRIBUTION OF $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.

This section describes the properties of the sampling distribution of $\bar{x}$. Just as with other probability distributions we studied, the sampling distribution of $\bar{x}$ has an expected value or mean, a standard deviation, and a characteristic shape or form. Let us begin by considering the mean of all possible $\bar{x}$ values, which is referred to as the expected value of $\bar{x}$.

Expected Value of $\bar{x}$

In the EAI sampling problem we saw that different simple random samples result in a variety of values for the sample mean $\bar{x}$. Because many different values of the random variable $\bar{x}$ are possible, we are often interested in the mean of all possible values of $\bar{x}$ that can be generated by the various simple random samples. The mean of the $\bar{x}$ random variable is the expected value of $\bar{x}$. Let $E(\bar{x})$ represent the expected value of $\bar{x}$ and μ represent the mean of the population from which we are selecting a simple random sample. It can be shown that with simple random sampling, $E(\bar{x})$ and μ are equal.

The expected value of $\bar{x}$ equals the mean of the population from which the sample is selected.

EXPECTED VALUE OF $\bar{x}$

$$E(\bar{x}) = \mu \tag{7.1}$$

where

$E(\bar{x})$ = the expected value of $\bar{x}$
μ = the population mean

This result shows that with simple random sampling, the expected value or mean of the sampling distribution of $\bar{x}$ is equal to the mean of the population. In Section 7.1 we saw that the mean annual salary for the population of EAI employees is $\mu = \$51{,}800$. Thus, according to equation (7.1), the mean of all possible sample means for the EAI study is also $51,800.

When the expected value of a point estimator equals the population parameter, we say the point estimator is **unbiased**. Thus, equation (7.1) shows that $\bar{x}$ is an unbiased estimator of the population mean μ.

Standard Deviation of $\bar{x}$

Let us define the standard deviation of the sampling distribution of $\bar{x}$. We will use the following notation.

$\sigma_{\bar{x}}$ = the standard deviation of $\bar{x}$
σ = the standard deviation of the population
n = the sample size
N = the population size

It can be shown that the formula for the standard deviation of $\bar{x}$ depends on whether the population is finite or infinite. The two formulas for the standard deviation of $\bar{x}$ follow.

STANDARD DEVIATION OF $\bar{x}$

Finite Population	*Infinite Population*	
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	(7.2)

In comparing the two formulas in equation (7.2), we see that the factor $\sqrt{(N-n)/(N-1)}$ is required for the finite population case but not for the infinite population case. This factor is commonly referred to as the **finite population correction factor**. In many practical sampling situations, we find that the population involved, although finite, is "large," whereas the sample size is relatively "small." In such cases the finite population correction factor $\sqrt{(N-n)/(N-1)}$ is close to 1. As a result, the difference between the values of the standard deviation of $\bar{x}$ for the finite and infinite population cases becomes negligible. Then, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ becomes a good approximation to the standard deviation of $\bar{x}$ even though the population is finite. This observation leads to the following general guideline, or rule of thumb, for computing the standard deviation of $\bar{x}$.

USE THE FOLLOWING EXPRESSION TO COMPUTE THE STANDARD DEVIATION OF $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (7.3)$$

whenever

1. The population is infinite; or
2. The population is finite *and* the sample size is less than or equal to 5% of the population size; that is, $n/N \leq .05$.

Exercise 17 shows that when $n/N \leq .05$, the finite population correction factor has little effect on the value of $\sigma_{\bar{x}}$.

In cases where $n/N > .05$, the finite population version of formula (7.2) should be used in the computation of $\sigma_{\bar{x}}$. Unless otherwise noted, throughout the text we will assume that the population size is "large," $n/N \leq .05$, and expression (7.3) can be used to compute $\sigma_{\bar{x}}$.

The term standard error *is used throughout statistical inference to refer to the standard deviation of a point estimator.*

To compute $\sigma_{\bar{x}}$, we need to know σ, the standard deviation of the population. To further emphasize the difference between $\sigma_{\bar{x}}$ and σ, we refer to the standard deviation of $\bar{x}$, $\sigma_{\bar{x}}$, as the **standard error** of the mean. In general, the term *standard error* refers to the standard deviation of a point estimator. Later we will see that the value of the standard error of the mean is helpful in determining how far the sample mean may be from the population mean. Let us now return to the EAI example and compute the standard error of the mean associated with simple random samples of 30 EAI employees.

In Section 7.1 we saw that the standard deviation of annual salary for the population of 2500 EAI employees is $\sigma = 4000$. In this case, the population is finite, with $N = 2500$. However, with a sample size of 30, we have $n/N = 30/2500 = .012$. Because the sample size is less than 5% of the population size, we can ignore the finite population correction factor and use equation (7.3) to compute the standard error.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$$

Form of the Sampling Distribution of $\bar{x}$

The preceding results concerning the expected value and standard deviation for the sampling distribution of $\bar{x}$ are applicable for any population. The final step in identifying the characteristics of the sampling distribution of $\bar{x}$ is to determine the form or shape of the sampling distribution. We will consider two cases: (1) The population has a normal distribution; and (2) the population does not have a normal distribution.

Population has a normal distribution In many situations it is reasonable to assume that the population from which we are selecting a random sample has a normal, or nearly normal, distribution. When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

Population does not have a normal distribution When the population from which we are selecting a random sample does not have a normal distribution, the **central limit theorem** is helpful in identifying the shape of the sampling distribution of $\bar{x}$. A statement of the central limit theorem as it applies to the sampling distribution of $\bar{x}$ follows.

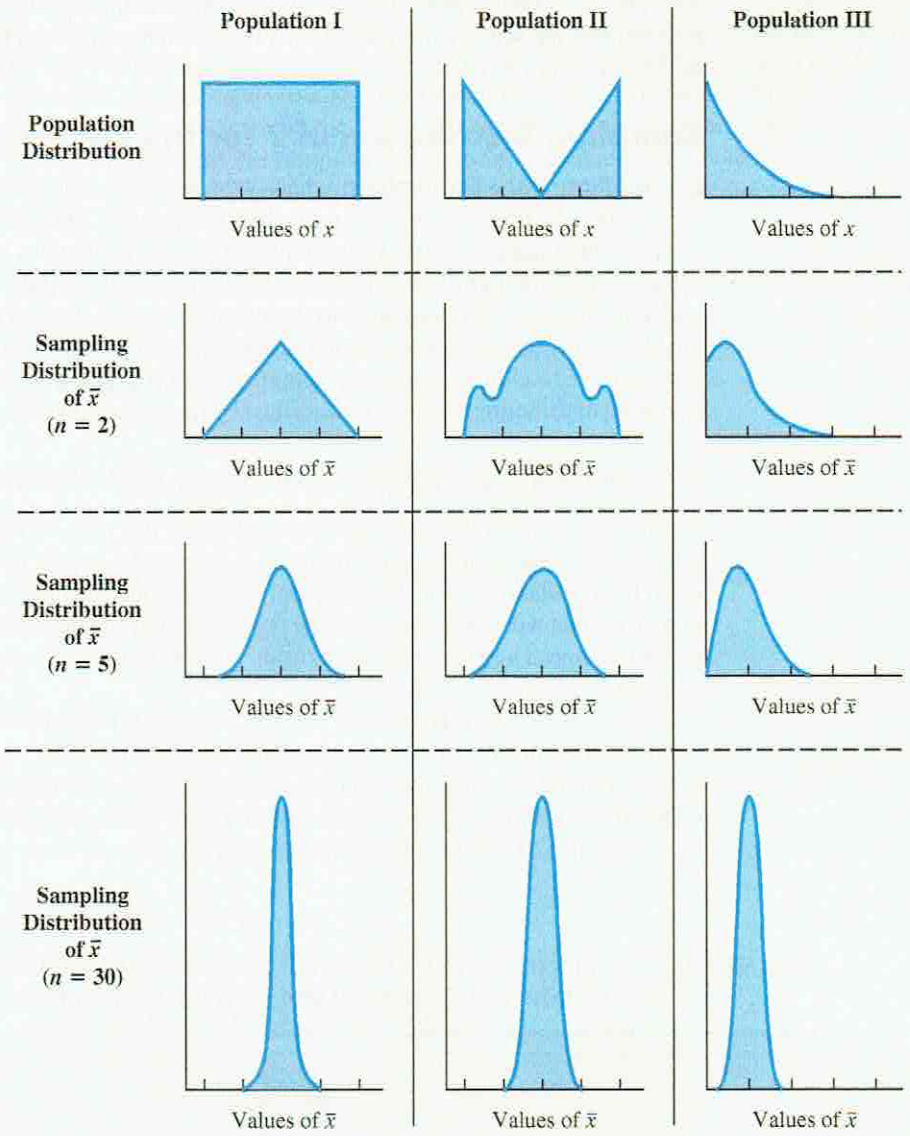
CENTRAL LIMIT THEOREM

In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a *normal distribution* as the sample size becomes large.

Figure 7.6 shows how the central limit theorem works for three different populations; each column refers to one of the populations. The top panel of the figure shows that none of the populations are normally distributed. Population I follows a uniform distribution. Population II is often called the rabbit-eared distribution. It is symmetric, but the more likely values fall in the tails of the distribution. Population III is shaped like the exponential distribution; it is skewed to the right.

The bottom three panels of Figure 7.6 show the shape of the sampling distribution for samples of size $n = 2$, $n = 5$, and $n = 30$. When the sample size is 2, we see that the shape of each sampling distribution is different from the shape of the corresponding population distribution. For samples of size 5, we see that the shapes of the sampling distributions

FIGURE 7.6 ILLUSTRATION OF THE CENTRAL LIMIT THEOREM FOR THREE POPULATIONS



for populations I and II begin to look similar to the shape of a normal distribution. Even though the shape of the sampling distribution for population III begins to look similar to the shape of a normal distribution, some skewness to the right is still present. Finally, for samples of size 30, the shapes of each of the three sampling distributions are approximately normal.

From a practitioner standpoint, we often want to know how large the sample size needs to be before the central limit theorem applies and we can assume that the shape of the sampling distribution is approximately normal. Statistical researchers have investigated this question by studying the sampling distribution of $\bar{x}$ for a variety of populations and a variety of sample sizes. General statistical practice is to assume that, for most

applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more. In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed. Finally, if the population is discrete, the sample size needed for a normal approximation often depends on the population proportion. We say more about this issue when we discuss the sampling distribution of $\bar{p}$ in Section 7.6.

Sampling Distribution of $\bar{x}$ for the EAI Problem

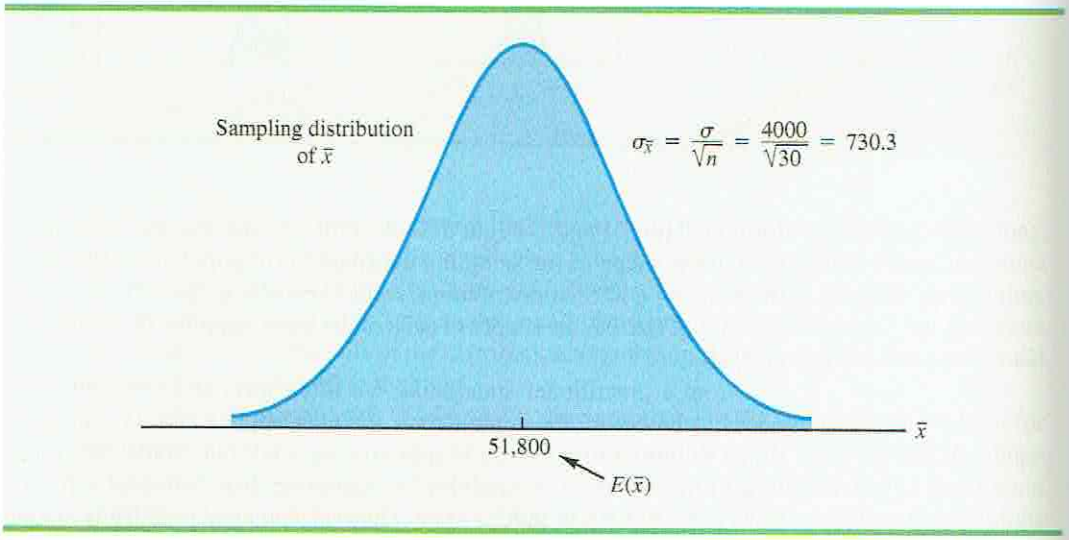
Let us return to the EAI problem where we previously showed that $E(\bar{x}) = \$51{,}800$ and $\sigma_{\bar{x}} = 730.3$. At this point, we do not have any information about the population distribution; it may or may not be normally distributed. If the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed. If the population does not have a normal distribution, the simple random sample of 30 employees and the central limit theorem enable us to conclude that the sampling distribution of $\bar{x}$ can be approximated by a normal distribution. In either case, we are comfortable proceeding with the conclusion that the sampling distribution of $\bar{x}$ can be described by the normal distribution shown in Figure 7.7.

Practical Value of the Sampling Distribution of $\bar{x}$

Whenever a simple random sample is selected and the value of the sample mean is used to estimate the value of the population mean μ, we cannot expect the sample mean to exactly equal the population mean. The practical reason we are interested in the sampling distribution of $\bar{x}$ is that it can be used to provide probability information about the difference between the sample mean and the population mean. To demonstrate this use, let us return to the EAI problem.

Suppose the personnel director believes the sample mean will be an acceptable estimate of the population mean if the sample mean is within \$500 of the population mean. However, it is not possible to guarantee that the sample mean will be within \$500 of the population mean. Indeed, Table 7.5 and Figure 7.4 show that some of the 500 sample means differed by more than \$2000 from the population mean. So we must think of the personnel director's

FIGURE 7.7 SAMPLING DISTRIBUTION OF $\bar{x}$ FOR THE MEAN ANNUAL SALARY OF A SIMPLE RANDOM SAMPLE OF 30 EAI EMPLOYEES



request in probability terms. That is, the personnel director is concerned with the following question: What is the probability that the sample mean computed using a simple random sample of 30 EAI employees will be within \$500 of the population mean?

Because we have identified the properties of the sampling distribution of $\bar{x}$ (see Figure 7.7), we will use this distribution to answer the probability question. Refer to the sampling distribution of $\bar{x}$ shown again in Figure 7.8. With a population mean of \$51,800, the personnel director wants to know the probability that $\bar{x}$ is between \$51,300 and \$52,300. This probability is given by the darkly shaded area of the sampling distribution shown in Figure 7.8. Because the sampling distribution is normally distributed, with mean 51,800 and standard error of the mean 730.3, we can use the standard normal probability table to find the area or probability.

We first calculate the z value at the upper endpoint of the interval (52,300) and use the table to find the cumulative probability at that point (left tail area). Then we compute the z value at the lower endpoint of the interval (51,300) and use the table to find the area under the curve to the left of that point (another left tail area). Subtracting the second tail area from the first gives us the desired probability.

At $\bar{x} = 52{,}300$, we have

$$z = \frac{52{,}300 - 51{,}800}{730.30} = .68$$

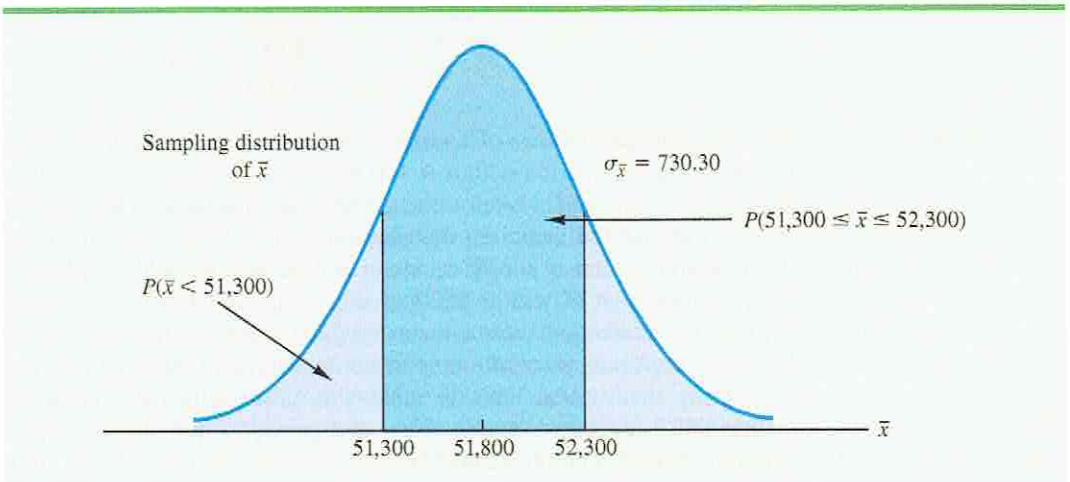
Referring to the standard normal probability table, we find a cumulative probability (area to the left of $z = .68$) of .7517.

At $\bar{x} = 51{,}300$, we have

$$z = \frac{51{,}300 - 51{,}800}{730.30} = -.68$$

The area under the curve to the left of $z = -.68$ is .2483. Therefore, $P(51{,}300 \leq \bar{x} \leq 52{,}300) = P(z \leq .68) - P(z < -.68) = .7517 - .2483 = .5034$.

FIGURE 7.8 PROBABILITY OF A SAMPLE MEAN BEING WITHIN \$500 OF THE POPULATION MEAN FOR A SIMPLE RANDOM SAMPLE OF 30 EAI EMPLOYEES



Using Excel's NORM.DIST function is easier and provides more accurate results than using the tables with rounded values for z.

The desired probability can also be computed using Excel's NORM.DIST function. The advantage of using the NORM.DIST function is that we do not have to make a separate computation of the z value. Evaluating the NORM.DIST function at the upper endpoint of the interval provides the cumulative probability at 52,300. Entering the formula =NORM.DIST(52300,51800,730.30,TRUE) into a cell of an Excel worksheet provides .7532 for this cumulative probability. Evaluating the NORM.DIST function at the lower endpoint of the interval provides the area under the curve to the left of 51,300. Entering the formula =NORM.DIST(51300,51800,730.30,TRUE) into a cell of an Excel worksheet provides .2468 for this cumulative probability. The probability of $\bar{x}$ being in the interval from 51,300 to 52,300 is then given by .7532 − .2468 = .5064. We note that this result is slightly different from the probability obtained using the table, because in using the normal table we rounded to two decimal places of accuracy when computing the z value. The result obtained using NORM.DIST is thus more accurate.

The sampling distribution of $\bar{x}$ can be used to provide probability information about how close the sample mean $\bar{x}$ is to the population mean μ.

The preceding computations show that a simple random sample of 30 EAI employees has a .5064 probability of providing a sample mean $\bar{x}$ that is within \$500 of the population mean. Thus, there is a 1 − .5064 = .4936 probability that the sampling error will be more than \$500. In other words, a simple random sample of 30 EAI employees has roughly a 50–50 chance of providing a sample mean within the allowable \$500. Perhaps a larger sample size should be considered. Let us explore this possibility by considering the relationship between the sample size and the sampling distribution of $\bar{x}$.

Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$

Suppose that in the EAI sampling problem we select a simple random sample of 100 EAI employees instead of the 30 originally considered. Intuitively, it would seem that with more data provided by the larger sample size, the sample mean based on $n = 100$ should provide a better estimate of the population mean than the sample mean based on $n = 30$. To see how much better, let us consider the relationship between the sample size and the sampling distribution of $\bar{x}$.

First note that $E(\bar{x}) = \mu$ regardless of the sample size. Thus, the mean of all possible values of $\bar{x}$ is equal to the population mean μ regardless of the sample size n. However, note that the standard error of the mean, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, is related to the square root of the sample size. Whenever the sample size is increased, the standard error of the mean $\sigma_{\bar{x}}$ decreases. With $n = 30$, the standard error of the mean for the EAI problem is 730.3. However, with the increase in the sample size to $n = 100$, the standard error of the mean is decreased to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400$$

The sampling distributions of $\bar{x}$ with $n = 30$ and $n = 100$ are shown in Figure 7.9. Because the sampling distribution with $n = 100$ has a smaller standard error, the values of $\bar{x}$ have less variation and tend to be closer to the population mean than the values of $\bar{x}$ with $n = 30$.

We can use the sampling distribution of $\bar{x}$ for the case with $n = 100$ to compute the probability that a simple random sample of 100 EAI employees will provide a sample mean that is within \$500 of the population mean. In this case the sampling distribution is normal with a mean of 51,800 and a standard deviation of 400 (see Figure 7.10). Again, we could compute the appropriate z values and use the standard normal probability distribution table to make this probability calculation. However, Excel's NORM.DIST function is easier to use and provides more accurate results. Entering the formula =NORM.DIST(52300,51800,400,TRUE) into a cell of an Excel worksheet provides the cumulative probability corresponding to $\bar{x} = 52{,}300$. The value provided

FIGURE 7.9 A COMPARISON OF THE SAMPLING DISTRIBUTIONS OF $\bar{x}$ FOR SIMPLE RANDOM SAMPLES OF $n = 30$ AND $n = 100$ EAI EMPLOYEES

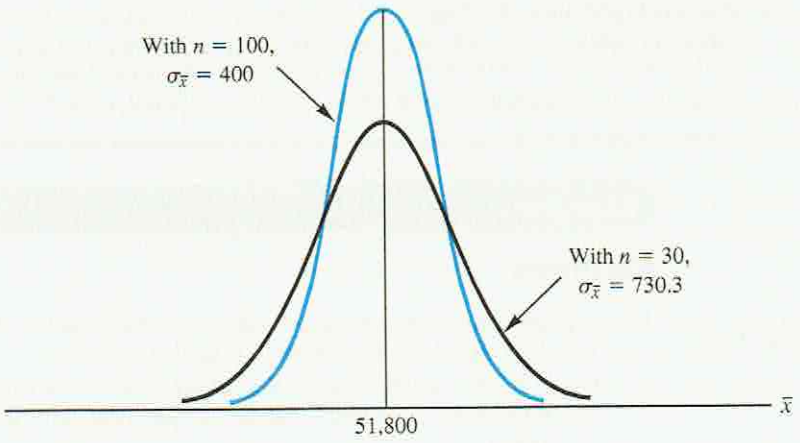
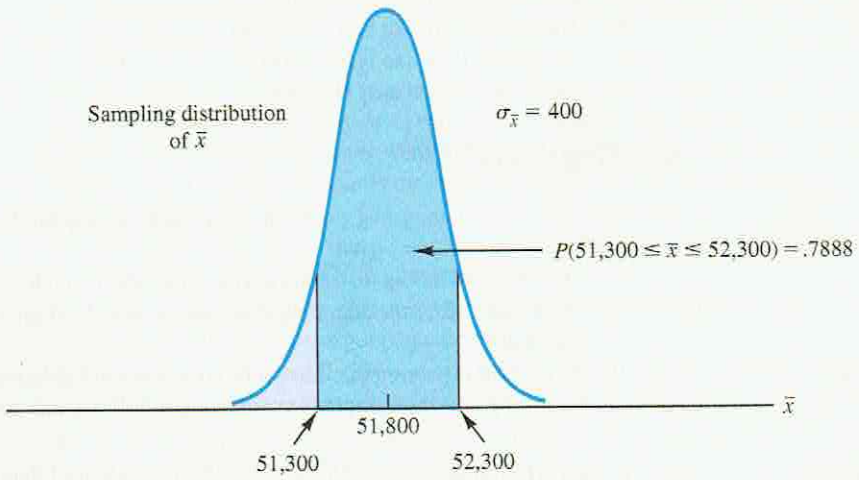


FIGURE 7.10 PROBABILITY OF A SAMPLE MEAN BEING WITHIN $500 OF THE POPULATION MEAN FOR A SIMPLE RANDOM SAMPLE OF 100 EAI EMPLOYEES



by Excel is .8944. Entering the formula = NORM.DIST (51300,51800,400,TRUE) into a cell of an Excel worksheet provides the cumulative probability corresponding to $\bar{x} = 51{,}300$. The value provided by Excel is .1056. Thus, the probability of $\bar{x}$ being in the interval from 51,300 to 52,300 is given by .8944 − .1056 = .7888. By increasing the sample size from 30 to 100 EAI employees, we increase the probability that the sampling error will be $500 or less; that is, the probability of obtaining a sample mean within $500 of the population mean increases from .5064 to .7888.

The important point in this discussion is that as the sample size increases, the standard error of the mean decreases. As a result, a larger sample size will provide a higher probability that the sample mean falls within a specified distance of the population mean.

NOTE AND COMMENT

In presenting the sampling distribution of $\bar{x}$ for the EAI problem, we took advantage of the fact that the population mean $\mu = 51{,}800$ and the population standard deviation $\sigma = 4000$ were known. However, usually the values of the population mean μ and the population standard deviation σ that are needed to determine the sampling distribution of $\bar{x}$ will be unknown. In Chapter 8 we show how the sample mean $\bar{x}$ and the sample standard deviation s are used when μ and σ are unknown.

Exercises

Methods

SELFtest

15. A population has a mean of 200 and a standard deviation of 50. Suppose a simple random sample of size 100 is selected and $\bar{x}$ is used to estimate μ.
    a. What is the probability that the sample mean will be within ± 5 of the population mean?
    b. What is the probability that the sample mean will be within ± 10 of the population mean?

16. Assume the population standard deviation is $\sigma = 25$. Compute the standard error of the mean, $\sigma_{\bar{x}}$, for sample sizes of 50, 100, 150, and 200. What can you say about the size of the standard error of the mean as the sample size is increased?

17. Suppose a random sample of size 50 is selected from a population with $\sigma = 10$. Find the value of the standard error of the mean in each of the following cases (use the finite population correction factor if appropriate).
    a. The population size is infinite.
    b. The population size is $N = 50{,}000$.
    c. The population size is $N = 5000$.
    d. The population size is $N = 500$.

Applications

18. Refer to the EAI sampling problem. Suppose a simple random sample of 60 employees is used.
    a. Sketch the sampling distribution of $\bar{x}$ when simple random samples of size 60 are used.
    b. What happens to the sampling distribution of $\bar{x}$ if simple random samples of size 120 are used?
    c. What general statement can you make about what happens to the sampling distribution of $\bar{x}$ as the sample size is increased? Does this generalization seem logical? Explain.

SELFtest

19. In the EAI sampling problem (see Figure 7.8), we showed that for $n = 30$, there was .5064 probability of obtaining a sample mean within $\pm \$500$ of the population mean.
    a. What is the probability that $\bar{x}$ is within \$500 of the population mean if a sample of size 60 is used?
    b. Answer part (a) for a sample of size 120.

20. *Barron's* reported that the average number of weeks an individual is unemployed is 17.5 weeks. Assume that for the population of all unemployed individuals the population mean length of unemployment is 17.5 weeks and that the population standard deviation is 4 weeks. Suppose you would like to select a random sample of 50 unemployed individuals for a follow-up study.
    a. Show the sampling distribution of $\bar{x}$, the sample mean average for a sample of 50 unemployed individuals.
    b. What is the probability that a simple random sample of 50 unemployed individuals will provide a sample mean within 1 week of the population mean?