CHAPTER 7

Distributions of Sampling Statistics

He uses statistics as a drunken man uses lampposts—for support rather than illumination.

Andrew Lang (Scottish author)

I could prove God statistically.

George Gallup, U.S. pollster

CONTENTS

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00007-7

We introduce the concept of sampling from a population distribution. The sample mean and sample variance are studied, and their expectations and variances are given. The central limit theorem is presented and applied to show that the distribution of the sample mean is approximately normal.

We consider samples taken from a finite population in which certain members have a particular characteristic of interest. We show that when the population size is large, the number of members of the sample who have the characteristic is approximately a binomial random variable. The central limit theorem is used to show that the probabilities of such a random variable can be approximated by the probabilities of a normal random variable.

We present the distribution of the sample variance in the case where the underlying population distribution is normal.

7.1 A PREVIEW

If you bet \$1 on a number at a roulette table in a U.S. casino, then either you will win \$35 if your number appears on the roulette wheel or you will lose \$1 if it does not. Since the wheel has 38 slots–numbered 0, 00, and each of the integers from 1 to 36—it follows that the probability that your number appears is 1/38. As a result, your expected gain on the bet is

$$E[\text{gain}] = 35\left(\frac{1}{38}\right) - 1\left(\frac{37}{38}\right) = -\frac{2}{38} = -0.0526$$

That is, your expected loss on each spin of the wheel is approximately 5.3 cents.

Suppose you continually place bets at the roulette table. How lucky do you have to be in order to be winning money at the end of your play? Well, it depends on how long you continue to play. Indeed, after 100 plays you will be ahead with probability 0.4916. On the other hand, after 1000 plays your chance of being ahead drops to 0.39. After 100,000 plays not only will you almost certainly be losing (your probability of being ahead is approximately 0.002), but also you can be 95 percent certain that your average loss per play will be 5.26 ± 1.13 cents (read as 5.26 plus or minus 1.13 cents). In other words, even if you did not know it to begin with, if you play long enough, you will learn that the average loss per game is around 5.26 cents.

7.2 INTRODUCTION

One of the key concerns of statistics is the drawing of conclusions from a set of observed data. These data will usually consist of a sample of certain elements of a population, and the objective will be to use the sample to draw conclusions about the entire population.

Suppose that each member of a population has a numerical value associated with it. To use sample data to make inferences about the values of the entire population, it is necessary to make some assumptions about the population values and about the relationship between the sample and the population. One such assumption is that there is an underlying probability distribution for the population values. That is, the values of different members of the population are assumed to be independent random variables having a common distribution. In addition, the sample data are assumed to be independent values from this distribution. Thus, by observing the sample data we are able to learn about this underlying population distribution.

Definition *If $X_1, \ldots, X_n$ are independent random variables having a common probability distribution, we say they constitute a* sample *from that distribution.*

In most applications, the population distribution will not be completely known, and one will attempt to use the sample data to make inferences about it. For instance, a manufacturer may be producing a new type of battery to be used in a particular electric-powered automobile. These batteries will each last for a random number of miles having some unknown probability distribution. To learn about this underlying probability distribution, the manufacturer could build and road-test a set of batteries. The resulting data, consisting of the number of miles of use obtained from each battery, would then constitute a sample from this distribution.

In this chapter we are concerned with the probability distributions of certain statistics that arise from a sample, where a statistic is a numerical quantity whose value is determined by the sample. Two important statistics that we will consider are the sample mean and the sample variance. In Sec. 7.3, we consider the sample mean and present formulas for the expectation and variance of this statistic. We also note that when the sample size is at least moderately large, the probability distribution of the sample mean can be approximated by a normal distribution. This result, which follows from one of the most important theoretical results in probability theory, known as the *central limit theorem*, will be discussed in Sec. 7.4. In Sec. 7.5 we concern ourselves with situations in which we are sampling from a finite population of objects, and we explain what it means for the sample to be a *random* sample. When the population size is large in relation to the sample size, then we often treat the population as if it were infinite. We illustrate and explain exactly when this can be done and what the consequences are. In Sec. 7.6, we consider the distribution of the sample variance from a sample chosen from a normal population.

7.3 SAMPLE MEAN

Consider a population of elements, each of which has a numerical value attached to it. For instance, the population might consist of the adults of a specified community, and the value attached to each adult might be her or his annual

income, or height, or age, and so on. We often suppose that the value associated with any member of the population can be regarded as being the value of a random variable having expectation μ and variance σ^2. The quantities μ and σ^2 are called the *population mean* and the *population variance*, respectively. Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

Since the value of the sample mean $\overline{X}$ is determined by the values of the random variables in the sample, it follows that $\overline{X}$ is also a random variable. Its expectation can be shown to be

$$E[\overline{X}] = \mu$$

That is, the expected value of the sample mean $\overline{X}$ is equal to the population mean μ.

In addition, it can be shown that the variance of the sample mean is

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Thus we see that the sample mean $\overline{X}$ has the same expected value as an individual data value, but its variance is smaller than that of an individual data value by the factor $1/n$, where n is the size of the sample. Therefore, we can conclude that $\overline{X}$ is also centered on the population mean μ, but its spread becomes more and more reduced as the sample size increases. Figure 7.1 plots the probability density

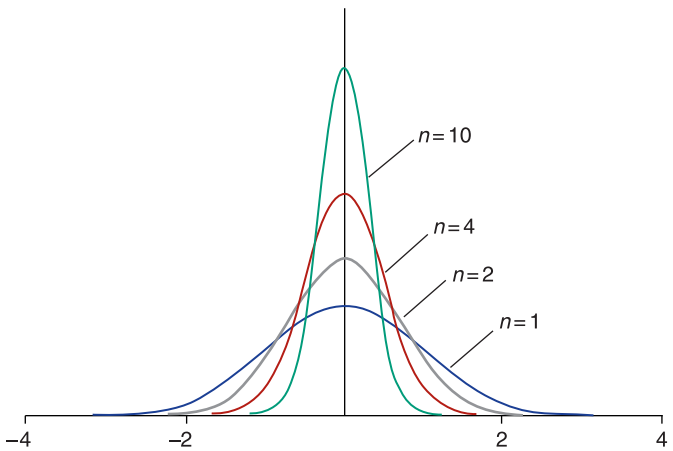

FIGURE 7.1

Densities of sample means from a standard normal population.

function of the sample mean from a standard normal population for a variety of sample sizes.

Example 7.1

Let us check the preceding formulas for the expected value and variance of the sample mean by considering a sample of size 2 from a population whose values are equally likely to be either 1 or 2. That is, if X is the value of a member of the population, then

$$P\{X = 1\} = \frac{1}{2}$$
$$P\{X = 2\} = \frac{1}{2}$$

The population mean and variance are obtained as follows:

$$\mu = E[X] = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{2}\right) = 1.5$$

and

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= E[(X - \mu)^2] \\ &= (1 - 1.5)^2\left(\frac{1}{2}\right) + (2 - 1.5)^2\left(\frac{1}{2}\right) \\ &= \frac{1}{4}\end{aligned}$$

To obtain the probability distribution of the sample mean $(X_1 + X_2)/2$, note that the pair of values X_1, X_2 can assume any of four possible pairs of values

$$(1, 1), (1, 2), (2, 1), (2, 2)$$

where the pair $(2, 1)$ means, for instance, that $X_1 = 2, X_2 = 1$. By the independence of X_1 and X_2 it follows that the probability of any given pair is $1/4$. Therefore, we see that the possible values of $\overline{X} = (X_1 + X_2)/2$ along with their respective probabilities are as follows:

$$P\{\overline{X} = 1\} = P\{(1, 1)\} = \frac{1}{4}$$
$$P\{\overline{X} = 1.5\} = P\{(1, 2) \text{ or } (2, 1)\} = \frac{2}{4} = \frac{1}{2}$$
$$P\{\overline{X} = 2\} = P\{(2, 2)\} = \frac{1}{4}$$

Therefore,

$$E[\overline{X}] = 1\left(\frac{1}{4}\right) + 1.5\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) = \frac{6}{4} = 1.5$$

Also

$$\begin{aligned}\text{Var}(\overline{X}) &= E\left[(\overline{X} - 1.5)^2\right] \\ &= (1 - 1.5)^2\left(\frac{1}{4}\right) + (1.5 - 1.5)^2\left(\frac{1}{2}\right) + (2 - 1.5)^2\left(\frac{1}{4}\right) \\ &= \frac{1}{16} + 0 + \frac{1}{16} = \frac{1}{8}\end{aligned}$$

which, since $\mu = 1.5$ and $\sigma^2 = 1/4$, verifies that $E[\overline{X}] = \mu$ and $\text{Var}(\overline{X}) = \sigma^2/2$.

Figure 7.2 plots the population probability distribution alongside the probability distribution of the sample mean of a sample of size 2. ■

The standard deviation of a random variable, which is equal to the square root of its variance, is a direct indicator of the spread in the distribution. It follows from the identity

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

that $\text{SD}(\overline{X})$, the standard deviation of the sample mean $\overline{X}$, is given by

$$\text{SD}(\overline{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

In the preceding formula, σ is the population standard deviation, and n is the sample size.

The *standard deviation* of the sample mean is equal to the population standard deviation divided by the square root of the sample size.

Summing up, we have seen in this section that the expectation of the sample mean from a sample of size n will equal the population mean, and the variance of the

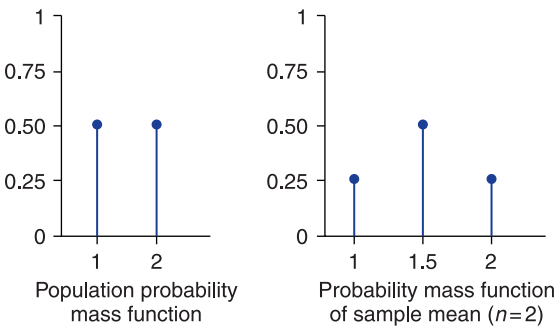

FIGURE 7.2

Probability mass functions.

sample mean will equal the population variance reduced by the factor $1/n$. Now, whereas knowledge of the mean and variance of a statistic tells us quite a bit about its probability distribution, it still leaves much unanswered. We will, however, show in Sec. 7.4 that the probability distribution of a sample mean is approximately normal, and, as we know, a normal distribution is completely specified by its mean and variance.

PROBLEMS

1. Consider the population described in Example 7.1. Plot the possible values along with their probabilities of the sample mean of a sample of size
   (a) $n = 3$
   (b) $n = 4$
   In both cases also derive the standard deviation of the sample mean.
2. Suppose that X_1 and X_2 constitute a sample of size 2 from a population in which a typical value X is equal to either 1 or 2 with respective probabilities

$$P\{X = 1\} = 0.7 \quad P\{X = 2\} = 0.3$$

   (a) Compute $E[X]$.
   (b) Compute $\text{Var}(X)$.
   (c) What are the possible values of $\overline{X} = (X_1 + X_2)/2$?
   (d) Determine the probabilities that $\overline{X}$ assumes the values in (c).
   (e) Using (d), directly compute $E[\overline{X}]$ and $\text{Var}(\overline{X})$.
   (f) Are your answers to (a), (b), and (e) consistent with the formulas presented in this section?
3. Consider a population whose probabilities are given by

$$p(1) = p(2) = p(3) = \frac{1}{3}$$

   (a) Determine $E[X]$.
   (b) Determine $\text{SD}(X)$.
   (c) Let $\overline{X}$ denote the sample mean of a sample of size 2 from this population. Determine the possible values of $\overline{X}$ along with their probabilities.
   (d) Use the result of part (c) to compute $E[\overline{X}]$ and $\text{SD}(\overline{X})$.
   (e) Are your answers consistent?
4. The amount of money withdrawn in each transaction at an automatic teller of a branch of the Bank of America has mean \$80 and standard

deviation $40. What are the mean and standard deviation of the average amount withdrawn in the next 20 transactions?

5. A producer of cigarettes claims that the mean nicotine content in its cigarettes is 2.4 milligrams with a standard deviation of 0.2 milligrams. Assuming these figures are correct, find the expected value and variance of the sample mean nicotine content of
   (a) 36 **(b)** 64 **(c)** 100 **(d)** 900
   randomly chosen cigarettes.
6. The lifetime of a certain type of electric bulb has expected value 475 hours and standard deviation 60 hours. Determine the expected value and standard deviation of the sample mean of
   (a) 100 **(b)** 200 **(c)** 400
   such lightbulbs.
7. The weight of a randomly chosen person riding a ferry has expected value 155 pounds and standard deviation 28 pounds. The ferry has the capacity to carry 100 riders. Find the expected value and standard deviation of the total passenger weight load of a ferry at capacity.

7.4 CENTRAL LIMIT THEOREM

In the previous section we showed that if we take a sample of size n from a population whose elements have mean μ and standard deviation σ then the sample mean $\overline{X}$ will have mean μ and standard deviation $\sigma/\sqrt{n}$. In this section, we consider one of the most important results in probability theory, known as the *central limit theorem*, which states that the sum (and thus also the average) of a large number of independent random variables is approximately normally distributed.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sample from a population having mean μ and standard deviation σ. For n large, the sum

$$X_1 + X_2 + \cdots + X_n$$

will approximately have a normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

■ Example 7.2

An insurance company has 10,000 $(=10^4)$ automobile policyholders. If the expected yearly claim per policyholder is $260 with a standard deviation of $800, approximate the probability that the total yearly claim exceeds $2.8 million $(=\$2.8 \times 10^6)$.

Solution

Number the policyholders, and let X_i denote the yearly claim of policyholder $i, i = 1, \ldots, 10^4$. By the central limit theorem, $X = \sum_{i=1}^{10^4} X_i$ will have an approximately normal distribution with mean $10^4 \times 260 = 2.6 \times 10^6$ and standard deviation $800\sqrt{10^4} = 800 \times 10^2 = 8 \times 10^4$. Hence,

$$\begin{aligned} P\{X > 2.8 \times 10^6\} &= P\left\{\frac{X - 2.6 \times 10^6}{8 \times 10^4} > \frac{2.8 \times 10^6 - 2.6 \times 10^6}{8 \times 10^4}\right\} \\ &\approx P\left\{Z > \frac{0.2 \times 10^6}{8 \times 10^4}\right\} \\ &= P\left\{Z > \frac{20}{8}\right\} \\ &= P\{Z > 2.5\} = 0.0062 \end{aligned}$$

where $\approx$ means "is approximately equal to." That is, there are only 6 chances out of 1000 that the total yearly claim will exceed \$2.8 million. ■

The preceding version of the central limit theorem is by no means the most general, for it can be shown that $\sum_{i=1}^{n} X_i$ will have an approximately normal distribution even in cases where the random variables X_i have different distributions. Indeed, provided that all the random variables tend to be of roughly the same magnitude so that none of them tends to dominate the value of the sum, it can be shown that the sum of a large number of independent random variables will have an approximately normal distribution.

Not only does the central limit theorem give us a method for approximating the distribution of the sum of random variables, but also it helps explain the remarkable fact that the empirical frequencies of so many naturally occurring populations exhibit a bell-shaped (that is, a normal) curve. Indeed, one of the first uses of the central limit theorem was to provide a theoretical justification of the empirical fact that measurement errors tend to be normally distributed. That is, by regarding an error in measurement as being composed of the sum of a large number of small independent errors, the central limit theorem implies that it should be approximately normal. For instance, the error in a measurement in astronomy can be regarded as being equal to the sum of small errors caused by such things as

1. Temperature effects on the measuring device
2. Bending of the device caused by the rays of the sun
3. Elastic effects
4. Air currents
5. Air vibrations
6. Human errors

Therefore, by the central limit theorem, the total measurement error will approximately follow a normal distribution. From this it follows that a histogram of errors resulting from a series of measurements of the *same* object will tend to follow a bell-shaped normal curve.

The central limit theorem also partially explains why many data sets related to biological characteristics tend to be approximately normal. For instance, consider a particular couple, call them Maria and Peter Fontanez, and consider the heights of their daughters (say, when they are 20 years old). Now, the height of a given daughter can be thought of as being composed of the sum of a large number of roughly independent random variables—relating, among other things, to the random set of genes that the daughter received from her parents as well as environmental factors. Since each of these variables plays only a small role in determining the total height, it seems reasonable, based on the central limit theorem, that the height of a Fontanez daughter will be normally distributed. If the Fontanez family has many daughters, then a histogram of their heights should roughly follow a normal curve. (The same thing is true for the sons of Peter and Maria, but the normal curve of the sons would have different parameters from the one of the daughters. The central limit theorem cannot be used to conclude that a plot of the heights of all the Fontanez children would follow a normal curve, since the gender factor does not play a "small" role in determining height.)

Thus, the central limit theorem can be used to explain why the heights of the many daughters of a particular pair of parents will follow a normal curve. However, by itself the theorem does not explain why a histogram of the heights of a collection of daughters from different parents will follow a normal curve. To see why not, suppose that this collection includes both a daughter of Maria and Peter Fontanez and a daughter of Henry and Catherine Silva. By the same argument given before, the height of the Silva daughter will be normally distributed, as will the height of the Fontanez daughter. However, the parameters of these two normal distributions—one for each family—will be different. (For instance, if Catherine and Henry are both around 6 feet tall while Maria and Peter are both about 5 feet tall, then it is clear that the heights of their daughters will have different normal distributions.) By the same reasoning, we can conclude that the heights of a collection of many women, from different families, will all come from different normal distributions. It is, therefore, by no means apparent that a plot of those heights would itself follow a normal curve. (A more complete explanation of why biological data sets often follow a normal curve will be given in Chap. 12.)

7.4.1 Distribution of the Sample Mean

The central limit theorem can be used to approximate the probability distribution of the sample mean. That is, let $X_1, \ldots, X_n$ be a sample from a population having

Historical Perspective

The application of the central limit theorem to show that measurement errors are approximately normally distributed is regarded as an important contribution to science. Indeed, in the 17th and 18th centuries, the central limit theorem was often called the *law of frequency of errors*.

The *law of frequency of errors* was considered a major advance by scientists. Listen to the words of Francis Galton (taken from his book *Natural Inheritance*, published in 1889):

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The Law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

mean μ and variance σ^2, and let

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

be the sample mean. Since a constant multiple of a normal random variable is also normal, it follows from the central limit theorem that $\overline{X}$ (which equals $\sum_{i=1}^{n} X_i$ multiplied by the constant $1/n$) also will be approximately normal when the sample size n is large. Since $\overline{X}$ has expectation μ and standard deviation $\sigma/\sqrt{n}$, the standardized variable

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

has an approximately standard normal distribution.

Let $\overline{X}$ be the sample mean of a sample of size n from a population having mean μ and variance σ^2. By the central limit theorem,

$$P\{\overline{X} \le a\} = P\left\{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le \frac{a - \mu}{\sigma/\sqrt{n}}\right\}$$
$$\approx P\left\{Z \le \frac{a - \mu}{\sigma/\sqrt{n}}\right\}$$

where Z is a standard normal.

Example 7.3

The blood cholesterol levels of a population of workers have mean 202 and standard deviation 14.

(a) If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

(b) Repeat (a) for a sample size of 64.

Solution

(a) It follows from the central limit theorem that $\overline{X}$ is approximately normal with mean $\mu = 202$ and standard deviation $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$. Thus the standardized variable

$$W = \frac{\overline{X} - 202}{7/3}$$

has an approximately standard normal distribution. To compute $P\{198 \leq \overline{X} \leq 206\}$, first we must write the inequality in terms of the standardized variable W. This results in the equality

$$\begin{aligned} P\{198 \leq \overline{X} \leq 206\} &= P\left\{\frac{198 - 202}{7/3} \leq \frac{\overline{X} - 202}{7/3} \leq \frac{206 - 202}{7/3}\right\} \\ &= P\{-1.714 \leq W \leq 1.714\} \\ &\approx P\{-1.714 \leq Z \leq 1.714\} \\ &= 2P\{Z \leq 1.714\} - 1 \\ &= 0.913 \end{aligned}$$

where Z is a standard normal random variable and the final equality follows from Table D.1 in App. D (or from Program 6-1).

(b) For a sample size of 64, the sample mean $\overline{X}$ will have mean 202 and standard deviation $14/\sqrt{64} = 7/4$. Hence, writing the desired probability in terms of the standardized variable

$$\frac{\overline{X} - 202}{7/4}$$

yields

$$\begin{aligned} P\{198 \leq \overline{X} \leq 206\} &= P\left\{\frac{198 - 202}{7/4} \leq \frac{\overline{X} - 202}{7/4} \leq \frac{206 - 202}{7/4}\right\} \\ &\approx P\{-2.286 \leq Z \leq 2.286\} \end{aligned}$$

$$= 2P\{Z \le 2.286\} - 1$$
$$= 0.978$$

Thus, we see that increasing the sample size from 36 to 64 increases the probability that the sample mean will be within 4 of the population mean from 0.913 to 0.978. ■

■ Example 7.4

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for the actual distance. If the values of the measurements constitute a sample from a population having mean d (the actual distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within 0.5 light-years of the actual distance.

Solution

The probability of interest is

$$P\{-0.5 < \overline{X} - d < 0.5\}$$

where $\overline{X}$ is the sample mean of the 10 measurements. Since $\overline{X}$ has mean d and standard deviation $3/\sqrt{10}$, this probability should be written in terms of the standardized variable

$$\frac{\overline{X} - d}{3/\sqrt{10}}$$

This gives

$$\begin{aligned} P\{-0.5 < \overline{X} - d < 0.5\} &= P\left\{\frac{-0.5}{3/\sqrt{10}} < \frac{\overline{X} - d}{3/\sqrt{10}} < \frac{0.5}{3/\sqrt{10}}\right\} \\ &\approx P\left\{\frac{-0.5}{3/\sqrt{10}} < Z < \frac{0.5}{3/\sqrt{10}}\right\} \\ &= P\{-0.527 < Z < 0.527\} \\ &= 2P\{Z < 0.527\} - 1 = 0.402 \end{aligned}$$

Therefore, we see that with 10 measurements there is a 40.2 percent chance that the estimated distance will be within plus or minus 0.5 light-years of the actual distance. ■

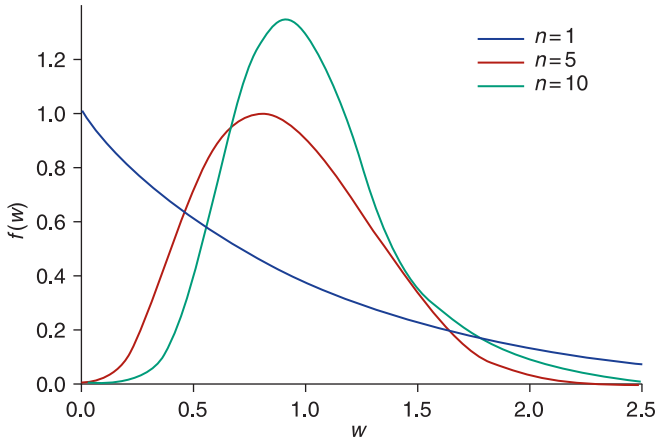

FIGURE 7.3

Density of the average of n exponential random variables.

7.4.2 How Large a Sample Is Needed?

The central limit theorem leaves open the question of how large the sample size n needs to be for the normal approximation to be valid, and indeed the answer depends on the population distribution of the sample data. For instance, if the underlying population distribution is normal, then the sample mean $\overline{X}$ will also be normal, no matter what the sample size is. A general rule of thumb is that you can be confident of the normal approximation whenever the sample size n is at least 30. That is, practically speaking, no matter how nonnormal the underlying population distribution is, the sample mean of a sample size of at least 30 will be approximately normal. In most cases the normal approximation is valid for much smaller sample sizes. Indeed, usually a sample size of 5 will suffice for the approximation to be valid. Figure 7.3 presents the distribution of the sample means from a certain underlying population distribution (known as the *exponential distribution*) for samples sizes $n = 1, 5$, and 10.

Historical Perspective

(Culver)

Pierre Simon, Marquis de Laplace

The central limit theorem was originally stated and proved by the French mathematician Pierre Simon, the Marquis de Laplace, who came to this theorem from his observations that errors of measurement (which usually can be regarded as being the sum of a large number of tiny forces) tend to be normally distributed. Laplace, who was also a famous astronomer (and indeed was called "the Newton of France"), was one of the great early contributors to both probability and statistics. Laplace was a popularizer of the uses of probability in everyday life. He strongly believed in its importance, as is indicated by the following quotation, taken from his published book *Analytical Theory of Probability*.

> We see that the theory of probability is at bottom only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it .... It is remarkable that this science, which originated in the consideration of games of chance, should become the most important object of human knowledge .... The most important questions of life are, for the most part, really only problems of probability.

An interesting footnote to the central limit theorem is that, because of it, most scientists in the late 19th and early 20th centuries believed that almost all data sets were normal. In the words of the famous French mathematician Henri Poincaré,

> Everyone believes it: experimentalists believe that it is a mathematical theorem, and mathematicians believe that it is an empirical fact.

PROBLEMS

1. Consider a sample from a population having mean 128 and standard deviation 16. Compute the approximate probability that the sample mean will lie between 124 and 132 when the sample size is
   (a) $n = 9$
   (b) $n = 25$
   (c) $n = 100$
2. Frequent fliers of a particular airline fly a random number of miles each year, having mean and standard deviation (in thousands of miles) of 23 and 11, respectively. As a promotional gimmick, the airline has decided to randomly select 20 of these fliers and give them, as a bonus, a check of \$10 for each 1000 miles flown. Approximate the probability that the total amount paid out is
   (a) Between \$4500 and \$5000
   (b) More than \$5200
3. In Example 7.2, approximate the probability that the yearly payout of the insurance company is between \$2.5 and \$2.7 million.
4. If you place a \$1 bet on a number of a roulette wheel, then either you win \$35, with probability 1/38, or you lose \$1, with probability 37/38. Let X denote your gain on a bet of this type.
   (a) Find $E[X]$ and $\mathrm{SD}(X)$.
   Suppose you continually place bets of the preceding type. Show that
   (b) The probability that you will be winning after 1000 bets is approximately 0.39.
   (c) The probability that you will be winning after 100,000 bets is approximately 0.002.

5. The time it takes to develop a photographic print is a random variable with mean 17 seconds and standard deviation 0.8 seconds. Approximate the probability that the total amount of time that it takes to process 100 prints is
   (a) More than 1720 seconds
   (b) Between 1690 and 1710 seconds
6. A zircon semiconductor is critical to the operation of a superconductor and must be immediately replaced upon failure. Its expected lifetime is 100 hours, and its standard deviation is 34 hours. If 22 of these semiconductors are available, approximate the probability that the superconductor can operate for the next 2000 hours. (That is, approximate the probability that the sum of the 22 lifetimes exceeds 2000.)
7. The amount of paper a print shop uses per job has mean 200 pages and standard deviation 50 pages. There are 2300 sheets of paper on hand and 10 jobs that need to be filled. What is the approximate probability that 10 jobs can be filled with the paper on hand?
8. A highway department has enough salt to handle a total of 80 inches of snowfall. Suppose the daily amount of snow has a mean of 1.5 inches and a standard deviation of 0.3 inches.
   (a) Approximate the probability that the salt on hand will suffice for the next 50 days.
   (b) What assumption did you make in solving part (a)?
   (c) Do you think this assumption is justified? Explain briefly!
9. Fifty numbers are rounded off to the nearest integer and then summed. If the individual roundoff errors are uniformly distributed between –0.5 and 0.5, what is the approximate probability that the resultant sum differs from the exact sum by more than 3? (Use the fact that the mean and variance of a random variable that is uniformly distributed between –0.5 and 0.5 are 0 and 1/12, respectively.)
10. A six-sided die, in which each side is equally likely to appear, is repeatedly rolled until the total of all rolls exceeds 400. What is the approximate probability that this will require more than 140 rolls? (*Hint*: Relate this to the probability that the sum of the first 140 rolls is less than 400.)
11. In Example 7.4, approximate the probability that the astronomer's estimate will be within 0.5 light-years of the true distance if
   (a) She makes a total of 100 observations.
   (b) She makes 10 observations but has found a way of improving the measurement technique so that the standard deviation of each observation is reduced from 3 to 2 light-years.
12. Suppose that the number of miles that an electric car battery functions has mean μ and standard deviation 100. Using the central limit

theorem, approximate the probability that the average number of miles per battery obtained from a set of n batteries will differ from μ by more than 20 if

(a) $n = 10$ **(b)** $n = 20$ **(c)** $n = 40$ **(d)** $n = 100$

13. A producer of cigarettes claims that the mean nicotine content in its cigarettes is 2.4 milligrams with a standard deviation of 0.2 milligrams. Assuming these figures are correct, approximate the probability that the sample mean of 100 randomly chosen cigarettes is

(a) Greater than 2.5 milligrams

(b) Less than 2.25 milligrams

14. The lifetime of a certain type of electric bulb has expected value 500 hours and standard deviation 60 hours. Approximate the probability that the sample mean of 20 such lightbulbs is less than 480 hours.

15. Consider a sample of size 16 from a population having mean 100 and standard deviation σ. Approximate the probability that the sample mean lies between 96 and 104 when

(a) $\sigma = 16$ **(b)** $\sigma = 8$ **(c)** $\sigma = 4$ **(d)** $\sigma = 2$ **(e)** $\sigma = 1$

16. An instructor knows from past experience that student examination scores have mean 77 and standard deviation 15. At present, the instructor is teaching two separate classes—one of size 25 and the other of size 64.

(a) Approximate the probability that the average test score in the class of size 25 lies between 72 and 82.

(b) Repeat (a) for the class of size 64.

(c) What is the approximate probability that the average test score in the class of size 25 is higher than that in the class of size 64?

(d) Suppose the average scores in the two classes are 76 and 83. Which class—the one of size 25 or the one of size 64—do you think was more likely to have averaged 83? Explain your intuition.

7.5 SAMPLING PROPORTIONS FROM A FINITE POPULATION

Consider a population of size N in which certain elements have a particular characteristic of interest. Let p denote the proportion of the population having this characteristic. So Np elements of the population have it and $N(1 - p)$ do not.

■ Example 7.5

Suppose that 60 out of a total of 900 students of a particular school are left-handed. If left-handedness is the characteristic of interest, then $N = 900$ and $p = 60/900 = 1/15$. ■

A sample of size n is said to be a *random sample* if it is chosen in a manner so that each of the possible population subsets of size n is equally likely to be in the sample. For instance, if the population consists of the three elements a, b, c, then a random sample of size 2 is one chosen so that it is equally likely to be any of the subsets $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$. A random subset can be chosen sequentially by letting its first element be equally likely to be any of the N elements of the population, then letting its second element be equally likely to be any of the remaining $N - 1$ elements of the population, and so on.

Definition *A sample of size* n *selected from a population of* N *elements is said to be a* random sample *if it is selected in such a manner that the sample chosen is equally likely to be any of the subsets of size* n.

The mechanics of using a computer to choose a random sample are explained in App. C. (In addition, Program A-1 on the enclosed disk can be used to accomplish this task.)

Suppose now that a random sample of size n has been chosen from a population of size N. For $i = 1, \ldots, n$, let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th member of the sample has the characteristic} \\ 0 & \text{otherwise} \end{cases}$$

Consider now the sum of the X_i; that is, consider

$$X = X_1 + X_2 + \cdots + X_n$$

Since the term X_i contributes 1 to the sum if the ith member of the sample has the characteristic and contributes 0 otherwise, it follows that the sum is equal to the number of members of the sample that possess the characteristic. (For instance, suppose $n = 3$ and $X_1 = 1, X_2 = 0$, and $X_3 = 1$. Then members 1 and 3 of the sample possess the characteristic, and member 2 does not. Hence, exactly 2 of the sample members possess it, as indicated by $X_1 + X_2 + X_3 = 2$.) Similarly, the sample mean

$$\overline{X} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

will equal the *proportion* of members of the sample who possess the characteristic. Let us now consider the probabilities associated with the statistic $\overline{X}$.

Since the ith member of the sample is equally likely to be any of the N members of the population, of which Np have the characteristic, it follows that

$$P\{X_i = 1\} = \frac{Np}{N} = p$$

Also

$$P\{X_i = 0\} = 1 - P\{X_i = 1\} = 1 - p$$

That is, each X_i is equal to either 1 or 0 with respective probabilities p and $1 - p$.

Note that the random variables $X_1, X_2, \ldots, X_n$ are not independent. For instance, since the second selection is equally likely to be any of the N members of the population, of which Np have the characteristic, it follows that the probability that the second selection has the characteristic is $Np/N = p$. That is, without any knowledge of the outcome of the first selection,

$$P\{X_2 = 1\} = p$$

However, the conditional probability that $X_2 = 1$, given that the first selection has the characteristic, is

$$P\{X_2 = 1|X_1 = 1\} = \frac{Np - 1}{N - 1}$$

which is seen by noting that if the first selection has the characteristic, then the second selection is equally likely to be any of the remaining $N - 1$ elements of which $Np - 1$ have the characteristic. Similarly, the probability that the second selection has the characteristic, given that the first one does not, is

$$P\{X_2 = 1|X_1 = 0\} = \frac{Np}{N - 1}$$

Thus, knowing whether the first element of the random sample has the characteristic changes the probability for the next element. However, when the population size N is large in relation to the sample size n, this change will be very slight. For instance, if $N = 1000$ and $p = 0.4$, then

$$P\{X_2 = 1|X_1 = 1\} = \frac{399}{999} = 0.3994$$

which is very close to the unconditional probability that $X_2 = 1$; namely,

$$P\{X_2 = 1\} = 0.4$$

Similarly, the probability that the second element of the sample has the characteristic, given that the first does not, will be given by

$$P\{X_2 = 1|X_1 = 0\} = \frac{400}{999} = 0.4004$$

which is again very close to 0.4.

Indeed, it can be shown that when the population size N is large with respect to the sample size n, then $X_1, X_2, \ldots, X_n$ are approximately independent. Now if we think of each X_i as representing the result of a trial that is a success if X_i equals 1 and a failure otherwise, it follows that $\sum_{i=1}^{n} X_i$ can be thought of as representing the total number of successes in n trials. Hence, if the X's are independent, then X represents the number of successes in n independent trials, where each trial is a success with probability p. In other words, X is a binomial random variable with parameters n and p.

If we let X denote the number of members of the population who have the characteristic, then it follows from the preceding that if the population size N is large in relation to the sample size n, then the distribution of X is approximately a binomial distribution with parameters n and p.

For the remainder of this text we will suppose that the underlying population is large in relation to the sample size, and we will take the distribution of X to be binomial.

By using the formulas given in Sec. 5.5.1 for the mean and standard deviation of a binomial random variable, we see that

$$E[X] = np \quad \text{and} \quad \text{SD}(X) = \sqrt{np(1-p)}$$

Since $\overline{X}$, the proportion of the sample that has the characteristic, is equal to X/n, we see that

$$E[\overline{X}] = \frac{E[X]}{n} = p$$

and

$$\text{SD}(\overline{X}) = \frac{\text{SD}(X)}{n} = \sqrt{\frac{p\,(1-p)}{n}}$$

Example 7.6

Suppose that 50 percent of the population is planning on voting for candidate A in an upcoming election. If a random sample of size 100 is chosen, then the proportion of those in the sample who favor candidate A has expected value

$$E[\overline{X}] = 0.50$$

and standard deviation

$$\text{SD}(\overline{X}) = \sqrt{\frac{0.50(1-0.50)}{100}} = \sqrt{\frac{1}{400}} = 0.05$$

■

7.5.1 Probabilities Associated with Sample Proportions: The Normal Approximation to the Binomial Distribution

Again, let $\overline{X}$ denote the proportion of members of a random sample of size n who have a certain characteristic. To determine the probabilities connected with the random variable $\overline{X}$, we make use of the fact that $X = n\overline{X}$ is binomial with parameters n and p. Now, binomial probabilities can be approximated by making use of the central limit theorem. Indeed, from an historical point of view, one of the most important applications of the central limit theorem was in computing binomial probabilities.

To see how this is accomplished, let X denote a binomial random variable having parameters n and p. Since X can be thought of as being equal to the number of successes in n independent trials when each trial is a success with probability p, it follows that it can be represented as

$$X = X_1 + X_2 + \cdots + X_n$$

where

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{if trial } i \text{ is a failure} \end{cases}$$

Now, in Examples 5.6 and 5.12, we showed that

$$E[X_i] = P \quad \text{and} \quad \text{Var}(X_i) = p(1-p)$$

Hence, it follows that X/n can be regarded as the sample mean of a sample of size n from a population having mean p and standard deviation $\sqrt{p(1-p)}$. Thus, from the central limit theorem, we see that for n large,

$$\frac{X/n - p}{\sqrt{p(1-p)/n}} = \frac{X - np}{\sqrt{np(1-p)}}$$

will have an approximately standard normal distribution. (Figure 7.4 graphically illustrates how the probability distribution of a binomial random variable with parameters n and p becomes more and more "normal" as n becomes larger and larger.)

From a practical point of view, the normal approximation to the binomial is quite good provided n is large enough that the quantities np and $n(1-p)$ are both greater than 5.

■ Example 7.7

Suppose that exactly 46 percent of the population favors a particular candidate. If a random sample of size 200 is chosen, what is the probability that at least 100 favor this candidate?

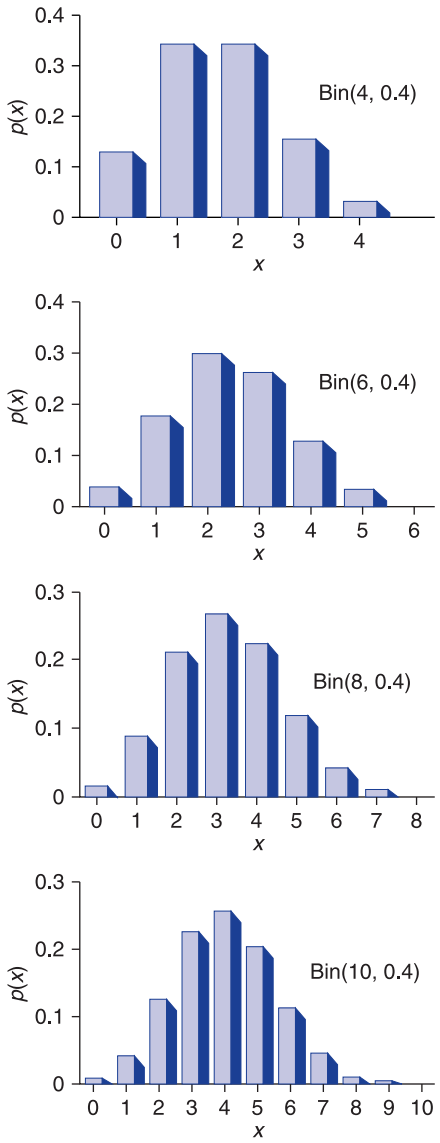

FIGURE 7.4

Probability mass functions of binomial random variables become more normal with increasing n.

Solution

If X is the number who favor the candidate, then X is a binomial random variable with parameters $n = 200$ and $p = 0.46$. The desired probability is $P\{X \geq 100\}$. To employ the normal approximation, first we note that since the binomial is a discrete and the normal is a continuous random variable,

it is best to compute $P\{X = i\}$ as $P\{i - 0.5 \le X \le i + 0.5\}$ when applying the normal approximation (this is called the *continuity correction*). Therefore, to compute $P\{X \ge 100\}$, we should use the normal approximation on the equivalent probability $P\{X \ge 99.5\}$. Considering the standardized variable

$$\frac{X - 200(0.46)}{\sqrt{200(0.46)(0.54)}} = \frac{X - 92}{7.0484}$$

we obtain the following normal approximation to the desired probability:

$$\begin{aligned} P\{X \ge 100\} &= P\{X \ge 99.5\} \\ &= P\left\{\frac{X - 92}{7.0484} \ge \frac{99.5 - 92}{7.0484}\right\} \\ &\approx P\{Z > 1.0641\} \\ &= 0.144 \text{(from Table D.1 or Program 6-1)} \end{aligned}$$

The exact value of the desired probability could, of course, have been obtained from Program 5-1. Running this program shows that the exact probability that a binomial random variable with parameters $n = 200$ and $p = 0.46$ is greater than or equal to 100 is 0.1437. Thus, in this problem, the normal approximation gives an answer that is correct to three decimal places. ■

PROBLEMS

1. Suppose that 60 percent of the residents of a city are in favor of teaching evolution in high school. Determine the mean and the standard deviation of the proportion of a random sample of size n that is in favor when
   (a) $n = 10$ **(b)** $n = 100$
   (c) $n = 1,000$ **(d)** $n = 10,000$
2. Ten percent of all electrical batteries are defective. In a random selection of 8 of these batteries, find the probability that
   (a) There are no defective batteries.
   (b) More than 15 percent of the batteries are defective.
   (c) Between 8 and 12 percent of the batteries are defective.
3. Suppose there was a random selection of $n = 50$ batteries in Prob. 2. Determine approximate probabilities for parts (a), (b), and (c) of that problem.
4. Consider Prob. 1. Find the probability that over 55 percent of the members of the sample are in favor of the proposal if the sample size is
   (a) $n = 10$ **(b)** $n = 100$
   (c) $n = 1000$ **(d)** $n = 10{,}000$

*A Cautionary Tale: Be Sure you are Sampling From the Right Population

Company X, which is not located near any public transportation and all of whose employees drive to work, is concerned that not enough people are utilizing carpools. The company has decided that if the average number of workers per car is less than 3, then it will organize its own carpool service and, at the same time, begin charging those employees who drive noncarpool automobiles a stiff parking fee.To determine if such a change is justified, 100 workers were chosen at random and were queried as to the number of workers in the car in which they drove to work that day. The average answer was 3.4; that is, the sum of the 100 answers divided by 100 was 3.4. On the basis of this, the company chose not to change its policy. Did the company make the correct decision?

This question is very tricky because the company, when selecting its random sample of 100 workers, has chosen a random sample from the wrong population. Since it wanted to learn about the average number of workers per car, the company should have chosen a random sample from the population of cars arriving in the parking lot—not from the population of workers. To see why, consider an extreme case where there are only 2 cars and 5 workers, with one of the cars containing 4 workers and the other containing 1 worker. Now, if we average over the 2 cars, then the average number of workers per car is clearly $(4 + 1)/2 = 2.5$. However, if we average over all the workers, then since 4 of the 5 workers ride in a car containing 4 workers, it follows that the average is $(4 + 4 + 4 + 4 + 1)/5 = 3.4$.

In general, by randomly choosing workers (as opposed to cars) it follows that cars containing more riders will tend to be more heavily represented (by their riders) in the sample than will those cars having fewer riders. As a result, the average number of riders in the cars of the randomly chosen workers will tend to be larger than the average number of workers per car.

To obtain a correct estimate of the average number of workers per car, the random sample should have been created by randomly choosing among the cars in the parking lot and then ascertaining how many workers were in each car.

Because the wrong random sample was chosen, the company cannot conclude that the average number of workers per car is at least 3. Indeed, a new sample chosen in the manner just noted will have to be taken before the company can decide whether any changes are needed.

The following table gives the 2003 first-quarter unemployment rates for a selection of countries. Problems 5, 6, and 7 are based on it.

United States	Australia	Canada	Germany	Italy	Japan	Sweden
6.2	6.1	6.9	9.2	8.9	5.4	6.1

5. Suppose that a random sample of 400 German workers was selected. Approximate the probability that

(a) Forty or fewer were unemployed
(b) More than 50 were unemployed

6. Suppose that a random sample of 600 Japanese workers was selected. Approximate the probability that
   (a) Thirty or fewer were unemployed
   (b) More than 40 were unemployed
7. Suppose that a random sample of 200 Canadian workers was selected. Approximate the probability that
   (a) Ten or fewer were unemployed
   (b) More than 25 were unemployed
8. If 65 percent of the population of a certain community is in favor of a proposed increase in school taxes, find the approximate probability that a random sample of 100 people will contain
   (a) At least 45 who are in favor of the proposition
   (b) Fewer than 60 who are in favor
   (c) Between 55 and 75 who are in favor
9. The ideal size of a first-year class at a particular college is 160 students. The college, from past experience, knows that on average only 40 percent of those accepted for admission will actually attend. Based on this, the college employs a policy of initially accepting 350 applicants. Find the normal approximation to the probability that this will result in
   (a) More than 160 accepted students attending
   (b) Fewer than 150 accepted students attending
10. An airline company experiences a 6 percent rate of no-shows among passengers holding reservations. If 260 people hold reservations on a flight in which the airplane can hold a maximum of 250 people, approximate the probability that the company will be able to accommodate everyone having a reservation who shows up.

The following table lists the likely fields of study as given by the entering college class of a large university system. Problems 11 through 14 are based on this table. In each of these problems suppose that a random sample of 200 entering students is chosen.

Field of study	Percentage
Arts and humanities	9
Biological sciences	4
Business	27
Education	9
Engineering	10

(Continued)

(Continued)

Field of study	Percentage
Physical sciences	2
Social sciences	9
Professional	11
Technical	3
Other	16

Source: Higher Educational Institute, University of California, Los Angeles, CA, *The American Freshman National Norms,* annual.

11. What is the probability that 22 or more students are planning to major in arts and humanities?

12. What is the probability that more than 60 students are planning to major in business?

13. What is the probability that 30 or more are planning to major in one of the sciences (biological, physical, or social)?

14. What is the probability that fewer than 15 students are planning to major in engineering?

15. Let X be a binomial random variable with parameters $n = 100$ and $p = 0.2$. Approximate the following probabilities.

(a) $P\{X \leq 25\}$

(b) $P\{X > 30\}$

(c) $P\{15 < X < 22\}$

16. Let X be a binomial random variable with parameters $n = 150$ and $p = 0.6$. Approximate the following probabilities.

(a) $P\{X \leq 100\}$

(b) $P\{X > 75\}$

(c) $P\{80 < X < 100\}$

17. A recent study has shown that 54 percent of all incoming first-year students at major universities do not graduate within 4 years of their entrance. Suppose a random sample of 500 entering first-year students is to be surveyed after 4 years.

(a) What is the approximate probability that fewer than half graduate within 4 years?

(b) What is the approximate probability that more than 175 but fewer than 225 students graduate within 4 years?

The following table gives the percentages of individuals, categorized by gender, who follow certain negative health practices. Problems 18, 19, and 20 are based on this table.

	Sleep 6 hours or less per night	Smoker	Never eat breakfast	Are 30% or more overweight
Males	22.7	32.6	25.2	12.1
Females	21.4	27.8	23.6	13.7

Source: U.S. National Center for Health Statistics, *Health Promotion and Disease Prevention*. 1985.

18. Suppose a random sample of 300 males is chosen. Approximate the probability that
   (a) At least 75 never eat breakfast.
   (b) Fewer than 100 smoke.

19. Suppose a random sample of 300 females is chosen. Approximate the probability that
   (a) At least 25 are overweight by 30 percent or more.
   (b) Fewer than 50 sleep 6 hours or less nightly.

20. Suppose random samples of 300 females and 300 males are chosen. Approximate the probability that there are more smokers in the sample of men than in the sample of woman. (*Hint*: Let X and Y denote, respectively, the numbers of men and women in the samples who are smokers. Write the desired probability as $P\{X - Y > 0\}$, and recall that the difference of two independent normal random variables is also a normal random variable.)

7.6 DISTRIBUTION OF THE SAMPLE VARIANCE OF A NORMAL POPULATION

Before discussing the distribution of the sample variance of a normal population, we need to introduce the concept of the chi-squared distribution, which is the distribution of the sum of the squares of independent standard normal random variables.

Definition *If* $Z_1, \ldots, Z_n$ *are independent standard normal random variables, then the random variable*

$$\sum_{i=1}^{n} Z_i^2$$

is said to be a chi-squared *random variable with* n degrees of freedom.

Figure 7.5 presents the chi-squared density functions for three different values of the degree of freedom parameter n.

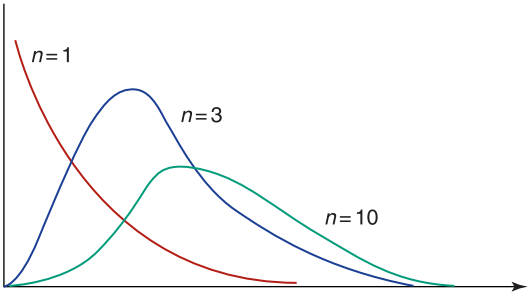

FIGURE 7.5

Chi-squared density function with n degrees of freedom, $n = 1, 3, 10$.

To determine the expected value of a chi-squared random variable, note first that for a standard normal random variable Z,

$$\begin{aligned} 1 &= \text{Var}(Z) \\ &= E\left[Z^2\right] - (E[Z])^2 \\ &= E\left[Z^2\right] \quad \text{since } E[Z] = 0 \end{aligned}$$

Hence, $E[Z^2] = 1$ and so

$$E\left[\sum_{i=1}^{n} Z_i^2\right] = \sum_{i=1}^{n} E\left[Z_i^2\right] = n$$

The expected value of a chi-squared random variable is equal to its number of degrees of freedom.

Suppose now that we have a sample $X_1, \ldots, X_n$ from a normal population having mean μ and variance σ^2. Consider the sample variance S^2 defined by

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

The following result can be proved:

Theorem

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2}$$

has a chi-squared distribution with $n-1$ degrees of freedom.

Although a mathematical proof of this theorem is beyond the scope of this text, we can obtain some understanding of why it is true. This understanding will also

be useful in guiding our intuition as we continue our studies in later chapters. To begin, let us consider the standardized variables $(X_i - \mu)/\sigma, i = 1, \ldots, n$, where μ is the population mean. Since these variables are independent standard normals, it follows that the sum of their squares,

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2}$$

has a chi-squared distribution with n degrees of freedom. Now, if we substitute the sample mean $\overline{X}$ for the population mean μ, then the new quantity,

$$\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2}$$

will remain a chi-squared random variable, but it will lose 1 degree of freedom because the population mean (μ) is replaced by its estimator (the sample mean $\overline{X}$).

PROBLEMS

1. The following data sets come from normal populations whose standard deviation σ is specified. In each case, determine the value of a statistic whose distribution is chi-squared, and tell how many degrees of freedom this distribution has.
   - **(a)** 104, 110, 100, 98, 106; $\sigma = 4$
   - **(b)** 1.2, 1.6, 2.0, 1.5, 1.3, 1.8; $\sigma = 0.5$
   - **(c)** 12.4, 14.0, 16.0; $\sigma = 2.4$
2. Explain why a chi-squared random variable having n degrees of freedom will approximately have the distribution of a normal random variable when n is large. (*Hint*: Use the central limit theorem.)

KEY TERMS

A sample from a population distribution: If $X_1, \ldots, X_n$ are independent random variables having a common distribution F, we say that they constitute a sample from the population distribution F.

Statistic: A numerical quantity whose value is determined by the sample.

Sample mean: If $X_1, \ldots, X_n$ are a sample, then the sample mean is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample variance: If $X_1, \ldots, X_n$ are a sample, then the sample variance is

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

Central limit theorem: A theorem stating that the sum of a sample of size n from a population will approximately have a normal distribution when n is large.

Random sample: A sample of n members of a population is a random sample if it is obtained in such a manner that each of the possible subsets of n members is equally likely to be the chosen sample.

Chi-squared distribution with n degrees of freedom: The distribution of the sum of the squares of n independent standard normals.

SUMMARY

If $\overline{X}$ is the sample mean of a sample of size n from a population having mean μ and standard deviation σ, then its mean and standard deviation are

$$E[\overline{X}] = \mu \quad \text{and} \quad \text{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

The central limit theorem states that the sample mean of a sample of size n from a population having mean μ and standard deviation σ will, for large n, have an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Consider a random sample of size n from a population of N individuals in which Np of them have a certain characteristic. Let X denote the number of members of the sample who have the characteristic. When N is large in relation to n, X will be an approximately binomial random variable with parameters n and p. In this text, we will always suppose that this is the case.

The proportion of the sample having the characteristic, namely, $\overline{X} = X/n$, has a mean and a standard deviation given by

$$E[\overline{X}] = p \quad \text{and} \quad \text{SD}(\overline{X}) = \sqrt{\frac{p(1-p)}{n}}$$

It follows from the central limit theorem that a binomial random variable with parameters n and p can, for reasonably large n, be approximated by a normal random variable with mean np and standard deviation $\sqrt{np(1-p)}$. The approximation should be quite accurate provided that n is large enough that both np and $n(1-p)$ are larger than 5.

If S^2 is the sample variance from a sample of size n from a normal population having variance σ^2, then $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom.

The expected value of a chi-squared random variable is equal to its number of degrees of freedom.

REVIEW PROBLEMS

1. The sample mean and sample standard deviation of all student scores on the last Scholastic Aptitude Test (SAT) examination were, respectively, 517 and 120. Find the approximate probability that a random sample of 144 students would have an average score exceeding
   (a) 507
   (b) 517
   (c) 537
   (d) 550
2. Let $\overline{X}$ denote the sample mean of a sample of size 10 from a population whose probability distribution is given by

$$P\{X = i\} = \begin{cases} 0.1 & \text{if } i = 1 \\ 0.2 & \text{if } i = 2 \\ 0.3 & \text{if } i = 3 \\ 0.4 & \text{if } i = 4 \end{cases}$$

   Compute
   (a) The population mean μ
   (b) The population standard deviation σ
   (c) $E[\overline{X}]$
   (d) $\text{Var}(\overline{X})$
   (e) $\text{SD}(\overline{X})$
3. In Prob. 2, suppose the sample size was 2. Find the probability distribution of $\overline{X}$, and use it to compute $E[\overline{X}]$ and $\text{SD}(\overline{X})$. Check your answers by using the values of μ and σ.
4. The mean and standard deviation of the lifetime of a type of battery used in electric cars are, respectively, 225 and 24 minutes. Approximate the probability that a set of 10 batteries, used one after the other, will last for more than
   (a) 2200 minutes
   (b) 2350 minutes
   (c) 2500 minutes
   (d) What is the probability they will last between 2200 and 2350 minutes?

5. Suppose that 12 percent of the members of a population are left-handed. In a random sample of 100 individuals from this population,
   (a) Find the mean and standard deviation of the number of left-handed people.
   (b) Find the probability that this number is between 10 and 14 inclusive.
6. The weight of a randomly chosen person riding a ferry has expected value 155 and standard deviation 28 pounds. The ferry's capacity is 100 riders. Find the probability that, at capacity, the total passenger load exceeds 16,000 pounds.
7. The monthly telephone bill of a student residing in a dormitory has an expected value of \$15 with a standard deviation of \$7. Let X denote the sum of the monthly telephone bills of a sample of 20 such students.
   (a) What is $E[X]$?
   (b) What is $\mathrm{SD}(X)$?
   (c) Approximate the probability that X exceeds \$300.
8. A recent newspaper article claimed that the average salary of newly graduated seniors majoring in chemical engineering is \$54,000, with a standard deviation of \$5000. Suppose a random sample of 12 such graduates revealed an average salary of \$45,000. How likely is it that an average salary as low as or lower than \$45,000 would have been observed from this sample if the newspaper article were correct?
9. An advertising agency ran a campaign to introduce a product. At the end of its campaign, it claimed that at least 25 percent of all consumers were now familiar with the product. To verify this claim, the producer randomly sampled 1000 consumers and found that 232 knew of the product. If 25 percent of all consumers actually knew of the product, what is the probability that as few as 232 (that is, 232 or less) in a random sample of 1000 consumers were familiar?
10. A club basketball team will play a 60-game season. Of these games 32 are against class A teams and 28 are against class B teams. The outcomes of all the games are independent. The team will win each game against a class A opponent with probability 0.5, and it will win each game against a class B opponent with probability 0.7. Let X denote the total number of victories in the season.
    (a) Is X a binomial random variable?
    (b) Let X_A and X_B denote, respectively, the number of victories against class A and class B teams. What are the distributions of X_A and X_B?
    (c) What is the relationship among X_A, X_B, and X?
    (d) Approximate the probability that the team wins 40 or more games. (*Hint*: Recall that the sum of independent normal random variables is also a normal random variable.)

11. If X is binomial with parameters $n = 80$ and $p = 0.4$, approximate the following probabilities.
- **(a)** $P\{X > 34\}$
- **(b)** $P\{X \leq 42\}$
- **(c)** $P\{25 \leq X \leq 39\}$

12. Consider the following simple model for daily changes in price of a stock. Suppose that on each day the price either goes up 1 with probability 0.52 or goes down 1 with probability 0.48. Suppose the price at the beginning of day 1 is 200. Let X denote the price at the end of day 100.
- **(a)** Define random variables $X_1, X_2, \ldots, X_{100}$ such that

$$X = 200 + \sum_{i=1}^{100} X_i$$

- **(b)** Determine $E[X_i]$
- **(c)** Determine $\text{Var}(X_i)$
- **(d)** Use the central limit theorem to approximate $P\{X \geq 210\}$

13. The following are the percentages of U.S. residents, classified by age, who were not covered by health insurance in 2002.

Age	Percentage not covered
under 18	11.6
18 to 24	29.6
25 to 34	24.9
35 to 44	17.7
45 to 64	13.5
65 and over	0.8

Suppose random samples of 1000 people in each age category are selected. Approximate the probability that
- **(a)** At least 100 of those under 18 are not covered.
- **(b)** Fewer than 260 of those 25 to 34 years old are uncovered.
- **(c)** At most 5 of those 65 and over and at most 120 of those 45 to 64 years old are uncovered.
- **(d)** More of those who are 18 to 24 years old than of those who are 25 to 34 years old are uncovered.

14. A university administrator wants a quick estimate of the average number of students enrolled per class. Because he does not want the faculty to be aware of his interest, he has decided to enlist the aid of students. He has decided to randomly choose 100 names from the roster of students and have them determine and then report to him the number

of students in each of their classes. His estimate of the average number of students per class will be the average number reported per class.

(a) Will this method achieve the desired goal?

(b) If the answer to part (a) is yes, explain why. If it is no, give a method that will work.