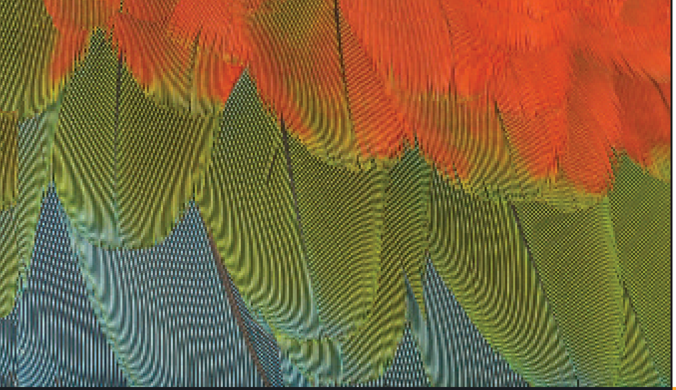


7

7.1 Sampling Distributions

7.2 The Central Limit Theorem

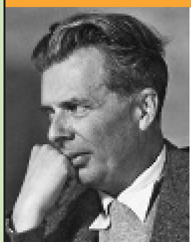
7.3 Sampling Distributions for Proportions

No one wants to learn by mistakes, but we cannot learn enough from success to go beyond the state of the art. . . . Such is the nature not only of science and engineering, but of all human endeavors.

—Henry Petroski

Experience isn't what happens to you. It's what you make out of what happens to you.

—Aldous Huxley



Henry Petroski is a professor of engineering at Duke University, and Aldous Huxley (1894–1963) was a well-known modern writer. Both quotes imply that experience, mistakes, information, and life itself are closely related. Life and uncertainty appear to be inseparable. Only those who are no longer living can escape chance happenings. Mistakes are bound to occur. However, not all mistakes are bad. The discovery of penicillin was a "mistake" when mold (penicillin) was accidentally introduced into a bacterial culture by Alexander Fleming in 1928.

Most of the really important decisions in life will involve incomplete information. In one lifetime, we simply cannot experience *everything*. Nor should we even want to! This is one reason why *experience by way of sampling* is so important. Statistics can help you have the experiences and yet maintain some control over mistakes. Remember, it is what you make out of experience (sample data) that is of real value.

In this chapter, we will study how information from samples relates to information about populations. We cannot be certain that the information from a sample reflects corresponding information about the entire population, but we can describe likely differences. Study this chapter and the following material carefully. We believe that your effort will be rewarded by helping you appreciate the joy and wonder of living in an uncertain universe.

For on-line student resources, visit the Brase/Brase, *Understandable Statistics*, 9th edition web site at **college.hmco.com/pic/braseUS9e**.

Introduction to Sampling Distributions

PREVIEW QUESTIONS

As humans, our experiences are finite and limited. Consequently, most of the important decisions in our lives are based on sample (incomplete) information. What is a probability sampling distribution? How will sampling distributions help us make good decisions based on incomplete information? (Section 7.1)

There is an old saying: All roads lead to Rome. In statistics, we could recast this saying: All probability distributions average out to be normal distributions (as the sample size increases). How can we take advantage of this in our study of sampling distributions? (Section 7.2)

Many issues in life come down to success or failure. In most cases, we will not be successful all the time, so proportions of successes are very important. What is the probability sampling distribution for proportions? (Section 7.3)

FOCUS PROBLEM

Impulse Buying

The Food Marketing Institute, Progressive Grocer, New Products News, and Point of Purchaser Advertising Institute are organizations that analyze supermarket sales. One of the interesting discoveries was that the average amount of impulse buying in a grocery store was very time-dependent. As reported in the *Denver Post,* "when you dilly dally in a store for 10 unplanned minutes, you can kiss nearly \$20 good-bye." For this reason, it is in the best interest of the supermarket to keep you in the store longer. In the *Post* article, it was pointed out that long checkout lines (near end-aisle displays), "samplefest" events of tasting free samples, video kiosks, magazine and book sections, and so on help keep customers in the store longer. On average, a single customer who strays from his or her grocery list can plan on impulse spending of \$20 for every 10 minutes spent wandering about in the supermarket.

Let x represent the dollar amount spent on supermarket impulse buying in a 10-minute (unplanned) shopping interval. Based on the *Post* article, the mean of the x distribution is about \$20 and the (estimated) standard deviation is about \$7.

(a) Consider a random sample of $n = 100$ customers, each of whom has 10 minutes of unplanned shopping time in a supermarket. From the central limit theorem, what can you say about the probability distribution of $\bar{x}$, the *average* amount spent by these customers due to impulse buying? Is the $\bar{x}$ distribution approximately normal? What are the mean and standard deviation of the $\bar{x}$ distribution? Is it necessary to make any assumption about the x distribution? Explain.

(b) What is the probability that $\bar{x}$ is between \$18 and \$22?

(c) Let us assume that x has a distribution that is approximately normal. What is the probability that x is between \$18 and \$22?

(d) In part (b), we used $\bar{x}$, the *average* amount spent, computed for 100 customers. In part (c), we used x, the amount spent by only *one* individual customer. The answers to parts (b) and (c) are very different. Why would this happen? In this example, $\bar{x}$ is a much more predictable or reliable statistic than x. Consider that almost all marketing strategies and sales pitches are designed for the *average* customer and *not* the *individual* customer. How does the central limit theorem tell us that the average customer is much more predictable than the individual customer? (See Problem 16 of Section 7.2.)

SECTION 7.1 Sampling Distributions

FOCUS POINTS

- Review such commonly used terms as *random sample, relative frequency, parameter, statistic, and sampling distribution.*
- From raw data, construct a relative frequency distribution for $\bar{x}$ values and compare the result to a theoretical sampling distribution.

Let us begin with some common statistical terms. Most of these have been discussed before, but this is a good time to review them.

From a statistical point of view, a *population* can be thought of as a set of measurements (or counts), either existing or conceptual. We discussed populations at some length in Chapter 1. A *sample* is a subset of measurements from the population. For our purposes, the most important samples are *random samples,* which were discussed in Section 1.2.

When we compute a descriptive measure such as an average, it makes a difference whether it was computed from a population or from a sample.

Statistic
Parameter

A **statistic** is a numerical descriptive measure of a *sample.*
A **parameter** is a numerical descriptive measure of a *population.*

For students wanting a memory device, point out that *population* and *parameter* both begin with the letter *p,* while *sample* and *statistic* both begin with the letter *s.*

It is important to notice that for a given population, a specified parameter is a fixed quantity. On the other hand, the value of a statistic might vary depending on which sample has been selected.

This section emphasizes some important statistical terminology. Students should realize we are "shifting gears" from a raw data distribution to a sampling distribution.

Some commonly used statistics and corresponding parameters

Measure	*Statistic*	*Parameter*
Mean	$\bar{x}$ (x bar)	μ (mu)
Variance	s^2	σ^2 (sigma squared)
Standard deviation	s	σ (sigma)
Proportion	$\hat{p}$ (p hat)	p

Often we do not have access to all the measurements of an entire population because of constraints on time, money, or effort. So, we must use measurements from a sample instead. In such cases, we will use a statistic (such as $\bar{x}$, s, or $\hat{p}$) to make *inferences* about a corresponding population parameter (e.g., μ, σ, or p). The principal types of inferences we will make are the following.

Types of inferences

1. **Estimation:** In this type of inference, we estimate the *value* of a population parameter.
2. **Testing:** In this type of inference, we formulate a *decision* about the value of a population parameter.
3. **Regression:** In this type of inference, we make *predictions* or *forecasts* about the value of a statistical variable.

Sampling distribution

To evaluate the reliability of our inferences, we will need to know the probability distribution for the statistic we are using. Such a probability distribution is called a *sampling distribution*. Perhaps Example 1 below will help clarify this discussion.

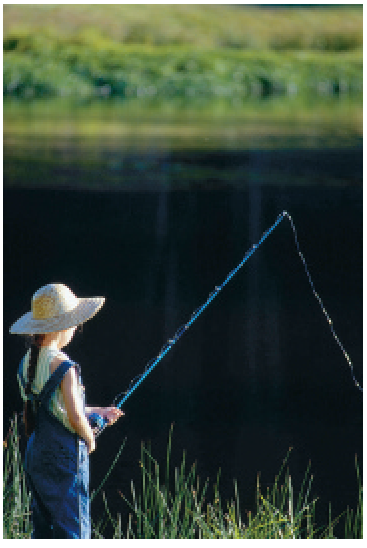
A **sampling distribution** is a probability distribution of a sample statistic based on all possible simple random samples of the *same* size from the same population.

EXAMPLE 1 Sampling distribution for $\bar{x}$

Pinedale, Wisconsin, is a rural community with a children's fishing pond. Posted rules state that all fish under 6 inches must be returned to the pond, only children under 12 years old may fish, and a limit of five fish may be kept per day. Susan is a college student who was hired by the community last summer to make sure the rules were obeyed and to see that the children were safe from accidents. The pond contains only rainbow trout and has been well stocked for many years. Each child has no difficulty catching his or her limit of five trout.

As a project for her biometrics class, Susan kept a record of the lengths (to the nearest inch) of all trout caught last summer. Hundreds of children visited the pond and caught their limit of five trout, so Susan has a lot of data. To make Table 7-1, Susan selected 100 children at random and listed the lengths of each of the five trout caught by a child in the sample. Then, for each child, she listed the mean length of the five trout that child caught.

Now let us turn our attention to the following question: What is the average (mean) length of a trout taken from the Pinedale children's pond last summer?



SOLUTION: We can get an idea of the average length by looking at the far-right column of Table 7-1. But just looking at 100 of the $\bar{x}$ values doesn't tell us much. Let's organize our $\bar{x}$ values into a frequency table. We used a class width of 0.38 to make Table 7-2.

Note: Techniques of Section 2.1 dictate a class width of 0.4. However, this choice results in the tenth class being beyond the data. Consequently, we shortened the class width slightly and also started the first class with a value slightly smaller than the smallest data value.

The far-right column of Table 7-2 contains relative frequencies $f/100$. Recall that the relative frequencies may be thought of as probabilities, so we effectively have a probability distribution. Because $\bar{x}$ represents the mean length of a trout

TABLE 7-1 **Length Measurements of Trout Caught by a Random Sample of 100 Children at the Pinedale Children's Pond**

Sample	Length (to nearest inch)					$\bar{x}$ = Sample Mean	Sample	Length (to nearest inch)					$\bar{x}$ = Sample Mean
1	11	10	10	12	11	10.8	51	9	10	12	10	9	10.0
2	11	11	9	9	9	9.8	52	7	11	10	11	10	9.8
3	12	9	10	11	10	10.4	53	9	11	9	11	12	10.4
4	11	10	13	11	8	10.6	54	12	9	8	10	11	10.0
5	10	10	13	11	12	11.2	55	8	11	10	9	10	9.6
6	12	7	10	9	11	9.8	56	10	10	9	9	13	10.2
7	7	10	13	10	10	10.0	57	9	8	10	10	12	9.8
8	10	9	9	9	10	9.4	58	10	11	9	8	9	9.4
9	10	10	11	12	8	10.2	59	10	8	9	10	12	9.8
10	10	11	10	7	9	9.4	60	11	9	9	11	11	10.2
11	12	11	11	11	13	11.6	61	11	10	11	10	11	10.6
12	10	11	10	12	13	11.2	62	12	10	10	9	11	10.4
13	11	10	10	9	11	10.2	63	10	10	9	11	7	9.4
14	10	10	13	8	11	10.4	64	11	11	12	10	11	11.0
15	9	11	9	10	10	9.8	65	10	10	11	10	9	10.0
16	13	9	11	12	10	11.0	66	8	9	10	11	11	9.8
17	8	9	7	10	11	9.0	67	9	11	11	9	8	9.6
18	12	12	8	12	12	11.2	68	10	9	10	9	11	9.8
19	10	8	9	10	10	9.4	69	9	9	11	11	11	10.2
20	10	11	10	10	10	10.2	70	13	11	11	9	11	11.0
21	11	10	11	9	12	10.6	71	12	10	8	8	9	9.4
22	9	12	9	10	9	9.8	72	13	7	12	9	10	10.2
23	8	11	10	11	10	10.0	73	9	10	9	8	9	9.0
24	9	12	10	9	11	10.2	74	11	11	10	9	10	10.2
25	9	9	8	9	10	9.0	75	9	11	14	9	11	10.8
26	11	11	12	11	11	11.2	76	14	10	11	12	12	11.8
27	10	10	10	11	13	10.8	77	8	12	10	10	9	9.8
28	8	7	9	10	8	8.4	78	8	10	13	9	8	9.6
29	11	11	8	10	11	10.2	79	11	11	11	13	10	11.2
30	8	11	11	9	12	10.2	80	12	10	11	12	9	10.8
31	11	9	12	10	10	10.4	81	10	9	10	10	13	10.4
32	10	11	10	11	12	10.8	82	11	10	9	9	12	10.2
33	12	11	8	8	11	10.0	83	11	11	10	10	10	10.4
34	8	10	10	9	10	9.4	84	11	10	11	9	9	10.0
35	10	10	10	10	11	10.2	85	10	11	10	9	7	9.4
36	10	8	10	11	13	10.4	86	7	11	10	9	11	9.6
37	11	10	11	11	10	10.6	87	10	11	10	10	10	10.2
38	7	13	9	12	11	10.4	88	9	8	11	10	12	10.0
39	11	11	8	11	11	10.4	89	14	9	12	10	9	10.8
40	11	10	11	12	9	10.6	90	9	12	9	10	10	10.0
41	11	10	9	11	12	10.6	91	10	10	8	6	11	9.0
42	11	13	10	12	9	11.0	92	8	9	11	9	10	9.4
43	10	9	11	10	11	10.2	93	8	10	9	9	11	9.4
44	10	9	11	10	9	9.8	94	12	11	12	13	10	11.6
45	12	11	9	11	12	11.0	95	11	11	9	9	9	9.8
46	13	9	11	8	8	9.8	96	8	12	8	11	10	9.8
47	10	11	11	11	10	10.6	97	13	11	11	12	8	11.0
48	9	9	10	11	11	10.0	98	10	11	8	10	11	10.0
49	10	9	9	10	10	9.6	99	13	10	7	11	9	10.0
50	10	10	6	9	10	9.0	100	9	9	10	12	12	10.4

TABLE 7-2 **Frequency Table for 100 Values of $\bar{x}$**

	Class Limits			
Class	Lower	Upper	f = Frequency	$f/100$ = Relative Frequency
1	8.39	8.76	1	0.01
2	8.77	9.14	5	0.05
3	9.15	9.52	10	0.10
4	9.53	9.90	19	0.19
5	9.91	10.28	27	0.27
6	10.29	10.66	18	0.18
7	10.67	11.04	12	0.12
8	11.05	11.42	5	0.05
9	11.43	11.80	3	0.03

(based on samples of five trout caught by each child), we estimate the probability of $\bar{x}$ falling into each class by using the relative frequencies. Figure 7-1 is a relative-frequency or probability distribution of the $\bar{x}$ values.

The bars of Figure 7-1 represent our estimated probabilities of $\bar{x}$ values based on the data of Table 7-1. The bell-shaped curve represents the theoretical probability distribution that would be obtained if the number of children (i.e., number of $\bar{x}$ values) were much larger.

FIGURE 7-1

Estimates of Probabilities of $\bar{x}$ Values

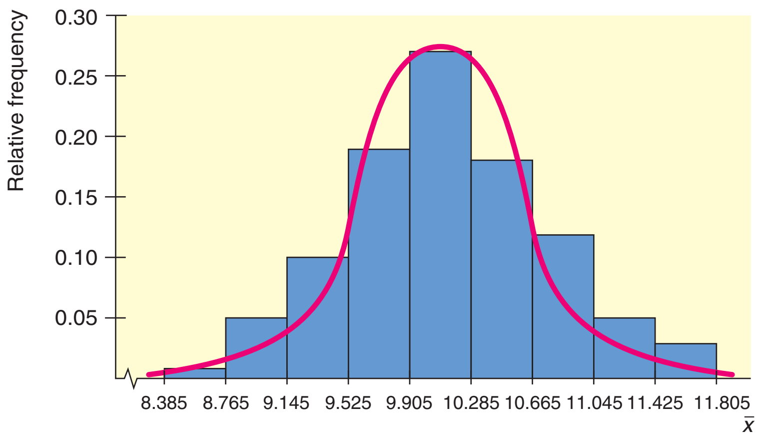


Figure 7-1 represents a *probability sampling distribution* for the sample mean $\bar{x}$ of trout lengths based on random samples of size 5. We see that the distribution is mound-shaped and even somewhat bell-shaped. Irregularities are due to the small number of samples used (only 100 sample means) and the rather small sample size (five trout per child). These irregularities would become less obvious and even disappear if the sample of children became much larger, if we used a larger number of classes in Figure 7-1, and if the number of trout in each sample became larger. In fact, the curve would eventually become a perfect bell-shaped curve. We will discuss this property at some length in the next section, which introduces the *central limit theorem*.

It is useful to remind students that in a sampling distribution, all samples must be of the *same size*.

There are other sampling distributions besides the $\bar{x}$ distribution. Section 7.3 shows the sampling distribution for $\hat{p}$. In the chapters ahead, we will see that other statistics have different sampling distributions. However, the $\bar{x}$ sampling distribution is very important. It will serve us well in our inferential work in Chapters 8 and 9 on estimation and testing.

Let us summarize the information about sampling distributions in the following exercise.

GUIDED EXERCISE 1 Terminology

(a) What is a population parameter? Give an example.

→ A population parameter is a numerical descriptive measure of a population. Examples are μ, σ, and p. (There are many others.)

(b) What is a sample statistic? Give an example.

→ A sample statistic or statistic is a numerical descriptive measure of a sample. Examples are $\bar{x}$, s, and $\hat{p}$.

(c) What is a sampling distribution?

→ A sampling distribution is a probability distribution for the sample statistic we are using.

(d) In Table 7-1, what makes up the members of the sample? What is the sample statistic corresponding to each sample? What is the sampling distribution? To which population parameter does this sampling distribution correspond?

→ There are 100 samples, each of which comprises five trout lengths. In the first sample, the five trout have lengths 11, 10, 10, 12, and 11. The sample statistic is the sample mean $\bar{x} = 10.8$. The sampling distribution is shown in Figure 7-1. This sampling distribution relates to the population mean μ of all lengths of trout taken from the Pinedale children's pond (i.e., trout over 6 inches long).

(e) Where will sampling distributions be used in our study of statistics?

→ Sampling distributions will be used for statistical inference. (Chapter 8 will concentrate on a method of inference called *estimation*. Chapter 9 will concentrate on a method of inference called *testing*.)

VIEWPOINT "Chance Favors the Prepared Mind"

—Louis Pasteur

It also has been said that a discovery is nothing more than an accident that meets a prepared mind. Sampling can be one of the best forms of preparation. In fact, sampling may be the primary way we humans venture into the unknown. Probability sampling distributions can provide new information for the sociologist, scientist, or economist. In addition, ordinary human sampling of life can help writers and artists develop preferences, style, and insight. Ansel Adams became famous for photographing lyrical, unforgettable landscapes such as "Moonrise, Hernandez, New Mexico." Adams claimed that he was a strong believer in the quote by Pasteur. In fact, he claims that the Hernandez photograph was just such a favored chance happening that his prepared mind readily grasped. During his lifetime, Adams made over $25 million from sales and royalties on the Hernandez photograph.

SECTION 7.1 PROBLEMS

These problems provide good topics for an in-class discussion of basic terminology.

1. Answers vary. Remind students to identify the individuals (subjects) and variable involved.
2. See Section 1.2.

This is a good time to review several important concepts, some of which we have studied earlier. Please write out a careful but brief answer to each of the following questions.

1. *Statistical Literacy* What is a population? Give three examples.
2. *Statistical Literacy* What is a random sample from a population? (*Hint:* See Section 1.2.)

3. A numerical descriptive measure of a population. Examples include μ, σ^2, σ, p, and ρ (rho) for those who have already studied linear regression from Chapter 10.
4. A numerical descriptive measure of a sample. Examples: $\bar{x}$, s, s^2, $\hat{p}$, and so forth.
5. A statistical inference is a conclusion about the value of a population parameter based on information about the corresponding sample statistic and probability. We will do both estimation and testing.
6. A probability distribution for a sample statistic.
7. They help us visualize the sampling distribution through tables and graphs that approximately represent the sampling distribution.
8. A relative frequency can be thought of as a measure or estimate of the likelihood of a certain statistic falling within the class bounds.
9. We studied the sampling distribution of mean trout lengths based on samples of size 5.

3. *Statistical Literacy* What is a population parameter? Give three examples.

4. *Statistical Literacy* What is a sample statistic? Give three examples.

5. *Statistical Literacy* What is the meaning of the term *statistical inference*? What types of inferences will we make about population parameters?

6. *Statistical Literacy* What is a sampling distribution?

7. *Critical Thinking* How do frequency tables, relative frequencies, and histograms showing relative frequencies help us understand sampling distributions?

8. *Critical Thinking* How can relative frequencies be used to help us estimate probabilities occurring in sampling distributions?

9. *Critical Thinking* Give an example of a specific sampling distribution we studied in this section. Outline other possible examples of sampling distributions from areas such as business administration, economics, finance, psychology, political science, sociology, biology, medical science, sports, engineering, chemistry, linguistics, and so on.

SECTION 7.2

The Central Limit Theorem

FOCUS POINTS

- For a normal distribution, use μ and σ to construct the theoretical sampling distribution for the statistic $\bar{x}$.
- For large samples, use sample estimates to construct a good approximate sampling distribution for the statistic $\bar{x}$.
- Learn the statement and underlying meaning of the central limit theorem well enough to explain it to a friend who is intelligent, but (unfortunately) doesn't know much about statistics.

The $\bar{x}$ Distribution, Given x Is Normal

In Section 7.1, we began a study of the distribution of $\bar{x}$ values, where $\bar{x}$ was the (sample) mean length of five trout caught by children at the Pinedale children's fishing pond. Let's consider this example again in the light of a very important theorem of mathematical statistics.

THEOREM 7.1 **For a Normal Probability Distribution** Let x be a random variable with a *normal distribution* whose mean is μ and whose standard deviation is σ. Let $\bar{x}$ be the sample mean corresponding to random samples of size n taken from the x distribution. Then the following are true:

(a) The $\bar{x}$ distribution is a *normal distribution*.

(b) The mean of the $\bar{x}$ distribution is μ.

(c) The standard deviation of the $\bar{x}$ distribution is $\sigma/\sqrt{n}$.

Linking Concepts, Problem 1, provides material for a class discussion of Theorem 7.1 and a discussion of why sampling distributions are important in statistical work.

We conclude from Theorem 7.1 that when x has a normal distribution, the $\bar{x}$ distribution will be normal *for any sample size n*. Furthermore, we can convert the $\bar{x}$ distribution to the standard normal z distribution using the following formulas.

Some students may need to be reminded to use parentheses in the numerator and denominator when they compute z on their calculators.

Some students might prefer to use the equivalent formula

$$z = \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma}$$

$$\mu_{\bar{x}} = \mu$$

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

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where n is the sample size,
μ is the mean of the $\bar{x}$ distribution, and
σ is the standard deviation of the x distribution.

Theorem 7.1 is a wonderful theorem! It states that the $\bar{x}$ distribution will be normal provided the x distribution is normal. The sample size n could be 2, 3, 4, or any (fixed) sample size we wish. Furthermore, the mean of the $\bar{x}$ distribution is μ (same as for the x distribution), but the standard deviation is $\sigma/\sqrt{n}$ (which is, of course, smaller than σ). The next example illustrates Theorem 7.1.

EXAMPLE 2 PROBABILITY REGARDING x AND $\bar{x}$

Suppose a team of biologists has been studying the Pinedale children's fishing pond. Let x represent the length of a single trout taken at random from the pond. This group of biologists has determined that x has a normal distribution with mean $\mu = 10.2$ inches and standard deviation $\sigma = 1.4$ inches.

(a) What is the probability that a *single trout* taken at random from the pond is between 8 and 12 inches long?

SOLUTION: We use the methods of Chapter 6, with $\mu = 10.2$ and $\sigma = 1.4$, to get

$$z = \frac{x - \mu}{\sigma} = \frac{x - 10.2}{1.4}$$

Therefore,

$$\begin{aligned} P(8 < x < 12) &= P\left(\frac{8 - 10.2}{1.4} < z < \frac{12 - 10.2}{1.4}\right) \\ &= P(-1.57 < z < 1.29) \\ &= 0.9015 - 0.0582 = 0.8433 \end{aligned}$$

Therefore, the probability is about 0.8433 that a *single* trout taken at random is between 8 and 12 inches long.

(b) What is the probability that the *mean length* $\bar{x}$ of five trout taken at random is between 8 and 12 inches?

SOLUTION: If we let $\mu_{\bar{x}}$ represent the mean of the distribution, then Theorem 7.1, part (b), tells us that

$$\mu_{\bar{x}} = \mu = 10.2$$

If $\sigma_{\bar{x}}$ represents the standard deviation of the $\bar{x}$ distribution, then Theorem 7.1, part (c), tells us that

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.4/\sqrt{5} \approx 0.63$$

To create a standard z variable from $\bar{x}$, we subtract $\mu_{\bar{x}}$ and divide by $\sigma_{\bar{x}}$:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 10.2}{0.63}$$

To standardize the interval $8 < \bar{x} < 12$, we use 8 and then 12 in place of $\bar{x}$ in the preceding formula for z.

$$8 < \bar{x} < 12$$
$$\frac{8 - 10.2}{0.63} < z < \frac{12 - 10.2}{0.63}$$
$$-3.49 < z < 2.86$$

Theorem 7.1, part (a), tells us that $\bar{x}$ has a normal distribution. Therefore,

$$P(8 < \bar{x} < 12) = P(-3.49 < z < 2.86) = 0.9979 - 0.0002 = 0.9977$$

The probability is about 0.9977 that the mean length based on a sample size of 5 is between 8 and 12 inches.

(c) Looking at the results of parts (a) and (b), we see that the probabilities (0.8433 and 0.9977) are quite different. Why is this the case?

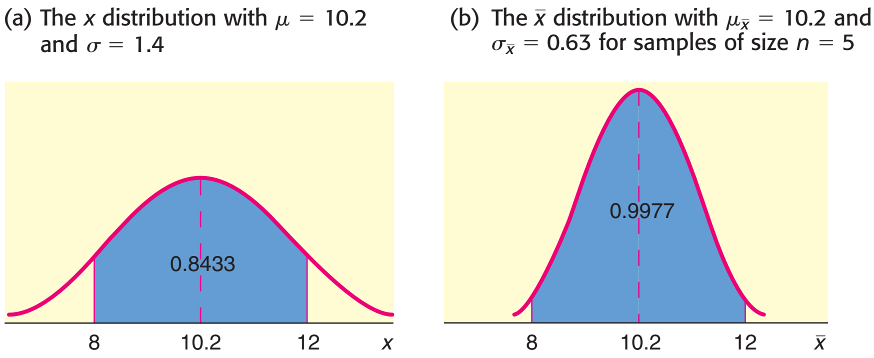
SOLUTION: According to Theorem 7.1, both x and $\bar{x}$ have a normal distribution, and both have the same mean of 10.2 inches. The difference is in the standard deviations for x and $\bar{x}$. The standard deviation of the x distribution is $\sigma = 1.4$. The standard deviation of the $\bar{x}$ distribution is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.4/\sqrt{5} \approx 0.63$$

The standard deviation of the $\bar{x}$ distribution is less than half the standard deviation of the x distribution. Figure 7-2 shows the distributions of x and $\bar{x}$.

FIGURE 7-2

General Shapes of the x and $\bar{x}$ Distributions



Looking at Figure 7-2(a) and (b), we see that both curves use the same scale on the horizontal axis. The means are the same, and the shaded area is above the interval from 8 to 12 on each graph. It becomes clear that the smaller standard deviation of the $\bar{x}$ distribution has the effect of gathering together much more of the total probability into the region over its mean. Therefore, the region from 8 to 12 has a much higher probability for the $\bar{x}$ distribution.

Standard error of the mean

Theorem 7.1 describes the distribution of a particular statistic: namely, the distribution of sample mean $\bar{x}$. The standard deviation of a statistic is referred to as the *standard error* of that statistic.

It is good to point out that the term *standard error* is very widely used in statistical literature.

The **standard error** is the standard deviation of a sampling distribution. For the $\bar{x}$ sampling distribution,

$$\textbf{standard error} = \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

Statistical software

The expression *standard error* appears commonly on printouts and refers to the standard deviation of the sampling distribution being used. (In Minitab, the expression SE MEAN refers to the standard error of the mean.)

The $\bar{x}$ Distribution, Given x Follows Any Distribution

Central limit theorem

Theorem 7.1 gives complete information about the $\bar{x}$ distribution, provided the original x distribution is known to be normal. What happens if we don't have information about the shape of the original x distribution? The *central limit theorem* tells us what to expect.

The central limit theorem is a generalization of the DeMoivre-Laplace theorem.

THEOREM 7.2 The Central Limit Theorem for Any Probability Distribution If x possesses *any* distribution with mean μ and standard deviation σ, then the sample mean $\bar{x}$ based on a random sample of size n will have a distribution that approaches the distribution of a normal random variable with mean μ and standard deviation $\sigma/\sqrt{n}$ as n increases without limit.

Using Technology has students use a random-number table as a tool for demonstrating the central limit theorem.

The central limit theorem is indeed surprising! It says that x can have *any* distribution whatsoever, but as the sample size gets larger and larger, the distribution of $\bar{x}$ will approach a *normal* distribution. From this relation, we begin to appreciate the scope and significance of the normal distribution.

Large sample

In the central limit theorem, the degree to which the distribution of $\bar{x}$ values fits a normal distribution depends on both the selected value of n and the original distribution of x values. A natural question is: How large should the sample size be if we want to apply the central limit theorem? After a great deal of theoretical as well as empirical study, statisticians agree that if n is 30 or larger, the $\bar{x}$ distribution will appear to be normal and the central limit theorem will apply. However, this rule should not be applied blindly. If the x distribution is definitely not symmetrical about its mean, then the $\bar{x}$ distribution also will display a lack of symmetry. In such a case, a sample size larger than 30 may be required to get a reasonable approximation to the normal.

In practice, it is a good idea, when possible, to make a histogram of sample x values. If the histogram is approximately mound-shaped, and if it is more or less symmetrical, then we may be assured that, for all practical purposes, the $\bar{x}$ distribution will be well approximated by a normal distribution and the central limit theorem will apply when the sample size is 30 or larger. The main thing to remember is that in almost all practical applications, a sample size of 30 or more is adequate for the central limit theorem to hold. However, in a few rare applications, you may need a sample size larger than 30 to get reliable results.

Let's summarize this information for convenient reference: For almost all x distributions, if we use a random sample of size 30 or larger, the $\bar{x}$ distribution will be approximately normal. The larger the sample size becomes, the closer the $\bar{x}$ distribution gets to the normal. Furthermore, we may convert the $\bar{x}$ distribution to a standard normal distribution using the following formulas.

Data Highlights (*Iris setosa*) and Linking Concepts, Problem 2, are good topics for a class discussion about the central limit theorem. Both features provide good material for an in-class demonstration.

Using the central limit theorem to convert the $\bar{x}$ distribution to the standard normal distribution

$$\mu_{\bar{x}} = \mu$$

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

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where n is the sample size ($n \geq 30$),
μ is the mean of the x distribution, and
σ is the standard deviation of the x distribution.

Guided Exercise 2 shows how to standardize $\bar{x}$ when appropriate. Then, Example 3 demonstrates the use of the central limit theorem in a decision-making process.

GUIDED EXERCISE 2 Central limit theorem

(a) Suppose x has a *normal* distribution with mean $\mu = 18$ and standard deviation $\sigma = 3$. If you draw random samples of size 5 from the x distribution and $\bar{x}$ represents the sample mean, what can you say about the $\bar{x}$ distribution? How could you standardize the $\bar{x}$ distribution?

Since the x distribution is given to be *normal,* the $\bar{x}$ distribution also will be normal even though the sample size is much less than 30. The mean is $\mu_{\bar{x}} = \mu = 18$. The standard deviation is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3/\sqrt{5} \approx 1.3$$

We could standardize $\bar{x}$ as follows:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 18}{1.3}$$

(b) Suppose you know that the x distribution has mean $\mu = 75$ and standard deviation $\sigma = 12$, but you have no information as to whether or not the x distribution is normal. If you draw samples of size 30 from the x distribution and $\bar{x}$ represents the sample mean, what can you say about the $\bar{x}$ distribution? How could you standardize the $\bar{x}$ distribution?

Since the sample size is large enough, the $\bar{x}$ distribution will be an approximately normal distribution. The mean of the $\bar{x}$ distribution is

$$\mu_{\bar{x}} = \mu = 75$$

The standard deviation of the $\bar{x}$ distribution is

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 12/\sqrt{30} \approx 2.2$$

We could standardize $\bar{x}$ as follows:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 75}{2.2}$$

(c) Suppose you did not know that x had a normal distribution. Would you be justified in saying that the $\bar{x}$ distribution is approximately normal if the sample size were $n = 8$?

No, the sample size should be 30 or larger if we don't know that x has a normal distribution.

EXAMPLE 3 CENTRAL LIMIT THEOREM

A certain strain of bacteria occurs in all raw milk. Let x be the bacteria count per milliliter of milk. The health department has found that if the milk is not contaminated, then x has a distribution that is more or less mound-shaped and symmetrical. The mean of the x distribution is $\mu = 2500$, and the standard deviation is $\sigma = 300$. In a large commercial dairy, the health inspector takes 42 random samples of the milk produced each day. At the end of the day, the bacteria count in each of the 42 samples is averaged to obtain the sample mean bacteria count $\bar{x}$.

(a) Assuming the milk is not contaminated, what is the distribution of $\bar{x}$?

SOLUTION: The sample size is $n = 42$. Since this value exceeds 30, the central limit theorem applies, and we know that $\bar{x}$ will be approximately normal with mean and standard deviation

$$\mu_{\bar{x}} = \mu = 2500$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 300/\sqrt{42} \approx 46.3$$

(b) Assuming the milk is not contaminated, what is the probability that the average bacteria count $\bar{x}$ for one day is between 2350 and 2650 bacteria per milliliter?

SOLUTION: We convert the interval

$$2350 \le \bar{x} \le 2650$$

to a corresponding interval on the standard z axis.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 2500}{46.3}$$

$$\bar{x} = 2350 \quad \text{converts to} \quad z = \frac{2350 - 2500}{46.3} \approx -3.24$$

$$\bar{x} = 2650 \quad \text{converts to} \quad z = \frac{2650 - 2500}{46.3} \approx 3.24$$

Therefore,

$$\begin{aligned} P(2350 \le \bar{x} \le 2650) &= P(-3.24 \le z \le 3.24) \\ &= 0.9994 - 0.0006 \\ &= 0.9988 \end{aligned}$$

The probability is 0.9988 that $\bar{x}$ is between 2350 and 2650.

(c) **INTERPRETATION** At the end of each day, the inspector must decide to accept or reject the accumulated milk that has been held in cold storage awaiting shipment. Suppose the 42 samples taken by the inspector have a mean bacteria count $\bar{x}$ that is *not* between 2350 and 2650. If you were the inspector, what would be your comment on this situation?

SOLUTION: The probability that $\bar{x}$ is between 2350 and 2650 is very high. If the inspector finds that the average bacteria count for the 42 samples is not between 2350 and 2650, then it is reasonable to conclude that there is something wrong with the milk. If $\bar{x}$ is less than 2350, you might suspect someone added chemicals to the milk to artificially reduce the bacteria count. If $\bar{x}$ is above 2650, you might suspect some other kind of biologic contamination.

PROCEDURE

HOW TO FIND PROBABILITIES REGARDING $\bar{x}$

Given a probability distribution of x values where

n = sample size

μ = mean of the x distribution

σ = standard deviation of the x distribution

1. If the x distribution is *normal*, then the $\bar{x}$ distribution is *normal*.
2. Even if the x distribution is *not* normal, if the *sample size* $n \ge 30$, then, by the central limit theorem, the $\bar{x}$ distribution is *approximately normal*.
3. Convert $\bar{x}$ to z using the formula

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

4. Use the standard normal distribution to find the corresponding probabilities of events regarding $\bar{x}$.

GUIDED EXERCISE 3 Probability regarding $\bar{x}$

In mountain country, major highways sometimes use tunnels instead of long, winding roads over high passes. However, too many vehicles in a tunnel at the same time can cause a hazardous situation. Traffic engineers are studying a long tunnel in Colorado. If x represents the time for a vehicle to go through the tunnel, it is known that the x distribution has mean $\mu = 12.1$ minutes and standard deviation $\sigma = 3.8$ minutes under ordinary traffic conditions. From a histogram of x values, it was found that the x distribution is mound-shaped with some symmetry about the mean.

Engineers have calculated that, *on average,* vehicles should spend from 11 to 13 minutes in the tunnel. If the time is less than 11 minutes, traffic is moving too fast for safe travel in the tunnel. If the time is more than 13 minutes, there is a problem of bad air quality (too much carbon monoxide and other pollutants).

Under ordinary conditions, there are about 50 vehicles in the tunnel at one time. What is the probability that the mean time for 50 vehicles in the tunnel will be from 11 to 13 minutes?

We will answer this question in steps.

(a) Let $\bar{x}$ represent the sample mean based on samples of size 50. Describe the $\bar{x}$ distribution.

From the central limit theorem, we expect the $\bar{x}$ distribution to be approximately normal with mean and standard deviation

$$\mu_{\bar{x}} = \mu = 12.1 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.8}{\sqrt{50}} \approx 0.54$$

(b) Find $P(11 < \bar{x} < 13)$.

We convert the interval

$11 < \bar{x} < 13$

to a standard z interval and use the standard normal probability table to find our answer. Since

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 12.1}{0.54}$$

$\bar{x} = 11$ converts to $z \approx \dfrac{11 - 12.1}{0.54} = -2.04$

and $\bar{x} = 13$ converts to $z \approx \dfrac{13 - 12.1}{0.54} = 1.67$

Therefore,

$$\begin{aligned} P(11 < \bar{x} < 13) &= P(-2.04 < z < 1.67) \\ &= 0.9525 - 0.0207 \\ &= 0.9318 \end{aligned}$$

(c) Interpret your answer to part (b).

It seems that about 93% of the time there should be no safety hazard for average traffic flow.

CRITICAL THINKING

Bias and Variability

Whenever we use a sample statistic as an estimate of a population parameter, we need to consider both *bias* and *variability* of the statistic.

A sample statistic is **unbiased** if the mean of its sampling distribution equals the value of the parameter being estimated.

The spread of the sampling distribution indicates the **variability of the statistic.** The spread is affected by the sampling method and the sample size. Statistics from larger random samples have spreads that are smaller.

We see from the central limit theorem that the sample mean $\bar{x}$ is an unbiased estimator of the mean μ when $n \geq 30$. The variability of $\bar{x}$ decreases as the sample size increases.

In Section 7.3, we will see that the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion of successes p in binomial experiments with sufficiently large numbers of trials n. Again, we will see that the variability of $\hat{p}$ decreases with increasing numbers of trials.

The sample variance s^2 is an unbiased estimator for the population variance σ^2.

VIEWPOINT

Chaos!

Is there a different side to random sampling? Can sampling be used as a weapon? According to The Wall Street Journal, *the answer could be yes! The acronym for* **C**reate **H**avoc **A**round **O**ur **S**ystem *is* **CHAOS.** *The Association of Flight Attendants (AFA) is a union that successfully used CHAOS against Alaska Airlines in 1994 as a negotiation tool.* ***CHAOS*** *involves a small sample of random strikes—a few flights at a time—instead of a mass walkout. The president of the AFA claims that by striking randomly, "we take control of the schedule." The entire schedule becomes unreliable, and that is something management cannot tolerate. In 1986, TWA flight attendants struck in a mass walkout, and all were permanently replaced! Using* ***CHAOS****, only a few jobs are put at risk, and these are usually not lost. It appears that random sampling can be used as a weapon.*

SECTION 7.2 PROBLEMS

Note: Answers may differ slightly depending on how many digits are carried in the standard deviation.

1. The standard deviation.
2. The standard error.
3. $\bar{x}$ is an unbiased estimator for μ; $\hat{p}$ is an unbiased estimator for p.
4. As the sample size increases, the variability decreases.
5. (a) 30 or more.
   (b) No.

In these problems, the word *average* refers to the arithmetic mean $\bar{x}$ or μ, as appropriate.

1. *Statistical Literacy* What is the standard error of a sampling distribution?
2. *Statistical Literacy* What is the standard deviation of a sampling distribution called?
3. *Statistical Literacy* List two unbiased estimators and their corresponding parameters.
4. *Statistical Literacy* Describe how the variability of the $\bar{x}$ distribution changes as the sample size increases.
5. *Statistical Literacy*
   (a) If we have a distribution of x values that is more or less mound-shaped and somewhat symmetrical, what is the sample size needed to claim that the distribution of sample means $\bar{x}$ from random samples of that size is approximately normal?

(b) If the original distribution of x values is known to be normal, do we need to make any restriction about sample size in order to claim that the distribution of sample means $\bar{x}$ taken from random samples of a given size is normal?

6. (a) No, sample size is too small.
(b) Normal with mean 72 and standard deviation 2; 0.6687.

7. The second. The standard error of the first is $\sigma/10$, while that of the second is $\sigma/15$, where σ is the standard deviation of the original x distribution.

8. (a) $n = 36$, since $\sigma/\sqrt{n} = 12/\sqrt{36} = 2$.
(b) $n = 144$, since $\sigma/\sqrt{n} = 12/\sqrt{144} = 1$.

9. (a) $\mu_{\bar{x}} = 15$; $\sigma_{\bar{x}} = 2.0$; 0.3413.
(b) $\mu_{\bar{x}} = 15$; $\sigma_{\bar{x}} = 1.75$; 0.3729.
(c) The standard deviation of part (b) is smaller, resulting in a narrower distribution.

10. (a) $\mu_{\bar{x}} = 5$.
(b) Distribution with $n = 81$, since the standard deviation is smaller, resulting in a distribution that is less spread out about the mean.
(c) Distribution with $n = 81$, since the standard deviation is smaller, resulting in a distribution that is less spread out about the mean.

11. (a) 0.2643.
(b) 0.0026.
(c) No; yes.

12. (a) 0.2586.
(b) 0.6826.
(c) Yes; the standard deviation is smaller for the $\bar{x}$ distribution.

13. (a) 0.0359.
(b) 0.0054.
(c) 0.0009.
(d) Less than 0.0002.
(e) Yes.

6. *Critical Thinking* Suppose x has a distribution with $\mu = 72$ and $\sigma = 8$.
(a) If random samples of size $n = 16$ are selected, can we say anything about the $\bar{x}$ distribution of sample means?
(b) If the original x distribution is *normal*, can we say anything about the $\bar{x}$ distribution of random samples of size 16? Find $P(68 \le \bar{x} \le 73)$.

7. *Critical Thinking* Consider two $\bar{x}$ distributions corresponding to the same x distribution. The first $\bar{x}$ distribution is based on samples of size $n = 100$ and the second is based on samples of size $n = 225$. Which $\bar{x}$ distribution has the smaller standard error? Explain.

8. *Critical Thinking* Consider an x distribution with standard deviation $\sigma = 12$.
(a) If specifications for a research project require the standard error of the corresponding $\bar{x}$ distribution to be 2, how large does the sample size need to be?
(b) If specifications for a research project require the standard error of the corresponding $\bar{x}$ distribution to be 1, how large does the sample size need to be?

9. *Critical Thinking* Suppose x has a distribution with $\mu = 15$ and $\sigma = 14$.
(a) If a random sample of size $n = 49$ is drawn, find $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$, and $P(15 \le \bar{x} \le 17)$.
(b) If a random sample of size $n = 64$ is drawn, find $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$, and $P(15 \le \bar{x} \le 17)$.
(c) Why should you expect the probability of part (b) to be higher than that of part (a)? (*Hint:* Consider the standard deviations in parts (a) and (b).)

10. *Critical Thinking* Suppose an x distribution has mean $\mu = 5$. Consider two corresponding $\bar{x}$ distributions, the first based on samples of size $n = 49$ and the second based on samples of size $n = 81$.
(a) What is the value of the mean of each of the two $\bar{x}$ distributions?
(b) For which $\bar{x}$ distribution is $P(\bar{x} > 6)$ smaller? Explain.
(c) For which $\bar{x}$ distribution is $P(4 < \bar{x} < 6)$ greater? Explain.

11. *Coal: Automatic Loader* Coal is carried from a mine in West Virginia to a power plant in New York in hopper cars on a long train. The automatic hopper car loader is set to put 75 tons of coal into each car. The actual weights of coal loaded into each car are *normally distributed,* with mean $\mu = 75$ tons and standard deviation $\sigma = 0.8$ ton.
(a) What is the probability that one car chosen at random will have less than 74.5 tons of coal?
(b) What is the probability that 20 cars chosen at random will have a mean load weight $\bar{x}$ of less than 74.5 tons of coal?
(c) *Interpretation:* Suppose the weight of coal in one car was less than 74.5 tons. Would that fact make you suspect that the loader had slipped out of adjustment? Suppose the weight of coal in 20 cars selected at random had an average $\bar{x}$ of less than 74.5 tons. Would that fact make you suspect that the loader had slipped out of adjustment? Why?

12. *Vital Statistics: Heights of Men* The heights of 18-year-old men are approximately *normally distributed,* with mean 68 inches and standard deviation 3 inches (based on information from *Statistical Abstract of the United States,* 112th Edition).
(a) What is the probability that an 18-year-old man selected at random is between 67 and 69 inches tall?
(b) If a random sample of nine 18-year-old men is selected, what is the probability that the mean height $\bar{x}$ is between 67 and 69 inches?
(c) *Interpretation:* Compare your answers to parts (a) and (b). Is the probability in part (b) much higher? Why would you expect this?

13. *Medical: Blood Glucose* Let x be a random variable that represents the level of glucose in the blood (milligrams per deciliter of blood) after a 12-hour fast.

Assume that for people under 50 years old, x has a distribution that is approximately normal, with mean $\mu = 85$ and estimated standard deviation $\sigma = 25$ (based on information from *Diagnostic Tests with Nursing Applications,* edited by S. Loeb, Springhouse). A test result $x < 40$ is an indication of severe excess insulin, and medication is usually prescribed.

(a) What is the probability that, on a single test, $x < 40$?

(b) Suppose a doctor uses the average $\bar{x}$ for two tests taken about a week apart. What can we say about the probability distribution of $\bar{x}$? *Hint:* See Theorem 7.1. What is the probability that $\bar{x} < 40$?

(c) Repeat part (b) for $n = 3$ tests taken a week apart.

(d) Repeat part (b) for $n = 5$ tests taken a week apart.

(e) ***Interpretation:*** Compare your answers to parts (a), (b), (c), and (d). Did the probabilities decrease as n increased? Explain what this might imply if you were a doctor or a nurse. If a patient had a test result of $\bar{x} < 40$ based on five tests, explain why either you are looking at an extremely rare event or (more likely) the person has a case of excess insulin.

14. (a) 0.0110.
(b) 0.0006.
(c) Less than 0.0002.
(d) The probabilities decreased as n increased. It would be an extremely rare event for a person to have two or three tests below 3500 purely by chance.

14. ***Medical: White Blood Cells*** Let x be a random variable that represents white blood cell count per cubic milliliter of whole blood. Assume that x has a distribution that is approximately normal, with mean $\mu = 7500$ and estimated standard deviation $\sigma = 1750$ (see reference in Problem 13). A test result of $x < 3500$ is an indication of leukopenia. This indicates bone marrow depression that may be the result of a viral infection.

(a) What is the probability that, on a single test, x is less than 3500?

(b) Suppose a doctor uses the average $\bar{x}$ for two tests taken about a week apart. What can we say about the probability distribution of $\bar{x}$? What is the probability of $\bar{x} < 3500$?

(c) Repeat part (b) for $n = 3$ tests taken a week apart.

(d) ***Interpretation:*** Compare your answers to parts (a), (b), and (c). How did the probabilities change as n increased? If a person had $\bar{x} < 3500$ based on three tests, what conclusion would you draw as a doctor or a nurse?

15. (a) 0.1020.
(b) 224.
(c) 0.0014.
(d) 0.8849. Unlikely.

15. ***Wildlife: Deer*** Let x be a random variable that represents the weights in kilograms (kg) of healthy adult female deer (does) in December in Mesa Verde National Park. Then x has a distribution that is approximately normal with mean $\mu = 63.0$ kg and standard deviation $\sigma = 7.1$ kg (Source: *The Mule Deer of Mesa Verde National Park,* by G. W. Mierau and J. L. Schmidt, Mesa Verde Museum Association). Suppose a doe that weighs less than 54 kg is considered undernourished.

(a) What is the probability that a single doe captured (weighed and released) at random in December is undernourished?

(b) If the park has about 2200 does, what number do you expect to be undernourished in December?

(c) ***Interpretation:*** To estimate the health of the December doe population, park rangers use the rule that the average weight of $n = 50$ does should be more than 60 kg. If the average weight is less than 60 kg, it is thought that the entire population of does might be undernourished. What is the probability that the average weight $\bar{x}$ for a random sample of 50 does is less than 60 kg (assume a healthy population)?

(d) ***Interpretation:*** Compute the probability that $\bar{x} < 64.2$ kg for 50 does (assume a healthy population). Suppose park rangers captured, weighed, and released 50 does in December, and the average weight was $\bar{x} = 64.2$ kg. Do you think the doe population is undernourished or not? Explain.

16. (a) Approximately normal, with mean \$20 and standard deviation \$0.70.
(b) 0.9958.
(c) 0.2282.
(d) Essay.

16. ***Focus Problem: Impulse Buying*** Let x represent the dollar amount spent on supermarket impulse buying in a 10-minute (unplanned) shopping interval. Based on a *Denver Post* article, the mean of the x distribution is about \$20 and the estimated standard deviation is about \$7.

(a) Consider a random sample of $n = 100$ customers, each of whom has 10 minutes of unplanned shopping time in a supermarket. From the central limit theorem, what can you say about the probability distribution of $\bar{x}$, the average amount spent by these customers due to impulse buying? What are the mean and standard deviation of the $\bar{x}$ distribution? Is it necessary to make any assumption about the x distribution? Explain.

(b) What is the probability that $\bar{x}$ is between \$18 and \$22?

(c) Let us assume that x has a distribution that is approximately normal. What is the probability that x is between \$18 and \$22?

(d) ***Interpretation:*** In part (b), we used $\bar{x}$, the *average* amount spent, computed for 100 customers. In part (c), we used x, the amount spent by only *one* customer. The answers to parts (b) and (c) are very different. Why would this happen? In this example, $\bar{x}$ is a much more predictable or reliable statistic than x. Consider that almost all marketing strategies and sales pitches are designed for the *average* customer and *not the individual* customer. How does the central limit theorem tell us that the average customer is much more predictable than the individual customer?

Problems 17 and 18 require a little extra attention to detail. Discussing one of the problems in class would be useful. In particular, point out that in part (a), the random variable x is itself an average based on the number of stocks or bonds in the fund. In part (b), the random variable is $\bar{x}$, an average based on the specified number of months or years. For part (b), point out that n = number of specified months or years.

17. (a) x is a mean of a sample of size $n = 250$. By the central limit theorem, the x distribution is approximately normal.
(b) 0.8105.
(c) 0.9849.
(d) Yes.
(e) 0.0005. Unlikely.

17. ***Finance: Templeton Funds*** Templeton World is a mutual fund that invests in both U.S. and foreign markets. Let x be a random variable that represents the monthly percentage return for the Templeton World fund. Based on information from the *Morningstar Guide to Mutual Funds* (available in most libraries), x has mean $\mu = 1.6\%$ and standard deviation $\sigma = 0.9\%$.

(a) Templeton World fund has over 250 stocks that combine together to give the overall monthly percentage return x. We can consider the monthly return of the stocks in the fund to be a sample from the population of monthly returns of all world stocks. Then we see that the overall monthly return x for Templeton World fund is itself an average return computed using all 250 stocks in the fund. Why would this indicate that x has an approximately normal distribution? Explain. *Hint:* See the discussion after Theorem 7.2.

(b) After 6 months, what is the probability that the *average* monthly percentage return $\bar{x}$ will be between 1% and 2%? *Hint:* See Theorem 7.1, and assume that x has a normal distribution as based on part (a).

(c) After 2 years, what is the probability that $\bar{x}$ will be between 1% and 2%?

(d) Compare your answers to parts (b) and (c). Did the probability increase as n (number of months) increased? Why would this happen?

(e) ***Interpretation:*** If after 2 years the average monthly percentage return $\bar{x}$ was less than 1%, would that tend to shake your confidence in the statement that $\mu = 1.6\%$? Might you suspect that μ has slipped below 1.6%? Explain.

18. (a) x is a mean of a sample of $n = 100$ stocks. By the central limit theorem, the x distribution is approximately normal.
(b) 0.9210.
(c) 0.9823.
(d) Yes, probability increases as the standard deviation decreases.
(e) 0.0007. Event is unlikely.

18. ***Finance: Dean Witter Funds*** Dean Witter European Growth is a mutual fund that specializes in stocks from the British Isles, continental Europe, and Scandinavia. The fund has over 100 stocks. Let x be a random variable that represents the monthly percentage return for this fund. Based on information from *Morningstar* (see Problem 17), x has mean $\mu = 1.4\%$ and standard deviation $\sigma = 0.8\%$.

(a) Let's consider the monthly return of the stocks in the Dean Witter fund to be a sample from the population of monthly returns of all European stocks. Is it reasonable to assume that x (the average monthly return on the 100 stocks in the Dean Witter European Growth fund) has a distribution that is approximately normal? Explain. *Hint:* See Problem 17, part (a).

(b) After 9 months, what is the probability that the *average* monthly percentage return $\bar{x}$ will be between 1% and 2%? *Hint:* See Theorem 7.1 and the results of part (a).

(c) After 18 months, what is the probability that the *average* monthly percentage return $\bar{x}$ will be between 1% and 2%?

(d) Compare your answers to parts (b) and (c). Did the probability increase as n (number of months) increased? Why would this happen?

(e) ***Interpretation:*** If after 18 months the average monthly percentage return $\bar{x}$ is more than 2%, would that tend to shake your confidence in the statement that $\mu = 1.4\%$? If this happened, do you think the European stock market might be heating up? Explain.

Problems 19, 20, and 21 use the central limit theorem or Theorem 7.1 to solve problems involving a *sum* of random variables.

19. (a) w is the sum of the waiting times for 30 customers, so we want to find $P(w < 90)$.
(b) If we divide both sides of $w < 90$ by 30, we get $w/30 < 3$. However, w is the sum of 30 waiting times, so $w/30$ is $\bar{x}$. Therefore, $P(w < 90) = P(\bar{x} < 3)$.
(c) Approximately normal, with mean 2.7 minutes and standard deviation 0.1095 minute.
(d) 0.9969.

19. ***Expand Your Knowledge: Totals Instead of Averages*** Let x be a random variable that represents checkout time (time spent in the actual checkout process) in minutes in the express lane of a large grocery. Based on a consumer survey, the mean of the x distribution is about $\mu = 2.7$ minutes, with standard deviation $\sigma = 0.6$ minute. Assume that the express lane always has customers waiting to be checked out and that the distribution of x values is more or less symmetrical and mound-shaped. What is the probability that the *total* checkout time for the next 30 customers is less than 90 minutes? Let us solve this problem in steps.
(a) Let x_i (for $i = 1, 2, 3, \ldots, 30$) represent the checkout time for each customer. For example, x_1 is the checkout time for the first customer, x_2 is the checkout time for the second customer, and so forth. Each x_i has mean $\mu = 2.7$ minutes and standard deviation $\sigma = 0.6$ minute. Let $w = x_1 + x_2 + \cdots + x_{30}$. Explain why the problem is asking us to compute the probability that w is less than 90.
(b) Use a little algebra and explain why $w < 90$ is mathematically equivalent to $w/30 < 3$. Since w is the total of the 30 x values, then $w/30 = \bar{x}$. Therefore, the statement $\bar{x} < 3$ is equivalent to the statement $w < 90$. From this we conclude that the probabilities $P(\bar{x} < 3)$ and $P(w < 90)$ are equal.
(c) What does the central limit theorem say about the probability distribution of $\bar{x}$? Is it approximately normal? What are the mean and standard deviation of the $\bar{x}$ distribution?
(d) Use the result of part (c) to compute $P(\bar{x} < 3)$. What does this result tell you about $P(w < 90)$?

20. (a) 0.8238.
(b) 0.9808.
(c) 0.8046.

20. ***Totals Instead of Averages: Airplane Takeoff Time*** The taxi and takeoff time for commercial jets is a random variable x with a mean of 8.5 minutes and a standard deviation of 2.5 minutes. Assume that the distribution of taxi and takeoff times is approximately normal. You may assume that the jets are lined up on a runway so that one taxies and takes off immediately after the other, and that they take off one at a time on a given runway. What is the probability that for 36 jets on a given runway, total taxi and takeoff time will be
(a) less than 320 minutes?
(b) more than 275 minutes?
(c) between 275 and 320 minutes?
Hint: See Problem 19.

21. (a) 0.2483.
(b) 0.2483.
(c) 0.5034.

21. ***Totals Instead of Averages: Escape Dunes*** It's true—sand dunes in Colorado rival sand dunes of the Great Sahara Desert! The highest dunes at Great Sand Dunes National Monument can exceed the highest dunes in the Great Sahara, extending over 700 feet in height. However, like all sand dunes, they tend to move around in the wind. This can cause a bit of trouble for temporary structures located near the "escaping" dunes. Roads, parking lots, campgrounds, small buildings, trees, and other vegetation are destroyed when a sand dune moves in and takes over. Such dunes are called "escape dunes" in the sense that they move out of the main body of sand dunes and, by the force of nature (prevailing winds), take over whatever space they choose to occupy. In most cases, dune movement does not occur quickly. An escape dune can take years to relocate itself. Just how fast does an escape dune move? Let x be a random variable representing movement (in feet per year) of such sand dunes (measured from the crest of the dune). Let us assume that x has a normal distribution with $\mu = 17$ feet per year and $\sigma = 3.3$ feet per year. (For more information, see *Hydrologic, Geologic, and Biologic Research at Great Sand Dunes National Monument and Vicinity, Colorado,* proceedings of the National Park Service Research Symposium.)

Under the influence of prevailing wind patterns, what is the probability that
(a) an escape dune will move a total distance of more than 90 feet in 5 years?
(b) an escape dune will move a total distance of less than 80 feet in 5 years?
(c) an escape dune will move a total distance of between 80 and 90 feet in 5 years?
Hint: See Problem 19 and Theorem 7.1.

SECTION 7.3

Sampling Distributions for Proportions

FOCUS POINTS

- Compute the mean and standard deviation for the sample proportion $\hat{p} = r/n$.
- Use the normal approximation to compute probabilities for proportions $\hat{p} = r/n$.
- Construct *P*-Charts and interpret their meaning.

In Section 6.4, we discussed the normal approximation to the binomial. There are many important situations in which we prefer to work with the *proportion* of successes r/n rather than the actual *number* of successes r in binomial experiments. With this in mind, we present the following summary about the *sampling distribution* of the proportion $\hat{p} = r/n$.

Sampling distribution for the proportion $\hat{p} = \frac{r}{n}$

Given n = number of binomial trials (fixed constant)

r = number of successes

p = probability of success on each trial

$q = 1 - p$ = probability of failure on each trial

If $np > 5$ and $nq > 5$, then the random variable $\hat{p} = r/n$ can be approximated by a normal random variable (x) with mean and standard deviation

$$\mu_{\hat{p}} = p \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

TERMINOLOGY The *standard error* for the $\hat{p}$ distribution is the standard deviation $\sigma_{\hat{p}}$ of the $\hat{p}$ sampling distribution.

COMMENT To obtain the information regarding the sampling distribution for the proportion $\hat{p} = r/n$, we consider the sampling distribution for r, the number of successes out of n binomial trials. In Section 6.4, we saw that when $np > 5$ and $nq > 5$, the r distribution is approximately normal, with mean $\mu_r = np$ and standard deviation $\sigma_r = \sqrt{npq}$. Notice that $\hat{p} = r/n$ is a linear function of r. This means that the $\hat{p}$ distribution is also approximately normal when np and nq are both greater than 5. In addition, from our work in Section 5.1 with linear functions of random variables, we know that $\mu_{\hat{p}} = \mu_r/n = np/n = p$ and $\sigma_{\hat{p}} = \sigma_r/n = \sqrt{npq}/n = \sqrt{pq/n}$.

This is a good time to review Section 6.4. Although the binomial distribution is used in both sections, the methods and formulas are different. It is especially useful to review the continuity correction process shown in Section 6.4.

If $np > 5$ and $nq > 5$, then $\hat{p} = r/n$ can be approximated by a normal random variable, which we will call x. However, $\hat{p}$ is *discrete* while x is *continuous*. To adjust for this discrepancy, we apply an appropriate *continuity correction*. In Section 6.4, we noted that in the probability histogram for the binomial random variable r, the bar centered over r actually starts at $r - 0.5$ and ends at $r + 0.5$. See Figure 7-3(a) on the next page.

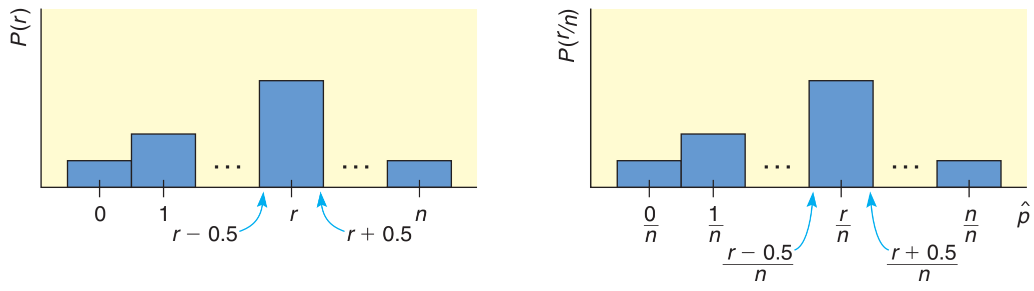
When we shift our focus from r to $\hat{p} = r/n$, the bar centered at r/n actually starts at $(r - 0.5)/n$ and ends at $(r + 0.5)/n$, as shown in Figure 7-3(b).

This leads us to the conclusion that the appropriate continuity correction for $\hat{p}$ is to add or subtract $0.5/n$ to the endpoints of a $\hat{p}$ (discrete) interval to convert it to an x (continuous normal) interval.

FIGURE 7-3

Distribution of r and Corresponding Distribution of $\hat{p} = r/n$

(a) Binomial Random Variable r (b) Sampling Distribution for $\hat{p} = r/n$



PROCEDURE

Emphasize that the continuity correction is more prominent for smaller n values.

HOW TO MAKE CONTINUITY CORRECTIONS TO $\hat{p}$ INTERVALS

1. If r/n is the *right* endpoint of a $\hat{p}$ interval, we *add* $0.5/n$ to get the corresponding right endpoint of the x interval.
2. If r/n is the *left* endpoint of a $\hat{p}$ interval, we *subtract* $0.5/n$ to get the corresponding left endpoint of the x interval.

The next example illustrates the process of using the continuity correction to convert a $\hat{p}$ interval to an x (normal) interval.

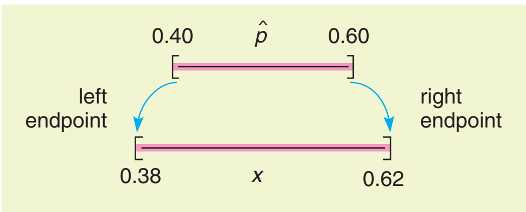
EXAMPLE 4 CONTINUITY CORRECTION

Suppose $n = 25$ and we have a $\hat{p}$ interval from $10/25 = 0.40$ to $15/25 = 0.60$. Use the continuity correction to convert this interval to an x interval.

SOLUTION: Since $n = 25$, then $0.5/n = 0.5/25 = 0.02$. This means that we subtract 0.02 from the left endpoint and add 0.02 to the right endpoint of the $\hat{p}$ interval (see Figure 7-4).

FIGURE 7-4

$\hat{p}$ and x intervals with Continuity Correction 0.02



$\hat{p}$ interval: 0.40 to 0.60

x interval: $0.40 - 0.02$ to $0.60 + 0.02$, or 0.38 to 0.62

COMMENT If n is large, the continuity correction for $\hat{p}$ won't change the x interval much. However, for smaller n values, it can make a difference. When we use the $\hat{p}$ distribution in later chapters, the sample size is sufficiently large that we can ignore the continuity correction.

EXAMPLE 5 Sampling distribution of $\hat{p}$

The annual crime rate in the Capitol Hill neighborhood of Denver is 111 victims per 1000 residents. This means that 111 out of 1000 residents have been the victim of at least one crime (Source: *Neighborhood Facts*, Piton Foundation). For more information, visit the Brase/Brase statistics site at **college.hmco.com/pic/braseUS9e** and find the link to the Piton Foundation. These crimes range from relatively minor crimes (stolen hubcaps or purse snatching) to major crimes (murder). The Arms is an apartment building in this neighborhood that has 50 year-round residents. Suppose we view each of the $n = 50$ residents as a binomial trial. The random variable r (which takes on values 0, 1, 2, . . . , 50) represents the number of victims of at least one crime in the next year.

(a) What is the population probability p that a resident in the Capitol Hill neighborhood will be the victim of a crime next year? What is the probability q that a resident will not be a victim?

SOLUTION: Using the Piton Foundation report, we take

$$p = 111/1000 = 0.111 \qquad \text{and} \qquad q = 1 - p = 0.889$$

(b) Consider the random variable

$$\hat{p} = \frac{r}{n} = \frac{r}{50}$$

Do you think we can approximate $\hat{p}$ with a normal distribution? Explain.

SOLUTION:

$$np = 50(0.111) = 5.55$$
$$nq = 50(0.889) = 44.45$$

Since both np and nq are greater than 5, we can approximate $\hat{p}$ with a normal distribution.

(c) What are the mean and standard deviation for $\hat{p}$?

SOLUTION:

$$\mu_{\hat{p}} = p = 0.111$$
$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.111)(0.889)}{50}} \approx 0.044$$

Compare this solution with the methods of Section 6.4. Here we talk about rates (between 10% and 20%). In Section 6.4, we talk about the number r of victims.

(d) What is the probability that between 10% and 20% of the Arms residents will be victims of a crime next year? Interpret the results.

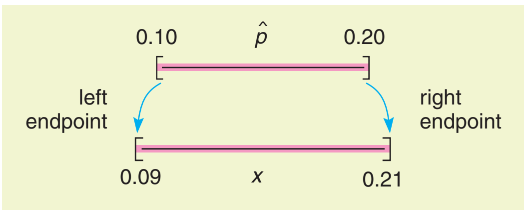
SOLUTION: First we find the continuity correction so that we can convert the $\hat{p}$ interval to an x interval. Since $n = 50$,

$$\text{Continuity correction} = 0.5/n = 0.5/50 = 0.01$$

We subtract 0.01 from the left $\hat{p}$ endpoint and add 0.01 to the right $\hat{p}$ endpoint (see Figure 7-5).

FIGURE 7-5

$\hat{p}$ and x Intervals with Continuity Correction 0.01



The x interval is from 0.09 to 0.21. Therefore,

$$\begin{aligned} P(0.10 \le \hat{p} \le 0.20) &\approx P(0.09 \le x \le 0.21) \\ &\approx P\left(\frac{0.09 - 0.111}{0.044} \le z \le \frac{0.21 - 0.111}{0.044}\right) \\ &\approx P(-0.48 \le z \le 2.25) \\ &\approx 0.6722 \end{aligned}$$

INTERPRETATION There is about a 67% chance that between 10% and 20% of the Arms residents will be crime victims next year.

GUIDED EXERCISE 4 Sampling distribution of $\hat{p}$

The general ethnic profile of Denver is about 42% minority and 58% Caucasian (Source: *Neighborhood Facts,* Piton Foundation). Suppose the city of Denver recently hired 56 new grounds and maintenance workers. It was claimed that the hiring practice was completely impartial with regard to ethnic background. However, only 27% of the new employees are minorities, and now there is a complaint. What is the probability that at most 27% of the new hires will be minorities if the selection process is impartial and the applicant pool reflects the ethnic profile of Denver?

(a) We take the point of view that each new hire is a binomial experiment, with success being a minority hire. What are the values of n, p, q?

→ $n = 56$; $p = 0.42$; $q = 0.58$

(b) For the new hires, what is the sample proportion $\hat{p}$ of minority hires?

→ $\hat{p} = 0.27$

(c) Is the normal approximation for the distribution of $\hat{p}$ appropriate? Explain.

→ $np = 56(0.42) \approx 23.5$;
$nq = 56(0.58) \approx 32.5$

Both products are larger than 5, so the normal approximation is appropriate.

(d) Compute $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.

→ $\mu_{\hat{p}} = p = 0.42$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.42)(0.58)}{56}} \approx 0.066$$

(e) Compute $P(\hat{p} \le 0.27)$.

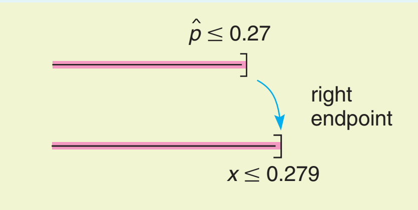
→ First find the continuity correction and convert $\hat{p} \le 0.27$ to an x interval (see Figure 7-6).

Continuity correction: $\frac{0.5}{56} \approx 0.009$

x interval: $x \le 0.279$

$$\begin{aligned} P(\hat{p} \le 0.27) &\approx P(x \le 0.279) \\ &\approx P\left(z \le \frac{0.279 - 0.42}{0.066}\right) \\ &\approx P(z \le -2.14) \\ &\approx 0.0162 \end{aligned}$$

FIGURE 7-6 $\hat{p}$ and x Intervals with Continuity Correction 0.009



(f) Interpret the results.

Ask the class to look in their local news media for similar examples.

→ Assuming all the conditions for binomial trials have been met, the probability is smaller than 2% that the proportion of minority hires would be 27% or less. It seems the hiring process might not be completely impartial or the applicant pool does not reflect the ethnic profile of Denver.

Control Charts for Proportions (P-Charts)

We conclude this section with an example of a control chart for proportions r/n. Such a chart is often called a *P-Chart*.

The control charts discussed in Section 6.1 were for *quantitative* data, where the *size* of something is being measured. There are occasions where we prefer to examine a *quality* or *attribute* rather than just size. One way to do this is to use a binomial distribution in which success is defined as the quality or attribute we wish to study.

The basic idea for using P-Charts is to select samples of a fixed size n at regular time intervals and count the number of successes r from the n trials. We use the normal approximation for r/n and methods of Section 6.1 to plot control limits and r/n values, and to interpret results.

This is a good time to review out-of-control signals from Section 6.1. Remind students that control charts often are made for specific attributes (not just size).

As in Section 6.1, we remind ourselves that control charts are used as warning devices tailored by a user for a particular need. Our assumptions and probability calculations need not be absolutely precise to achieve our purpose. For example, $\hat{p} = r/n$ need not follow a normal distribution exactly. A mound-shaped and more or less symmetric distribution to which the empirical rule applies will be sufficient.

EXAMPLE 6 *P*-Chart

Anatomy and Physiology is taught each semester. The course is required for several popular health-science majors, so it always fills up to its maximum of 60 students. The dean of the college asked the biology department to make a control chart for the proportion of A's given in the course each semester for the past 14 semesters. Using information from the registrar's office, the following data were obtained. Make a control chart and interpret the result.

Semester	1	2	3	4	5	6	7
r = no. of A's	9	12	8	15	6	7	13
$\hat{p} = r/60$	0.15	0.20	0.13	0.25	0.10	0.12	0.22

Semester	8	9	10	11	12	13	14
r = no. of A's	7	11	9	8	21	11	10
$\hat{p} = r/60$	0.12	0.18	0.15	0.13	0.35	0.18	0.17

SOLUTION: Let us view each student as a binomial trial, where success is the quality or attribute we wish to study. Success means the student got an A, and failure is not getting an A. Since the class size is 60 students each semester, the number of trials is $n = 60$.

(a) The first step is to use the data to estimate the overall proportion of successes. To do this, we pool the data for all 14 semesters, and use the symbol $\bar{p}$ (not to be confused with $\hat{p}$) to designate the *pooled proportion* of success.

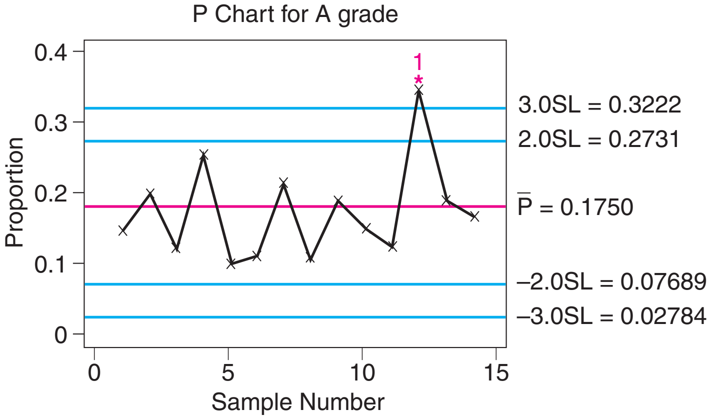
$$\bar{p} = \frac{\text{Total number of A's from all 14 semesters}}{\text{Total number of students from all 14 semesters}}$$

$$\bar{p} = \frac{9 + 12 + 8 + \cdots + 10}{14(60)} = \frac{147}{840} = 0.175$$

Since the pooled estimate for the proportion of successes is $\bar{p} = 0.175$, the estimate for the proportion of failures is $\bar{q} = 1 - \bar{p} = 0.825$.

FIGURE 7-7

P-Chart for Proportion of A's (Minitab generated)



(b) For the random variable $\hat{p} = r/n$, we know the mean is $\mu_{\hat{p}} = p$ and the standard deviation is $\sigma_{\hat{p}} = \sqrt{pq/n}$. In our case, we don't have given values for p and q, so we use the pooled estimates $\bar{p} = 0.175$ and $\bar{q} = 0.825$. The number of trials is the class size $n = 60$. Therefore,

$$\mu_{\hat{p}} = p \approx \bar{p} = 0.175$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\bar{p}\,\bar{q}}{n}} = \sqrt{\frac{(0.175)(0.825)}{60}} \approx 0.049$$

(c) Since a control chart is a *warning device*, it is not necessary that our probability calculations be absolutely precise. For instance, the empirical rule would substitute quite well for the normal distribution. However, it is a good idea to check that both $n\bar{p} = 60(0.175) = 10.5$ and $n\bar{q} = 60(0.825) = 49.5$ are larger than 5. This means the normal approximation should be reasonably good.

(d) Now we use the basic methods of Section 6.1 to construct the control limits and control chart shown in Figure 7-7. The center line is at $\bar{p} = 0.175$.

$$\text{Control limits at } \bar{p} \pm 2\sqrt{\frac{\bar{p}\,\bar{q}}{n}} = 0.175 \pm 2(0.049), \text{ or } 0.077 \text{ and } 0.273$$

$$\text{Control limits at } \bar{p} \pm 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} = 0.175 \pm 3(0.049), \text{ or } 0.028 \text{ and } 0.322$$

(e) Interpretation: We use the three out-of-control signals discussed in Section 6.1.

Signal I—beyond the 3σ level.
We see semester number 12 was above the 3σ level. That semester the class must have been very good indeed!

Signal II—run of nine *consecutive* points on one side of center line.
Since this did not happen, there is no slow drift either up or down.

Signal III—at least two out of three *consecutive* points beyond the 2σ level (on the same side of center).
This out-of-control signal did not occur.

(f) Conclusion: The biology department can tell the dean that the proportion of A's given in Anatomy and Physiology is in statistical control, with the exception of one unusually good class two semesters ago.

PROCEDURE

HOW TO MAKE A *P*-CHART

1. Estimate $\bar{p}$, the overall proportion of successes.

$$\bar{p} = \frac{\text{Total number of observed successes in all samples}}{\text{Total number of trials in all samples}}$$

2. The center line of the control chart is assigned to be $\mu_{\hat{p}} = \bar{p}$.
3. Control limits are located at

$$\bar{p} \pm 2\sqrt{\frac{\bar{p}\bar{q}}{n}} \quad \text{and} \quad \bar{p} \pm 3\sqrt{\frac{\bar{p}\bar{q}}{n}} \quad \text{where } \bar{q} = 1 - \bar{p}$$

4. Interpretation: Out-of-control signals
   (a) Signal I: any point beyond a $\bar{p} \pm 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$ control limit.
   (b) Signal II: run of nine *consecutive* points on *one side* of the center line $\mu_{\hat{p}} = \bar{p}$.
   (c) Signal III: at least two out of three *consecutive* points beyond a $\bar{p} \pm 2\sqrt{\frac{\bar{p}\bar{q}}{n}}$ control limit (on the same side).

If no out-of-control signals occur, we say that the process is "in control," while keeping a watchful eye on what occurs next.

COMMENT In some *P*-Charts, the value of $\bar{p}$ may be near 0 or near 1. In this case, the control limits may drop below 0 or rise above 1. If this happens, we follow the usual convention of rounding negative control limits to 0 and rounding control limits above 1 to 1.

VIEWPOINT Happy Memories! False Memories!!!

"Memory isn't a record; it's an interpretation," says psychologist Mark Reinitz of the University of Puget Sound, Washington. Research indicates that about 68% of all people occasionally "fill in the blanks" in their memories. They claim to remember things that did not actually occur. In another study, it was found that 35% of visitors to Disneyland claimed they shook hands with Bugs Bunny, who welcomed them at the entrance. However, Bugs Bunny is not *a Disney character and was* not *at the Disneyland entrance.*

Psychologists say memory is malleable for a reason. It helps us view ourselves in a more positive light. Statistical interpretation of proportions $\hat{p}$ *are useful in this memory study, as well as all areas of natural science, business, linguistics, and social science. For more information about the memory study, see the July 2001 issue of the* Journal of Experimental Psychology.

SECTION 7.3 PROBLEMS

Tables and art to accompany margin answers may be found in the back of the book.

1. *Statistical Literacy* Under what conditions is it appropriate to use a normal distribution to approximate the $\hat{p}$ distribution?
2. *Statistical Literacy* What is the formula for the standard error of the normal approximation to the $\hat{p}$ distribution? What is the mean of the $\hat{p}$ distribution?
3. *Statistical Literacy* Is $\hat{p}$ an unbiased estimator for p when $np > 5$ and $nq > 5$? Recall that a statistic is an unbiased estimator of the corresponding parameter if the mean of the sampling distribution equals the parameter in question.

1. $np > 5$ and $nq > 5$.
2. $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$; $\mu_{\hat{p}} = p$.
3. Yes, since the mean of the approximate sampling distribution is $\mu_{\hat{p}} = p$.
4. (a) 0.5/25, or 0.02.
   (b) 0.5/100, or 0.005.
   (c) As n increases, the continuity correction value decreases.
5. (a) Yes; both np and nq exceed 5. $\mu_{\hat{p}} = 0.21$; $\sigma_{\hat{p}} = 0.071$; continuity correction ≈ 0.015; 0.6348.
   (b) No; $np < 5$.
   (c) Yes; both np and nq exceed 5. $\mu_{\hat{p}} = 0.15$; $\sigma_{\hat{p}} = 0.052$; continuity correction $= 0.010$; 0.1251.

4. *Critical Thinking* Consider a binomial experiment with n trials for which $np > 5$ and $nq > 5$. What is the value of the continuity correction when
   (a) $n = 25$?
   (b) $n = 100$?
   (c) As the value of n increases, does the continuity correction value increase, decrease, or stay the same?

5. *Critical Thinking* Suppose we have a binomial experiment in which success is defined to be a particular quality or attribute that interests us.
   (a) Suppose $n = 33$ and $p = 0.21$. Can we approximate $\hat{p}$ by a normal distribution? Why? What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$? What is the value of the continuity correction? Compute $P(0.15 \le \hat{p} \le 0.25)$.
   (b) Suppose $n = 25$ and $p = 0.15$. Can we safely approximate $\hat{p}$ by a normal distribution? Why or why not?
   (c) Suppose $n = 48$ and $p = 0.15$. Can we approximate $\hat{p}$ by a normal distribution? Why? What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$? If a survey, experiment, or laboratory work gives us a $\hat{p}$ value of 0.22, what is the probability of getting a $\hat{p}$ value this high or higher?

6. (a) 0.7777.
   (b) 0.1075.
   (c) No; $np < 5$.

6. *Critical Thinking* Suppose we have a binomial distribution with n trials and probability of success p. The random variable r is the number of successes in the n trials, and the random variable representing the proportion of successes is $\hat{p} = r/n$.
   (a) $n = 50$; $p = 0.36$; Compute $P(0.30 \le \hat{p} \le 0.45)$.
   (b) $n = 38$; $p = 0.25$; Compute the probability that $\hat{p}$ will exceed 0.35.
   (c) $n = 41$; $p = 0.09$; Can we approximate $\hat{p}$ by a normal distribution? Explain.

7. (a) 0.9049.
   (b) 0.2877.
   (c) 0.0025.
   (d) Yes; both np and nq exceed 5.

7. *Sociology: Criminal Justice* Courts sometimes make mistakes, but which do you believe is the worse mistake: convicting an innocent person or letting a guilty person go free? It turns out that about 60% of all Americans believe that convicting an innocent person is the worse mistake. (Source: *American Attitudes* by S. Mitchell, Sociology Department, Ithaca College.) Suppose you are taking a sociology class with 30 students enrolled. The question discussed today is: Do you agree with the statement that convicting an innocent person is worse than letting the guilty go free? What is the probability that the proportion of the class who agree is
   (a) at least one half?
   (b) at least two thirds?
   (c) no more than one third?
   (d) Is the normal approximation to the proportion $\hat{p} = r/n$ valid? Explain.

8. (a) 0.8554.
   (b) Approximately 1.
   (c) Yes; both np and nq exceed 5 for men and women.

8. *Sociology: Gun Permits* Would you favor a law requiring a police permit to buy a gun? About 73% of American men and 86% of American women would favor such a law. (Source: See Problem 7.)
   (a) A candidate for city council is speaking to a breakfast group of 38 men. The topic of gun permits comes up. What is the probability that the majority of the audience (at least two thirds) will support gun permits? Assume the group is representative of all U.S. men.
   (b) Answer part (a) if our candidate is speaking to a women's seminar with 45 women in the audience. Assume the group is representative of all U.S. women.
   (c) Is the normal approximation to the proportion $\hat{p} = r/n$ valid in both applications? Explain.

9. (a) Both np and nq exceed 5; $\mu_{\hat{p}} = 0.06$; $\sigma_{\hat{p}} = 0.024$.
   (b) 0.4168.
   (c) 0.0301. Yes; the probability of this proportion of defective toys is only about 3%.

9. *Manufacturing: Defective Toys* A mechanical press is used to mold shapes for plastic toys. When the machine is adjusted and working well, it still produces about 6% defective toys. The toys are manufactured in lots of $n = 100$. Let r be a random variable representing the number of defective toys in a lot. Then $\hat{p} = r/n$ is the proportion of defective toys in a lot.
   (a) Explain why $\hat{p}$ can be approximated by a normal random variable. What are $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

(b) Suppose a lot of 100 toys had a 7% proportion of defective toys. What is the probability that a situation this bad or worse could occur? Compute $P(0.07 \le \hat{p})$.

(c) *Interpretation:* Suppose a lot of 100 toys had an 11% proportion of defective toys. What is the probability that a situation this bad or worse could occur? Compute $P(0.11 \le \hat{p})$. Do you think the machine might need an adjustment? Explain.

10. (a) Both np and nq exceed 5; $\mu_{\hat{p}} = 0.565$; $\sigma_{\hat{p}} = 0.070$.
(b) 0.3594.
(c) 0.0192.
(d) Meredith; the probability of having such a low reading in a healthy person is less than 2%.

10. *Medical Tests: Leukemia* Healthy adult bone marrow contains about 56.5% neutrophils (a particular type of white blood cell). However, if this level is significantly reduced, it may be an early indicator of leukemia (Reference: *Diagnostic Tests with Nursing Implications,* edited by S. Loeb, Springhouse).

(a) In a laboratory biopsy, a field of $n = 50$ bone marrow cells are observed under a microscope. A special dye is inserted, which only the neutrophils absorb. Then the number r of neutrophils in the field is counted. Although the field size $n = 50$ is fixed, the number of neutrophils r is a random variable, so the proportion $\hat{p} = r/n$ is also a random variable. Explain why $\hat{p}$ can be approximated by a normal random variable. What are $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

(b) Suppose Jan had a bone marrow biopsy and $\hat{p}$ was observed to be 0.53. Assuming nothing is wrong (no leukemia), what is the probability of getting a biopsy result this low or lower? Compute $P(\hat{p} \le 0.53)$.

(c) Suppose Meredith had a bone marrow biopsy and $\hat{p}$ was observed to be 0.41. Assuming nothing is wrong (no leukemia), what is the probability of getting a biopsy result this low or lower? Compute $P(\hat{p} \le 0.41)$.

(d) *Interpretation:* Based on the probability estimates in parts (b) and (c), which do you think is the more serious case, Jan or Meredith? Explain.

11. No out-of-control signals.

11. *P-Chart: Property Crime* Lee is a cadet at the Honolulu Police Academy. He was asked to make a *P*-Chart for reported (minor) property crimes. Lee chose a small neighborhood with 92 families. Each family is viewed as a binomial trial. Success means that the family was the victim of at least one minor property crime in the past 3 months. Police reports gave the following data for the past 12 quarters (4 years). Assume the 92 families lived in the neighborhood all 4 years.

Quarter	1	2	3	4	5	6
r = no. of successes	11	14	18	23	19	15
$\hat{p} = r/92$	0.12	0.15	0.20	0.25	0.21	0.16

Quarter	7	8	9	10	11	12
r = no. of successes	12	16	13	22	24	19
$\hat{p} = r/92$	0.13	0.17	0.14	0.24	0.26	0.21

Make a *P*-Chart and list any out-of-control signals by type (I, II, or III).

12. No out-of-control signals. It appears that the production process is in reasonable control.

12. *P-Chart: Aluminum Cans* A high-speed metal stamp machine produces 12-ounce aluminum beverage cans. The cans are mass produced in lots of 110 cans for each square sheet of aluminum fed into the machine. However, some of the cans come out of the die stamp with folds and wrinkles. These are defective cans that must be recycled. Let us view each can as a binomial trial, where success is defined to mean the can is defective. So we have $n = 110$ trials (cans), and the random variable r is the number of defective cans. A test run of 15 consecutive aluminum sheets gave the following numbers r of defective cans.

Test sheet	1	2	3	4	5	6	7	8
r	8	11	6	9	12	8	7	11
$\hat{p} = r/110$	0.07	0.10	0.05	0.08	0.11	0.07	0.06	0.10

Test sheet	9	10	11	12	13	14	15
r	10	7	9	6	12	7	10
$\hat{p} = r/110$	0.09	0.06	0.08	0.05	0.11	0.06	0.09

Make a *P*-Chart and list any out-of-control signals by type (I, II, or III). Does it appear from the sequential test runs that the production process is in reasonable control? Explain.

13. Out-of-control signal III occurs on days 4 and 5; out-of-control signal I occurs on day 11 on the low side and day 14 on the high side. Out-of-control signals on the low side are of most concern for the homeless seeking work. The foundation should look to see what happened on that day. The foundation might take a look at the out-of-control periods on the high side to see if there is a possibility of cultivating more jobs.

13. ***P-Chart: Temporary Work*** Jobs for the homeless! A philanthropic foundation bought a used school bus that stops at homeless shelters early every weekday morning. The bus picks up people looking for temporary, unskilled day jobs. The bus delivers these people to a work center and later picks them up after work. The bus can hold 75 people, and it fills up every morning. Not everyone finds work, so at 11 A.M. the bus goes to a soup kitchen where those not finding work that day volunteer their time. Let us view each person on the bus looking for work as a binomial trial. Success means he or she got a day job. The random variable r represents the number who got jobs. The foundation requested a *P*-Chart for the success ratios. For the past 3 weeks, we have the following data.

Day	1	2	3	4	5	6	7	8
r	60	53	61	66	67	55	53	58
$\hat{p} = r/75$	0.80	0.71	0.81	0.88	0.89	0.73	0.71	0.77

Day	9	10	11	12	13	14	15
r	60	52	46	52	61	70	58
$\hat{p} = r/75$	0.80	0.69	0.61	0.69	0.81	0.93	0.77

Make a *P*-Chart, list any out-of-control signals, and interpret the results.

VIEWPOINT Why Wait? Apply Now for a College Loan!

The cost of education is high. The cost of not having an education is higher! What kinds of education costs can you expect? What about tuition and student fees? What about room and board? What is the total cost for 1 year at college? Perhaps some averages based on random samples of colleges would be useful. For more information, visit the Brase/Brase statistics site at **college.hmco.com/pic/braseUS9e** *and find the link to the U.S. News site. Then select education. Search for the geographic regions of your colleges of interest.*

Chapter Review

SUMMARY

Sampling distributions give us the basis for inferential statistics. By studying the distribution of a sample statistic, we can learn about the corresponding population parameter.

- For random samples of size n, the $\bar{x}$ distribution is the sampling distribution for the sample mean of an x distribution with population mean μ and population standard deviation σ. If the x distribution is normal, then the corresponding $\bar{x}$ distribution is normal.

  By the central limit theorem, when n is sufficiently large ($n \geq 30$), the $\bar{x}$ distribution is approximately normal even if the original x distribution is not normal.

In both cases,

$$\mu_{\bar{x}} = \mu$$

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

- For n binomial trials with probability of success p on each trial, the $\hat{p}$ distribution is the sampling distribution of the sample proportion of successes. When $np > 5$ and $nq > 5$, the $\hat{p}$ distribution is approximately normal with

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

- P-Charts are control charts for proportions.

IMPORTANT WORDS & SYMBOLS

Section 7.1
Population parameter
Statistic
Sampling distribution

Section 7.2
$\mu_{\bar{x}}$
$\sigma_{\bar{x}}$
Standard error of the mean
Central limit theorem

Section 7.3
Sampling distribution for $\hat{p}$
$\mu_{\hat{p}}$
$\sigma_{\hat{p}}$
Standard error of a proportion
Continuity correction, $0.5/n$
P-Chart

CHAPTER REVIEW PROBLEMS

Tables and art to accompany margin answers may be found in the back of the book.

1. (a) A normal distribution.
   (b) The mean μ of the x distribution.
   (c) $\sigma/\sqrt{n}$, where σ is the standard deviation of the x distribution.
   (d) Approximately normal with the same mean, but the standard deviations will be $\sigma/\sqrt{50}$ and $\sigma/\sqrt{100}$, respectively.
2. All the $\bar{x}$ distributions will be normal with mean 15. The standard deviations will be 3/2, 3/4, and 3/10, respectively.

1. *Critical Thinking* Let x be a random variable representing the amount of sleep each adult in New York City got last night. Consider a sampling distribution of sample means $\bar{x}$.
   (a) As the sample size becomes increasingly large, what distribution does the $\bar{x}$ distribution approach?
   (b) As the sample size becomes increasingly large, what value will the mean $\mu_{\bar{x}}$ of the $\bar{x}$ distribution approach?
   (c) What value will the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution approach?
   (d) How do the two $\bar{x}$ distributions for sample size $n = 50$ and $n = 100$ compare?

2. *Critical Thinking* If x has a normal distribution with mean $\mu = 15$ and standard deviation $\sigma = 3$, describe the distribution of $\bar{x}$ values for sample size n, where $n = 4$, $n = 16$, and $n = 100$. How do the $\bar{x}$ distributions compare for the various sample sizes?

3. (a) 0.2389.
(b) 0.0162.

3. ***Job Interview: Length*** The personnel office at a large electronics firm regularly schedules job interviews and maintains records of the interviews. From the past records, they have found that the length of a first interview is normally distributed, with mean $\mu = 35$ minutes and standard deviation $\sigma = 7$ minutes.
(a) What is the probability that a first interview will last 40 minutes or longer?
(b) Nine first interviews are usually scheduled per day. What is the probability that the average length of time for the nine interviews will be 40 minutes or longer?

4. (a) 0.2743.
(b) 0.0287.
(c) The standard deviation of $\bar{x}$ is smaller.

4. ***Drugs: Effects*** A new muscle relaxant is available. Researchers from the firm developing the relaxant have done studies that indicate that the time lapse between administration of the drug and beginning effects of the drug is normally distributed, with mean $\mu = 38$ minutes and standard deviation $\sigma = 5$ minutes.
(a) The drug is administered to one patient selected at random. What is the probability that the time it takes to go into effect is 35 minutes or less?
(b) The drug is administered to a random sample of 10 patients. What is the probability that the average time before it is effective for all 10 patients is 35 minutes or less?
(c) Comment on the differences of the results in parts (a) and (b).

5. 0.8164.

5. ***Psychology: IQ Scores*** Assume that IQ scores are normally distributed, with a standard deviation of 15 points and a mean of 100 points. If 100 people are chosen at random, what is the probability that the sample mean of IQ scores will not differ from the population mean by more than 2 points?

6. 0.8664.

6. ***Hatchery Fish: Length*** A large tank of fish from a hatchery is being delivered to a lake. The hatchery claims that the mean length of fish in the tank is 15 inches, and the standard deviation is 2 inches. A random sample of 36 fish is taken from the tank. Let $\bar{x}$ be the mean sample length of these fish. What is the probability that $\bar{x}$ is within 0.5 inch of the claimed population mean?

7. (a) 0.8790.
(b) 0.5526.
(c) Yes, both *np* and *nq* exceed 5.

7. ***Who's Who: Misinformation*** About 11% of Americans believe that Joan of Arc was Noah's wife. (Source: *Harper's Index,* Volume 3.) At a large freshman symposium, a college professor (jokingly) says that Joan of Arc was Noah's wife. Assume that college freshmen are representative of the general American population regarding biblical knowledge.
(a) If the symposium was attended by 55 freshmen, what is the probability that up to 15% of the freshmen believe the professor's claim?
(b) What is the probability that between 10% and 15% (including 10% and 15%) of the freshmen believe the professor's claim?
(c) Is the normal approximation to the proportion $\hat{p} = r/n$ valid in this application? Explain.

8. (a) 0.8159.
(b) 0.8075.
(c) Yes, both *np* and *nq* exceed 5.

8. ***Grand Canyon: Boating Accidents*** Thomas Myers is a staff physician at the clinic in Grand Canyon Village. Based on reports in recent years, Dr. Myers estimates that about 31% of the boating accidents on the Colorado River in Grand Canyon National Park occur at Crystal Rapids (mile 98). These range from small accidents (a few bruises) to major accidents (death). (Source: *Fateful Journey,* Myers, Becker, and Stevens.) In the next 28 boating accidents to be reported,
(a) what is the probability that at least 25% of these accidents occurred at Crystal Rapids?
(b) what is the probability that between 25% and 50% (including 25% and 50%) of these accidents occurred at Crystal Rapids?
(c) Is the normal approximation to the proportion $\hat{p} = r/n$ valid in this application? Explain.

9. (a) 0.4467.
(b) 0.1762.
(c) No, $np < 5$.

9. ***Critical Thinking*** Suppose we have a binomial distribution with n trials and probability of success p. The random variable r is the number of successes in the n trials, and the random variable representing the proportion of successes is $\hat{p} = r/n$.
(a) $n = 50$; $p = 0.22$; Compute $P(0.20 \leq \hat{p} \leq 0.25)$.
(b) $n = 38$; $p = 0.27$; Compute the probability that $\hat{p}$ will equal or exceed 0.35.
(c) $n = 51$; $p = 0.05$; Can we approximate $\hat{p}$ by a normal distribution? Explain.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Iris setosa is a beautiful wildflower that is found in such diverse places as Alaska, the Gulf of St. Lawrence, much of North America, and even in English meadows and parks. R. A. Fisher, with his colleague Dr. Edgar Anderson, studied these flowers extensively. Dr. Anderson described how he collected information on irises:

Wild iris

> I have studied such irises as I could get to see, in as great detail as possible, measuring iris standard after iris standard and iris fall after iris fall, sitting squat-legged with record book and ruler in mountain meadows, in cypress swamps, on lake beaches, and in English parks. [Anderson, E., "The Irises of the Gaspé Peninsula," *Bulletin, American Iris Society*, 59:2–5, 1935.]

The data in Table 7-3 were collected by Dr. Anderson and were published by his friend and colleague R. A. Fisher in a paper entitled "The Use of Multiple Measurements in Taxonomic Problems" (*Annals of Eugenics,* part II, 179–188, 1936). To find these data, visit the Brase/Brase statistics site at **college.hmco.com/pic/braseUS9e** and find the link to DASL, the Carnegie Mellon University Data and Story Library. From the DASL site, look under famous data sets.

Let x be a random variable representing petal length. Using a TI-84Plus/TI-83Plus calculator, it was found that the sample mean is $\bar{x} = 1.46$ centimeters (cm) and the sample standard deviation is $s = 0.17$ cm. Figure 7-8 shows a histogram for the given data generated on a TI-84Plus/TI-83Plus calculator.

(a) Examine the histogram for petal lengths. Would you say that the distribution is approximately mound-shaped and symmetrical? Our sample has only 50 irises; if many thousands of irises had been used, do you think the distribution would look even more like a normal curve? Let x be the petal length of Iris setosa. Research has shown that x has an approximately normal distribution, with mean $\mu = 1.5$ cm and standard deviation $\sigma = 0.2$ cm.

(b) Use the empirical rule with $\mu = 1.5$ and $\sigma = 0.2$ to get an interval in which approximately 68% of the petal lengths will fall. Repeat this for 95% and 99.7%. Examine the raw data and compute the percentage of the raw data that actually falls into each of these intervals (the 68% interval, the 95% interval, and the 99.7% interval). Compare your computed percentages with those given by the empirical rule.

(c) Compute the probability that a petal length is between 1.3 and 1.6 cm. Compute the probability that a petal length is greater than 1.6 cm.

(d) Suppose that a random sample of 30 irises is obtained. Compute the probability that the average petal length for this sample is between 1.3 and 1.6 cm. Compute the probability that the average petal length is greater than 1.6 cm.

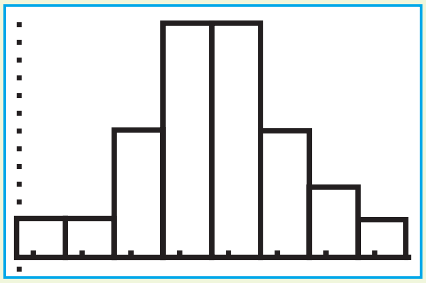
(e) Compare your answers to parts (c) and (d). Do you notice any differences? Why would these differences occur?

TABLE 7-3 Petal Length in Centimeters for *Iris setosa*

1.4	1.4	1.3	1.5	1.4
1.7	1.4	1.5	1.4	1.5
1.5	1.6	1.4	1.1	1.2
1.5	1.3	1.4	1.7	1.5
1.7	1.5	1	1.7	1.9
1.6	1.6	1.5	1.4	1.6
1.6	1.5	1.5	1.4	1.5
1.2	1.3	1.4	1.3	1.5
1.3	1.3	1.3	1.6	1.9
1.4	1.6	1.4	1.5	1.4

FIGURE 7-8

Petal Length (cm) for *Iris setosa* (TI-84Plus/TI-83Plus)



LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class, or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. Most people would agree that increased information should give better predictions. Discuss how sampling distributions actually enable better predictions by providing more information. Examine Theorem 7.1 again. Suppose that x is a random variable with a *normal* distribution. Then $\bar{x}$, the sample mean based on random samples of size n, also will have a normal distribution for *any* value of $n = 1, 2, 3, \ldots$

   What happens to the standard deviation of the $\bar{x}$ distribution as n (the sample size) increases? Consider the following table for different values of n.

n	1	2	3	4	10	50	100
$\sigma/\sqrt{n}$,	1σ	0.71σ	0.58σ	0.50σ	0.32σ	0.14σ	0.10σ

   In this case, "increased information" means a larger sample size n. Give a brief explanation as to why a *large* standard deviation will usually result in poor statistical predictions, whereas a *small* standard deviation usually results in much better predictions. Since the standard deviation of the sampling distribution $\bar{x}$ is $\sigma/\sqrt{n}$, we can decrease the standard deviation by increasing n. In fact, if we look at the preceding table, we see that if we use a sample size of only $n = 4$, we cut the standard deviation of $\bar{x}$ by 50% of the standard deviation σ of x. If we were to use a sample of size $n = 100$, we would cut the standard deviation of $\bar{x}$ to 10% of the standard deviation σ of x.

   Give the preceding discussion some thought and explain why you should get much better predictions for μ by using $\bar{x}$ from a sample of size n rather than by just using x. Write a brief essay in which you explain why sampling distributions are an important tool in statistics.

2. In a way, the central limit theorem can be thought of as a kind of "grand central station." It is a connecting hub or center for a great deal of statistical work. We will use it extensively in Chapters 8, 9, and 10. Put in a very elementary way, the central limit theorem states that as the sample size n increases, the distribution of the sample mean $\bar{x}$ will always approach a normal distribution, no matter where the original x variable came from. For most people, it is the complete generality of the central limit theorem that is so awe inspiring: It applies to practically everything. List and discuss at least three variables from everyday life for which you expect the variable x itself does *not* follow a normal or bell-shaped distribution. Then discuss what would happen to the sampling distribution $\bar{x}$ if the sample size were increased. Sketch diagrams of the $\bar{x}$ distributions as the sample size n increases.

Using Technology

As we have seen in this chapter, the value of a sample statistic such as $\bar{x}$ varies from one sample to another. The central limit theorem describes the distribution of the sample statistic $\bar{x}$ when samples are sufficiently large.

We can use technology tools to generate samples of the same size from the same population. Then we can look at the statistic $\bar{x}$ for each sample, and the resulting $\bar{x}$ distribution.

Project Illustrating the Central Limit Theorem

Step 1: Generate random samples of specified size n from a population.

The random-number table enables us to sample from the uniform distribution of digits 0 through 9. Use either the random-number table or a random-number generator to generate 30 samples of size 10.

Step 2: Compute the sample mean $\bar{x}$ of the digits in each sample.

Step 3: Compute the sample mean of the means (i.e., $\bar{x}_{\bar{x}}$) as well as the standard deviation $s_{\bar{x}}$ of the sample means.

The population mean of the uniform distribution of digits from 0 through 9 is 4.5. How does $\bar{x}_{\bar{x}}$ compare to this value?

Step 4: Compare the sample distribution of $\bar{x}$ values to a normal distribution having the mean and standard deviation computed in Step 3.

(a) Use the values of $\bar{x}_{\bar{x}}$ and $s_{\bar{x}}$ computed in Step 3 to create the intervals shown in column 1 of Table 7-4.

(b) Tally the sample means computed in Step 2 to determine how many fall into each interval of column 2. Then compute the percent of data in each interval and record the results in column 3.

(c) The percentages listed in column 4 are those from a normal distribution (see Figure 6-3 showing the empirical rule). Compare the percentages in column 3 to those in column 4. How do the sample percentages compare with the hypothetical normal distribution?

Step 5: Create a histogram showing the sample means computed in Step 2.

Look at the histogram and compare it to a normal distribution with the mean and standard deviation of the $\bar{x}$s (as computed in Step 3).

Step 6: Compare the results of this project to the central limit theorem.

Increase the sample size of Step 1 to 20, 30, and 40 and repeat Steps 1 to 5.

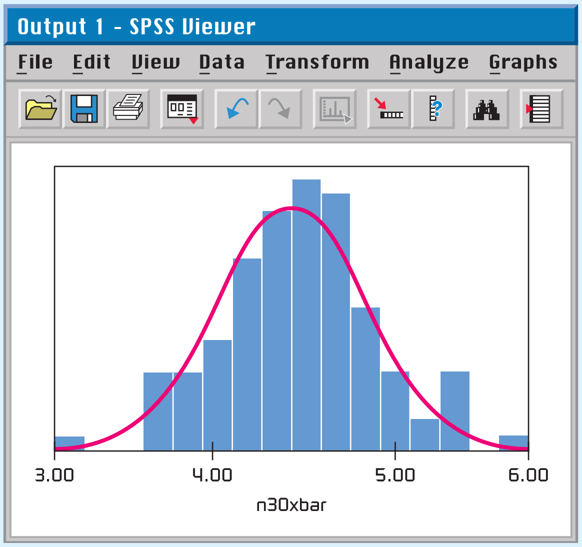
TABLE 7-4 **Frequency Table of Sample Means**

1. Interval	2. Frequency	3. Percent	4. Hypothetical Normal Distribution
$\bar{x} - 3s$ to $\bar{x} - 2s$	Tally the sample means computed in step 2 and place here.	Compute percents from column 2 and place here.	2 or 3%
$\bar{x} - 2s$ to $\bar{x} - s$			13 or 14%
$\bar{x} - s$ to $\bar{x}$			About 34%
$\bar{x}$ to $\bar{x} + s$			About 34%
$\bar{x} + s$ to $\bar{x} + 2s$			13 or 14%
$\bar{x} + 2s$ to $\bar{x} + 3s$			2 or 3%

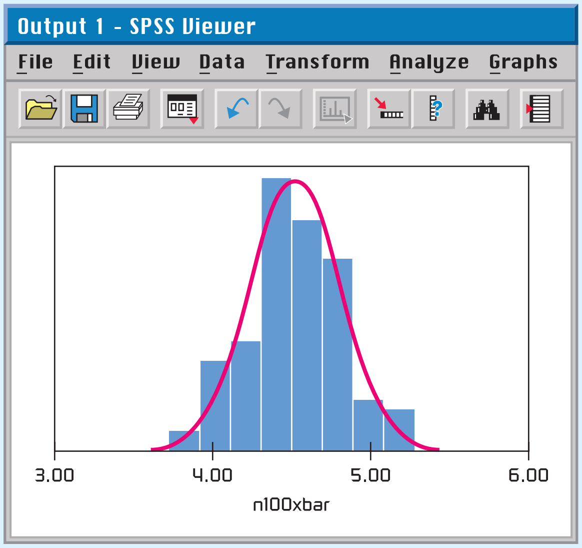
FIGURE 7-9

SPSS-Generated Histograms for Samples of Size 30 and Size 100

(a) $n = 30$



(b) $n = 100$



Technology Hints

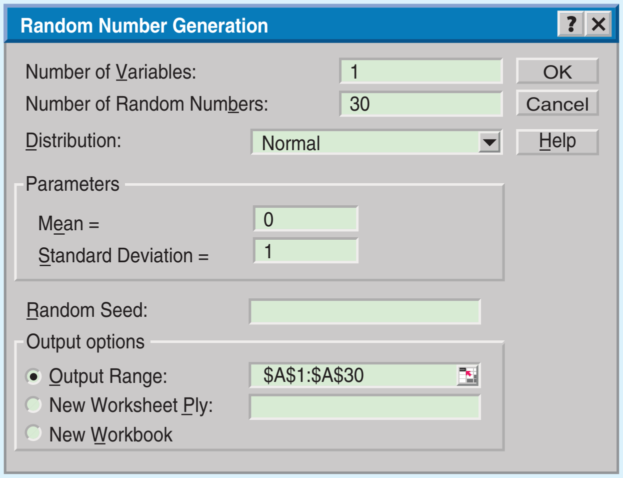
The TI-84Plus and TI-83Plus calculators, Excel, Minitab, and SPSS all support the process of drawing random samples from a variety of distributions. Macros can be written in Excel, Minitab, and the professional version of SPSS to repeat the six steps of the project. Figure 7-9 shows histograms generated by SPSS for random samples of size 30 and size 100. The samples are taken from a uniform probability distribution.

TI-84Plus/TI-83Plus

You can generate random samples from uniform, normal, and binomial distributions. Press **MATH** and select **PRB.** Selection **5:randInt(lower, upper, sample size m)** generates m random integers from the specified interval. Selection **6:randNorm(**μ**,** σ**, sample size m)** generates m random numbers from a normal distribution with mean μ and standard deviation σ. Selection **7:randBin(number of trials n, p, sample size m)** generates m random values (number of successes out of n trials) for a binomial distribution with probability of success p on each trial. You can put these values in lists by using **Edit** under **Stat.** Highlight the list header, press Enter, and then select one of the options discussed.

Excel

Use the menu selection **Tools ➤ Data Analysis ➤ Random Number Generator.** The dialogue box provides choices for the population distribution, including uniform, binomial, and normal distributions. Fill in the required parameters and designate the location for the output.



Minitab

Use the menu selections **Calc ➤ Random Data.** Then select the population distribution. The choices include uniform, binomial, and normal distributions. Fill in the dialogue box, where the number of rows indicates the number of data in the sample.

SPSS

SPSS supports random samples from a variety of distributions, including binomial, normal, and uniform. In data view, generate a column of consecutive integers from 1 to *n*, where *n* is the sample size. In variable view, name the variables sample1, sample2, and so on, through sample30. These variables head the columns containing each of the 30 samples of size *n*. Then use the menu choices **Transform ➤ Compute.** In the dialogue box, use sample 1 as the target variable for the first sample, and so forth. In the function box, select **RV.UNIFORM(min,max)** for samples from a uniform distribution. Functions **RV.NORMAL(mean,stddev)** and **RV.BINOM(n,p)** provide random samples from normal and binomial distributions, respectively.