6

Historical Pictures/Stock Montage

One cannot escape the feeling that these mathematical formulas have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them.

—Heinrich Hertz

How can it be that mathematics, a product of human thought independent of experience, is so admirably adapted to the objects of reality?

—Albert Einstein

For online student resources, visit The Brase/Brase, *Understandable Statistics,* 10th edition web site at **http://www.cengage.com/statistics/brase**.

Heinrich Hertz (1857–1894) was a pioneer in the study of radio waves. His work and the later work of Maxwell and Marconi led the way to modern radio, television, and radar. Albert Einstein is world renowned for his great discoveries in relativity and quantum mechanics. Everyone who has worked in both mathematics and real-world applications cannot help but marvel at how the "pure thought" of the mathematical sciences can predict and explain events in other realms. In this chapter, we will study the most important type of probability distribution in all of mathematical statistics: the normal distribution. Why is the normal distribution so important? Two of the reasons are that it applies to a wide variety of situations and that other distributions tend to become normal under certain conditions.

Normal Curves and Sampling Distributions

Dana White/PhotoEdit

PREVIEW QUESTIONS

What are some characteristics of a normal distribution? What does the empirical rule tell you about data spread around the mean? How can this information be used in quality control? (Section 6.1)

Can you compare apples and oranges, or maybe elephants and butterflies? In most cases, the answer is no—unless you first standardize your measurements. What are a standard normal distribution and a standard z score? (Section 6.2)

How do you convert any normal distribution to a standard normal distribution? How do you find probabilities of "standardized events"? (Section 6.3)

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 6.4)

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 6.5)

The binomial and normal distributions are two of the most important probability distributions in statistics. Under certain limiting conditions, the binomial can be thought to evolve (or envelope) into the normal distribution. How can you apply this concept in the real world? (Section 6.6)

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 6.6)

Syracuse Newspapers/David Lassman/The Image Works Image

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 is 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 free tasting, 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 18 of Section 6.5.)

SECTION 6.1 Graphs of Normal Probability Distributions

FOCUS POINTS

- Graph a normal curve and summarize its important properties.
- Apply the empirical rule to solve real-world problems.
- Use control limits to construct control charts. Examine the chart for three possible out-of-control signals.

Normal distribution

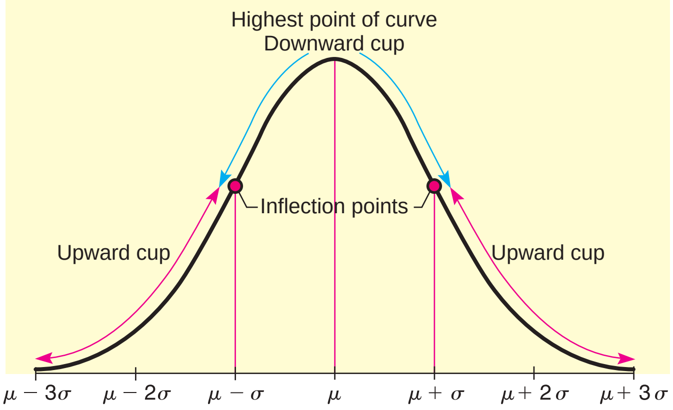
One of the most important examples of a continuous probability distribution is the *normal distribution*. This distribution was studied by the French mathematician Abraham de Moivre (1667–1754) and later by the German mathematician Carl Friedrich Gauss (1777–1855), whose work is so important that the normal distribution is sometimes called *Gaussian*. The work of these mathematicians provided a foundation on which much of the theory of statistical inference is based.

Applications of a normal probability distribution are so numerous that some mathematicians refer to it as "a veritable Boy Scout knife of statistics." However, before we can apply it, we must examine some of the properties of a normal distribution.

A rather complicated formula, presented later in this section, defines a normal distribution in terms of μ and σ, the mean and standard deviation of the population distribution. It is only through this formula that we can verify if a distribution is normal. However, we can look at the graph of a normal

FIGURE 6-1

A Normal Curve



Two other continuous probability distributions are discussed in the section problems. Problems 16 and 17 discuss the *uniform probability distribution*. Problems 18 through 20 discuss the *exponential probability distribution*.

distribution and get a good pictorial idea of some of the essential features of any normal distribution.

Normal curve

The graph of a normal distribution is called a *normal curve*. It possesses a shape very much like the cross section of a pile of dry sand. Because of its shape, blacksmiths would sometimes use a pile of dry sand in the construction of a mold for a bell. Thus the normal curve is also called a *bell-shaped curve* (see Figure 6-1).

We see that a general normal curve is smooth and symmetrical about the vertical line extending upward from the mean μ. Notice that the highest point of the curve occurs over μ. If the distribution were graphed on a piece of sheet metal, cut out, and placed on a knife edge, the balance point would be at μ. We also see that the curve tends to level out and approach the horizontal (x axis) like a glider making a landing. However, in mathematical theory, such a glider would never quite finish its landing because a normal curve never touches the horizontal axis.

Downward cup
Upward cup

The parameter σ controls the spread of the curve. The curve is quite close to the horizontal axis at $\mu + 3\sigma$ and $\mu - 3\sigma$. Thus, if the standard deviation σ is large, the curve will be more spread out; if it is small, the curve will be more peaked. Figure 6-1 shows the normal curve cupped downward for an interval on either side of the mean μ. Then it begins to cup upward as we go to the lower part of the bell. The exact places where the *transition* between the upward and downward cupping occur are above the points $\mu + \sigma$ and $\mu - \sigma$. In the terminology of calculus, transition points such as these are called *inflection points*.

Symmetry of normal curves

Important properties of a normal curve

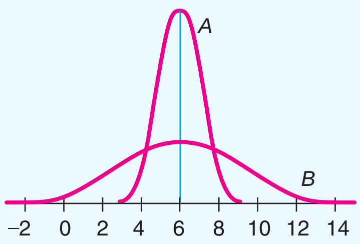
1. The curve is bell-shaped, with the highest point over the mean μ.
2. The curve is symmetrical about a vertical line through μ.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur above $\mu + \sigma$ and $\mu - \sigma$.
5. The area under the entire curve is 1.

The parameters that control the shape of a normal curve are the mean μ and the standard deviation σ. When both μ and σ are specified, a specific normal curve is determined. In brief, μ locates the balance point and σ determines the extent of the spread.

GUIDED EXERCISE 1 Identify μ and σ on a normal curve

Look at the normal curves in Figure 6-2.

FIGURE 6-2



(a) Do these distributions have the same mean? If so, what is it?

→ The means are the same, since both graphs have the high point over 6. $\mu = 6$.

(b) One of the curves corresponds to a normal distribution with $\sigma = 3$ and the other to one with $\sigma = 1$. Which curve has which σ?

→ Curve A has $\sigma = 1$ and curve B has $\sigma = 3$. (Since curve B is more spread out, it has the larger σ value.)

COMMENT The normal distribution curve is always above the horizontal axis. The area beneath the curve and above the axis is exactly 1. As such, the normal distribution curve is an example of a *density curve*. The formula used to generate the shape of the normal distribution curve is called the *normal density function*. If x is a normal random variable with mean μ and standard deviation σ, the formula for the normal density function is

$$f(x) = \frac{e^{(-1/2)((x-\mu)/\sigma)^2}}{\sigma\sqrt{2\pi}}$$

In this text, we will not use this formula explicitly. However, we will use tables of areas based on the normal density function.

The total area under any normal curve studied in this book will *always* be 1. The graph of the normal distribution is important because the portion of the *area* under the curve above a given interval represents the *probability* that a measurement will lie in that interval.

Empirical rule

In Section 3.2, we studied Chebyshev's theorem. This theorem gives us information about the *smallest* proportion of data that lies within 2, 3, or k standard deviations of the mean. This result applies to *any* distribution. However, for normal distributions, we can get a much more precise result, which is given by the *empirical rule*.

Empirical rule

For a distribution that is symmetrical and bell-shaped (in particular, for a normal distribution):

Approximately 68% of the data values will lie within 1 standard deviation on each side of the mean.

Approximately 95% of the data values will lie within 2 standard deviations on each side of the mean.

Approximately 99.7% (or almost all) of the data values will lie within 3 standard deviations on each side of the mean.

The preceding statement is called the *empirical rule* because, for symmetrical, bell-shaped distributions, the given percentages are observed in practice. Furthermore, for the normal distribution, the empirical rule is a direct consequence of the very nature of the distribution (see Figure 6-3). Notice that the empirical rule is a stronger statement than Chebyshev's theorem in that it gives *definite percentages,* not just lower limits. Of course, the empirical rule applies only to normal or symmetrical, bell-shaped distributions, whereas Chebyshev's theorem applies to all distributions.

FIGURE 6-3

Area Under a Normal Curve

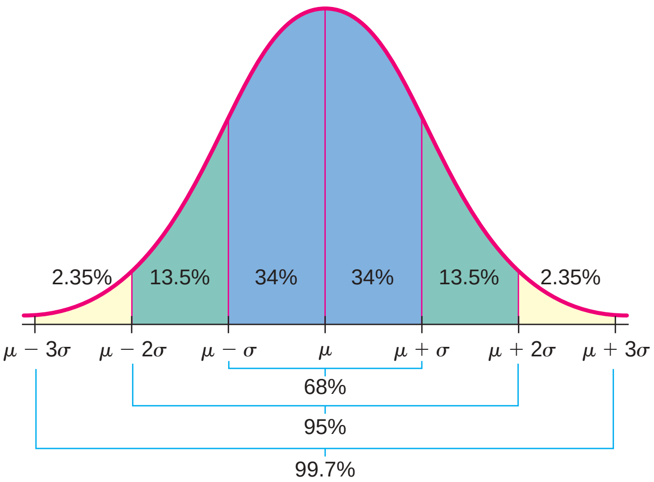
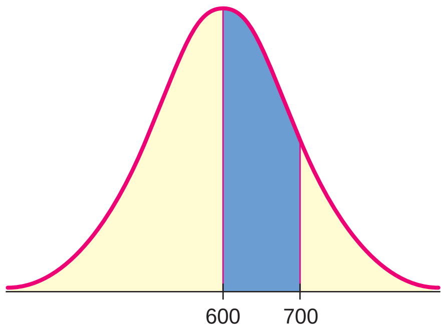


FIGURE 6-4

Distribution of Playing Times



EXAMPLE 1 EMPIRICAL RULE

The playing life of a Sunshine radio is normally distributed with mean $\mu = 600$ hours and standard deviation $\sigma = 100$ hours. What is the probability that a radio selected at random will last from 600 to 700 hours?

SOLUTION: The probability that the playing life will be between 600 and 700 hours is equal to the percentage of the total area under the curve that is shaded in Figure 6-4. Since $\mu = 600$ and $\mu + \sigma = 600 + 100 = 700$, we see that the shaded area is simply the area between μ and $\mu + \sigma$. The area from μ to $\mu + \sigma$ is 34% of the total area. This tells us that the probability a Sunshine radio will last between 600 and 700 playing hours is about 0.34.

GUIDED EXERCISE 2 *Empirical rule*

The yearly wheat yield per acre on a particular farm is normally distributed with mean $\mu = 35$ bushels and standard deviation $\sigma = 8$ bushels.

(a) Shade the area under the curve in Figure 6-5 (next page) that represents the probability that an acre will yield between 19 and 35 bushels. → See Figure 6-6.

(b) Is the area the same as the area between $\mu - 2\sigma$ and μ? → Yes, since $\mu = 35$ and $\mu - 2\sigma = 35 - 2(8) = 19$.

Continued

GUIDED EXERCISE 2 *continued*

FIGURE 6-5

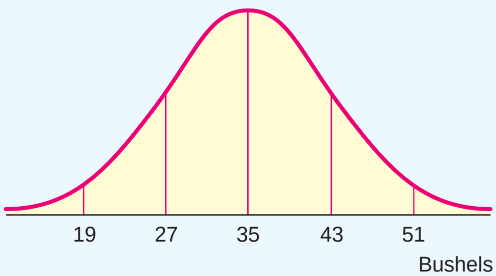
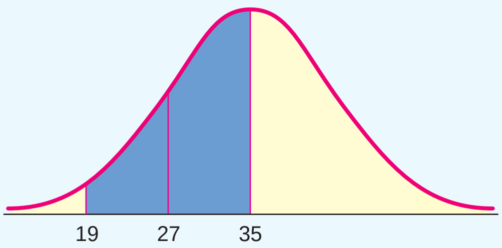


FIGURE 6-6 Completion of Figure 6-5



(c) Use Figure 6-3 to find the percentage of area over the interval between 19 and 35.

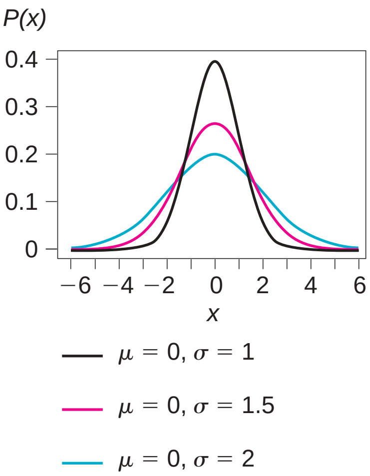
→ The area between the values $\mu - 2\sigma$ and μ is 47.5% of the total area.

(d) What is the probability that the yield will be between 19 and 35 bushels per acre?

→ It is 47.5% of the total area, which is 1. Therefore, the probability is 0.475 that the yield will be between 19 and 35 bushels.

TECH NOTES

We can graph normal distributions using the TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab. In each technology, set the range of x values between $\mu - 3.5\sigma$ and $\mu + 3.5\sigma$. Then use the built-in normal density functions to generate the corresponding y values.



TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Press the Y = key. Then, under **DISTR**, select **1:normalpdf** (x,μ,σ) and fill in desired μ and σ values. Press the **WINDOW** key. Set **Xmin** to $\mu - 3\sigma$ and **Xmax** to $\mu + 3\sigma$. Finally, press the **ZOOM** key and select option **0:ZoomFit.**

Excel 2007 In one column, enter x values from $\mu - 3.5\sigma$ to $\mu + 3.5\sigma$ in increments of 0.2σ. In the next column, generate y values by using the ribbon choices **Insert function** f_x

In the dialogue box, select **Statistical** for the Category and then for the Function, select **NORMDIST** (x,μ,σ,false). Next click the **Insert** tab and in the **Charts** group, click **Scatter**. Select the scatter diagram with smooth lines (third grouping).

Minitab In one column, enter x values from -3.5σ to 3.5σ in increments of 0.2σ. In the next column, enter y values by using the menu choices **Calc ➤ Probability Distribution ➤ Normal.** Fill in the dialogue box. Next, use menu choices **Graph ➤ Plot.** Fill in the dialogue box. Under Display, select connect.

LOOKING FORWARD

Normal probability distributions will be used extensively in our later work. For instance, when we repeatedly take samples of the same size from a distribution and compute the same mean for each sample, we'll find that the sample means follow a distribution that is normal or approximately normal (Section 6.5). Also, when the number of trials is sufficiently large, the binomial distribution can be approximated by a normal distribution (Section 6.6). The distribution of the sample proportion of successes in a fixed number of binomial trials also can be approximated by a normal distribution (Section 6.6).

Control Charts

Control charts

If we are examining data over a period of equally spaced time intervals or in some sequential order, then *control charts* are especially useful. Business managers and people in charge of production processes are aware that there exists an inherent amount of variability in any sequential set of data. The sugar content of bottled drinks taken sequentially off a production line, the extent of clerical errors in a bank from day to day, advertising expenses from month to month, or even the number of new customers from year to year are examples of sequential data. There is a certain amount of variability in each.

A random variable x is said to be in *statistical control* if it can be described by the *same* probability distribution when it is observed at successive points in time. Control charts combine graphic and numerical descriptions of data with probability distributions.

Control charts were invented in the 1920s by Walter Shewhart at Bell Telephone Laboratories. Since a control chart is a *warning device,* it is not absolutely necessary that our assumptions and probability calculations be precisely correct. For example, the x distributions need not follow a normal distribution exactly. Any mound-shaped and more or less symmetrical distribution will be good enough.

PROCEDURE

How to make a control chart for the random variable x

A control chart for a random variable x is a plot of observed x values in time sequence order.

1. Find the mean μ and standard deviation σ of the x distribution by
   (a) using past data from a period during which the process was "in control" or
   (b) using specified "target" values for μ and σ.
2. Create a graph in which the vertical axis represents x values and the horizontal axis represents time.
3. Draw a horizontal line at height μ and horizontal, dashed control-limit lines at $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$.
4. Plot the variable x on the graph in time sequence order. Use line segments to connect the points in time sequence order.

How do we pick values for μ and σ? In most practical cases, values for μ (population mean) and σ (population standard deviation) are computed from past data for which the process we are studying was known to be *in control*. Methods for choosing the sample size to fit given error tolerances can be found in Chapter 7.

Sometimes values for μ and σ are chosen as *target values*. That is, μ and σ values are chosen as set goals or targets that reflect the production level or service level at which a company hopes to perform. To be realistic, such target assignments for μ and σ should be reasonably close to actual data taken when the process was operating at a satisfactory production level. In Example 2, we will make a control chart; then we will discuss ways to analyze it to see if a process or service is "in control."

EXAMPLE 2

Control chart

Susan Tamara is director of personnel at the Antlers Lodge in Denali National Park, Alaska. Every summer Ms. Tamara hires many part-time employees from all over the United States. Most are college students seeking summer employment. One of the biggest activities for the lodge staff is that of "making up" the rooms each day. Although the rooms are supposed to be ready by 3:30 P.M., there are always some rooms not made up by this time because of high personnel turnover.

Every 15 days Ms. Tamara has a general staff meeting at which she shows a control chart of the number of rooms not made up by 3:30 P.M. each day. From extensive experience, Ms. Tamara is aware that the distribution of rooms not made up by 3:30 P.M. is approximately normal, with mean $\mu = 19.3$ rooms and standard deviation $\sigma = 4.7$ rooms. This distribution of x values is acceptable to the top administration of Antlers Lodge. For the past 15 days, the housekeeping unit has reported the number of rooms not ready by 3:30 P.M. (Table 6-1). Make a control chart for these data.

Mt. McKinley, Denali National Park

TABLE 6-1 Number of Rooms x Not Made Up by 3:30 P.M.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	11	20	25	23	16	19	8	25	17	20	23	29	18	14	10

SOLUTION: A control chart for a variable x is a plot of the observed x values (vertical scale) in time sequence order (the horizontal scale represents time). Place horizontal lines at

the mean $\mu = 19.3$

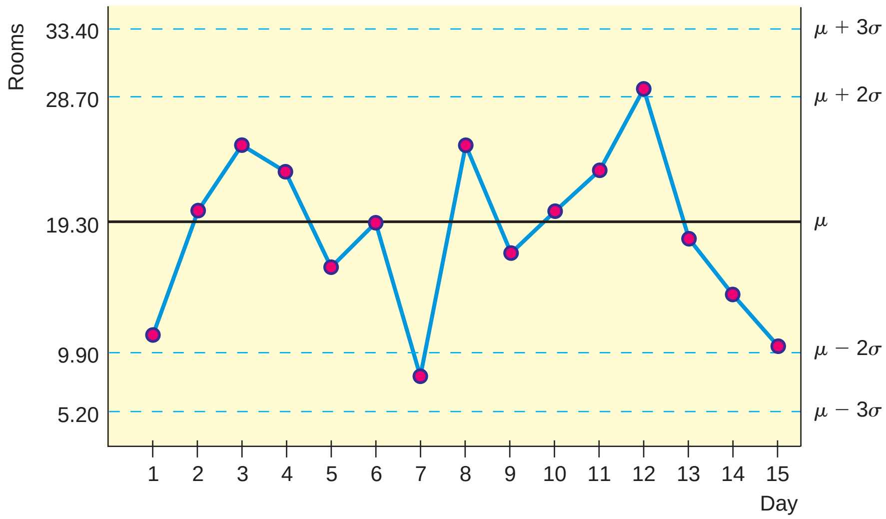
the control limits $\mu \pm 2\sigma = 19.3 \pm 2(4.7)$, or 9.90 and 28.70

the control limits $\mu \pm 3\sigma = 19.3 \pm 3(4.7)$, or 5.20 and 33.40

Then plot the data from Table 6-1. (See Figure 6-7.)

FIGURE 6-7

Number of Rooms Not Made Up by 3:30 P.M.



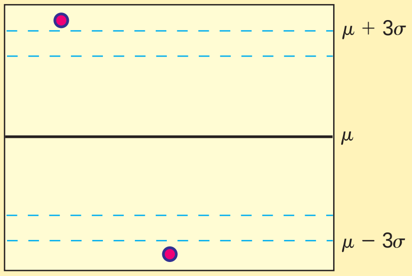
Once we have made a control chart, the main question is the following: As time goes on, is the x variable continuing in this same distribution, or is the distribution of x values changing? If the x distribution is continuing in more or less the same manner, we say it is *in statistical control*. If it is not, we say it is *out of control*.

Out-of-control warning signals

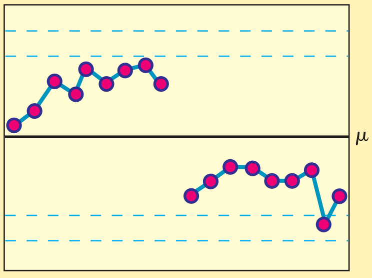
Many popular methods can set off a warning signal that a process is out of control. Remember, a random variable x is said to be *out of control* if successive time measurements of x indicate that it is no longer following the target probability distribution. We will assume that the target distribution is (approximately) normal and has (user-set) target values for μ and σ.

Three of the most popular warning signals are described next.

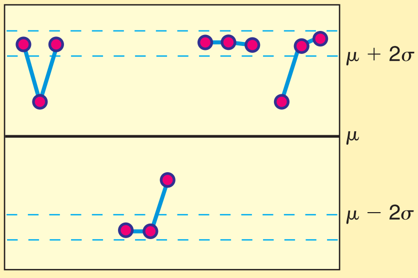
Out-of-Control Signal I



Out-of-Control Signal II



Out-of-Control Signal III



Out-of-control signals

1. **Out-of-Control Signal I: One point falls beyond the 3σ level**
   What is the probability that signal I will be a false alarm? By the empirical rule, the probability that a point lies within 3σ of the mean is 0.997. The probability that signal I will be a false alarm is $1 - 0.997 = 0.003$. Remember, a false alarm means that the x distribution is really on the target distribution, and we simply have a very rare (probability of 0.003) event.
2. **Out-of-Control Signal II: A run of nine consecutive points on one side of the center line (the line at target value μ)**
   To find the probability that signal II is a false alarm, we observe that if the x distribution and the target distribution are the same, then there is a 50% chance that the x values will lie above or below the center line at μ. Because the samples are (time) independent, the probability of a run of nine points on one side of the center line is $(0.5)^9 = 0.002$. If we consider both sides, this probability becomes 0.004. Therefore, the probability that signal II is a false alarm is approximately 0.004.
3. **Out-of-Control Signal III: At least two of three consecutive points lie beyond the 2σ level on the same side of the center line**
   To determine the probability that signal III will produce a false alarm, we use the empirical rule. By this rule, the probability that an x value will be above the 2σ level is about 0.023. Using the binomial probability distribution (with success being the point is above 2σ), the probability of two or more successes out of three trials is

$$\frac{3!}{2!1!}(0.023)^2(0.997) + \frac{3!}{3!0!}(0.023)^3 \approx 0.002$$

   Taking into account *both* above and below the center line, it follows that the probability that signal III is a false alarm is about 0.004.

Remember, a control chart is only a warning device, and it is possible to get a false alarm. A false alarm happens when one (or more) of the out-of-control signals occurs, but the x distribution is really on the target or assigned distribution. In this case, we simply have a rare event (probability of 0.003 or 0.004). In practice, whenever a control chart indicates that a process is out of control, it is usually a good precaution to examine what is going on. If the process is out of control, corrective steps can be taken before things get a lot worse. The rare false alarm is a small price to pay if we can avert what might become real trouble.

Type of Warning Signal	Probability of a False Alarm
Type I: Point beyond 3σ	0.003
Type II: Run of nine consecutive points, all below center line μ or all above center line μ	0.004
Type III: At least two out of three consecutive points beyond 2σ	0.004

From an intuitive point of view, signal I can be thought of as a blowup, something dramatically out of control. Signal II can be thought of as a slow drift out of control. Signal III is somewhere between a blowup and a slow drift.

EXAMPLE 3 INTERPRETING A CONTROL CHART

Ms. Tamara of the Antlers Lodge examines the control chart for housekeeping. During the staff meetings, she makes recommendations about improving service or, if all is going well, she gives her staff a well-deserved "pat on the back." Look at the control chart created in Example 2 (Figure 6-7 on page 256) to determine if the housekeeping process is out of control.

SOLUTION: The x values are more or less evenly distributed about the mean $\mu = 19.3$. None of the points are outside the $\mu \pm 3\sigma$ limit (i.e., above 33.40 or below 5.20 rooms). There is no run of nine consecutive points above or below μ. No two of three consecutive points are beyond the $\mu \pm 2\sigma$ limit (i.e., above 28.7 or below 9.90 rooms).

It appears that the x distribution is "in control." At the staff meeting, Ms. Tamara should tell her employees that they are doing a reasonably good job and they should keep up the fine work!

GUIDED EXERCISE 3 *Interpreting a control chart*

Figures 6-8 and 6-9 show control charts of housekeeping reports for two other 15-day periods.

FIGURE 6-8 Report II

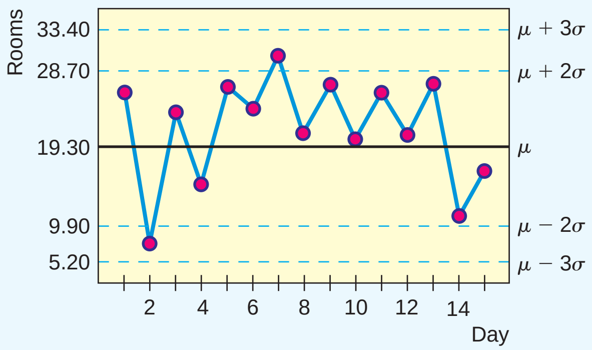
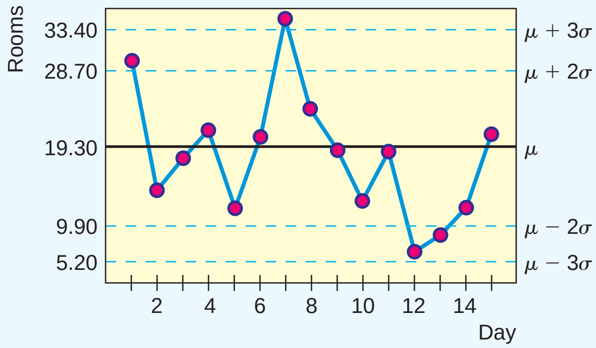


FIGURE 6-9 Report III



(a) *Interpret* the control chart in Figure 6-8.

Days 5 to 13 are above $\mu = 19.3$. We have nine consecutive days on one side of the mean. This is a warning signal! It would appear that the mean μ is slowly drifting up beyond the target value of 19.3. The chart indicates that housekeeping is "out of control." Ms. Tamara should take corrective measures at her staff meeting.

(b) *Interpret* the control chart in Figure 6-9.

On day 7, we have a data value beyond $\mu + 3\sigma$ (i.e., above 33.40). On days 11, 12, and 13, we have two of three data values beyond $\mu - 2\sigma$ (i.e., below 9.90). The occurrences during both of these periods are out-of-control warning signals. Ms. Tamara might ask her staff about both of these periods. There may be a lesson to be learned from day 7, when housekeeping apparently had a lot of trouble. Also, days 11, 12, and 13 were very good days. Perhaps a lesson could be learned about why things went so well.

COMMENT Uniform Distribution and Exponential Distribution

Normal distributions are the central topic of this chapter. Normal distributions are very important in general probability and statistics. However, there are other, more specialized distributions that also have many applications. Two such distributions are the *uniform distribution* and the *exponential distribution*. To learn more about these distributions, please see Problems 16 to 20 at the end of this section.

VIEWPOINT In Control? Out of Control?

If you care about quality, you also must care about control! Dr. Walter Shewhart invented control charts when he was working for Bell Laboratories. The great contribution of control charts is that they separate variation into two sources: (1) random or chance causes (in control) and (2) special or assignable causes (out of control). A process is said to be in statistical control *when it is no longer afflicted with special or assignable causes. The performance of a process that is in statistical control is predictable. Predictability and quality control tend to be closely associated.*

(Source: Adapted from the classic text *Statistical Methods from the Viewpoint of Quality Control,* by W. A. Shewhart, with foreword by W. E. Deming, Dover Publications.)

SECTION 6.1 PROBLEMS

1. *Statistical Literacy* Which, if any, of the curves in Figure 6-10 look(s) like a normal curve? If a curve is not a normal curve, tell why.

2. *Statistical Literacy* Look at the normal curve in Figure 6-11, and find μ, $\mu + \sigma$, and σ.

FIGURE 6-10

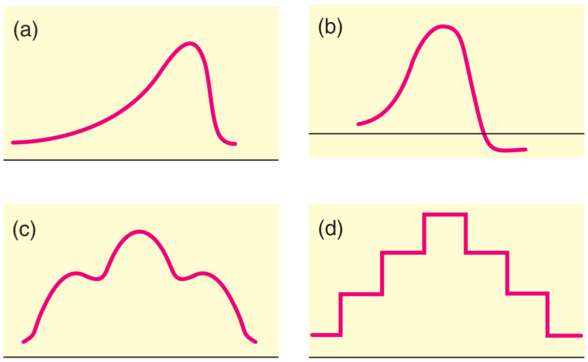
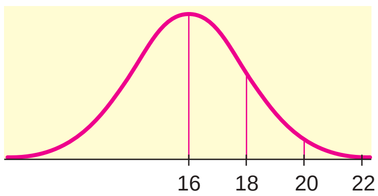


FIGURE 6-11



3. *Critical Thinking* Look at the two normal curves in Figures 6-12 and 6-13. Which has the larger standard deviation? What is the mean of the curve in Figure 6-12? What is the mean of the curve in Figure 6-13?

FIGURE 6-13

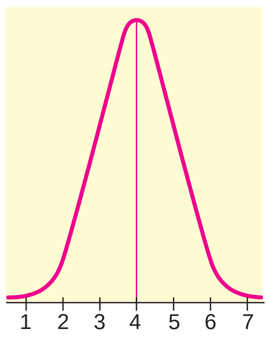
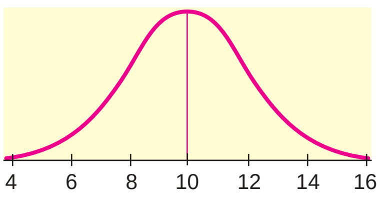


FIGURE 6-12



4. ***Critical Thinking*** Sketch a normal curve
   (a) with mean 15 and standard deviation 2.
   (b) with mean 15 and standard deviation 3.
   (c) with mean 12 and standard deviation 2.
   (d) with mean 12 and standard deviation 3.
   (e) Consider two normal curves. If the first one has a larger mean than the second one, must it have a larger standard deviation as well? Explain your answer.

5. ***Basic Computation: Empirical Rule*** What percentage of the area under the normal curve lies
   (a) to the left of μ?
   (b) between $\mu - \sigma$ and $\mu + \sigma$?
   (c) between $\mu - 3\sigma$ and $\mu + 3\sigma$?

6. ***Basic Computation: Empirical Rule*** What percentage of the area under the normal curve lies
   (a) to the right of μ?
   (b) between $\mu - 2\sigma$ and $\mu + 2\sigma$?
   (c) to the right of $\mu + 3\sigma$?

7. ***Distribution: Heights of Coeds*** Assuming that the heights of college women are normally distributed with mean 65 inches and standard deviation 2.5 inches (based on information from *Statistical Abstract of the United States,* 112th Edition), answer the following questions. *Hint:* Use Problems 5 and 6 and Figure 6-3.
   (a) What percentage of women are taller than 65 inches?
   (b) What percentage of women are shorter than 65 inches?
   (c) What percentage of women are between 62.5 inches and 67.5 inches?
   (d) What percentage of women are between 60 inches and 70 inches?

8. ***Distribution: Rhode Island Red Chicks*** The incubation time for Rhode Island Red chicks is normally distributed with a mean of 21 days and standard deviation of approximately 1 day (based on information from *World Book Encyclopedia*). Look at Figure 6-3 and answer the following questions. If 1000 eggs are being incubated, how many chicks do we expect will hatch
   (a) in 19 to 23 days?
   (b) in 20 to 22 days?
   (c) in 21 days or fewer?
   (d) in 18 to 24 days? (Assume all eggs eventually hatch.)

   Note: In this problem, let us agree to think of a single day or a succession of days as a continuous interval of time.

9. ***Archaeology: Tree Rings*** At Burnt Mesa Pueblo, archaeological studies have used the method of tree-ring dating in an effort to determine when prehistoric people lived in the pueblo. Wood from several excavations gave a mean of (year) 1243 with a standard deviation of 36 years (*Bandelier Archaeological Excavation Project: Summer 1989 Excavations at Burnt Mesa Pueblo,* edited by Kohler, Washington State University Department of Anthropology). The distribution of dates was more or less mound-shaped and symmetrical about the mean. Use the empirical rule to
   (a) estimate a range of years centered about the mean in which about 68% of the data (tree-ring dates) will be found.
   (b) estimate a range of years centered about the mean in which about 95% of the data (tree-ring dates) will be found.
   (c) estimate a range of years centered about the mean in which almost all the data (tree-ring dates) will be found.

Raul Touzon/National Geographic Stock

10. ***Vending Machine: Soft Drinks*** A vending machine automatically pours soft drinks into cups. The amount of soft drink dispensed into a cup is normally

distributed with a mean of 7.6 ounces and standard deviation of 0.4 ounce. Examine Figure 6-3 and answer the following questions.

(a) Estimate the probability that the machine will overflow an 8-ounce cup.
(b) Estimate the probability that the machine will not overflow an 8-ounce cup.
(c) The machine has just been loaded with 850 cups. How many of these do you expect will overflow when served?

11. ***Pain Management: Laser Therapy*** "Effect of Helium-Neon Laser Auriculotherapy on Experimental Pain Threshold" is the title of an article in the journal *Physical Therapy* (Vol. 70, No. 1, pp. 24–30). In this article, laser therapy was discussed as a useful alternative to drugs in pain management of chronically ill patients. To measure pain threshold, a machine was used that delivered low-voltage direct current to different parts of the body (wrist, neck, and back). The machine measured current in milliamperes (mA). The pretreatment experimental group in the study had an average threshold of pain (pain was first detectable) at $\mu = 3.15$ mA with standard deviation $\sigma = 1.45$ mA. Assume that the distribution of threshold pain, measured in milliamperes, is symmetrical and more or less mound-shaped. Use the empirical rule to
(a) estimate a range of milliamperes centered about the mean in which about 68% of the experimental group had a threshold of pain.
(b) estimate a range of milliamperes centered about the mean in which about 95% of the experimental group had a threshold of pain.

12. ***Control Charts: Yellowstone National Park*** Yellowstone Park Medical Services (YPMS) provides emergency health care for park visitors. Such health care includes treatment for everything from indigestion and sunburn to more serious injuries. A recent issue of *Yellowstone Today* (National Park Service Publication) indicated that the average number of visitors treated each day by YPMS is 21.7. The estimated standard deviation is 4.2 (summer data). The distribution of numbers treated is approximately mound-shaped and symmetrical.
(a) For a 10-day summer period, the following data show the number of visitors treated each day by YPMS:

Day	1	2	3	4	5	6	7	8	9	10
Number treated	25	19	17	15	20	24	30	19	16	23

Make a control chart for the daily number of visitors treated by YPMS, and plot the data on the control chart. Do the data indicate that the number of visitors treated by YPMS is "in control"? Explain your answer.
(b) For another 10-day summer period, the following data were obtained:

Day	1	2	3	4	5	6	7	8	9	10
Number treated	20	15	12	21	24	28	32	36	35	37

Make a control chart, and plot the data on the chart. ***Interpretation*** Do the data indicate that the number of visitors treated by YPMS is "in control" or "out of control"? Explain your answer. Identify all out-of-control signals by type (I, II, or III). If you were the park superintendent, do you think YPMS might need some (temporary) extra help? Explain.

13. ***Control Charts: Bank Loans*** Tri-County Bank is a small independent bank in central Wyoming. This is a rural bank that makes loans on items as small as horses and pickup trucks to items as large as ranch land. Total monthly loan requests are used by bank officials as an indicator of economic business conditions in this rural community. The mean monthly loan request for the past several years has been 615.1 (in thousands of dollars) with a standard deviation of 11.2 (in thousands of dollars). The distribution of loan requests is approximately mound-shaped and symmetrical.


WendellFranks/iStockphoto.com

(a) For 12 months, the following monthly loan requests (in thousands of dollars) were made to Tri-County Bank:

Month	1	2	3	4	5	6
Loan request	619.3	625.1	610.2	614.2	630.4	615.9

Month	7	8	9	10	11	12
Loan request	617.2	610.1	592.7	596.4	585.1	588.2

Make a control chart for the total monthly loan requests, and plot the preceding data on the control chart. *Interpretation* From the control chart, would you say the local business economy is heating up or cooling down? Explain your answer by referring to any trend you may see on the control chart. Identify all out-of-control signals by type (I, II, or III).

(b) For another 12-month period, the following monthly loan requests (in thousands of dollars) were made to Tri-County Bank:

Month	1	2	3	4	5	6
Loan request	608.3	610.4	615.1	617.2	619.3	622.1

Month	7	8	9	10	11	12
Loan request	625.7	633.1	635.4	625.0	628.2	619.8

Make a control chart for the total monthly loan requests, and plot the preceding data on the control chart. *Interpretation* From the control chart, would you say the local business economy is heating up, cooling down, or about normal? Explain your answer by referring to the control chart. Identify all out-of-control signals by type (I, II, or III).

14. *Control Charts: Motel Rooms* The manager of Motel 11 has 316 rooms in Palo Alto, California. From observation over a long period of time, she knows that on an average night, 268 rooms will be rented. The long-term standard deviation is 12 rooms. This distribution is approximately mound-shaped and symmetrical.

(a) For 10 consecutive nights, the following numbers of rooms were rented each night:

Night	1	2	3	4	5	6
Number of rooms	234	258	265	271	283	267

Night	7	8	9	10
Number of rooms	290	286	263	240

Make a control chart for the number of rooms rented each night, and plot the preceding data on the control chart. *Interpretation* Looking at the control chart, would you say the number of rooms rented during this 10-night period has been unusually low? unusually high? about what you expected? Explain your answer. Identify all out-of-control signals by type (I, II, or III).

(b) For another 10 consecutive nights, the following numbers of rooms were rented each night:

Night	1	2	3	4	5	6
Number of rooms	238	245	261	269	273	250

Night	7	8	9	10
Number of rooms	241	230	215	217

Make a control chart for the number of rooms rented each night, and plot the preceding data on the control chart. ***Interpretation*** Would you say the room occupancy has been high? low? about what you expected? Explain your answer. Identify all out-of-control signals by type (I, II, or III).

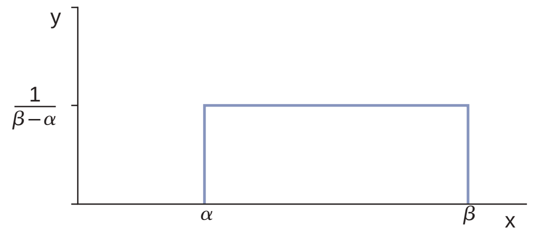
15. ***Control Chart: Air Pollution*** The visibility standard index (VSI) is a measure of Denver air pollution that is reported each day in the *Rocky Mountain News*. The index ranges from 0 (excellent air quality) to 200 (very bad air quality). During winter months, when air pollution is higher, the index has a mean of about 90 (rated as fair) with a standard deviation of approximately 30. Suppose that for 15 days, the following VSI measures were reported each day:

Day	1	2	3	4	5	6	7	8	9
VSI	80	115	100	90	15	10	53	75	80

Day	10	11	12	13	14	15
VSI	110	165	160	120	140	195

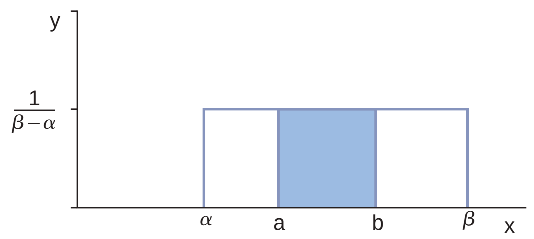
Make a control chart for the VSI, and plot the preceding data on the control chart. Identify all out-of-control signals (high or low) that you find in the control chart by type (I, II, or III).

16. ***Expand Your Knowledge: Continuous Uniform Probability Distribution*** Let α and β be any two constants such that $\alpha < \beta$. Suppose we choose a point x at random in the interval from α to β. In this context the phrase *at random* is taken to mean that the point x is as likely to be chosen from one particular part of the interval as any other part. Consider the rectangle.



The base of the rectangle has length $\beta - \alpha$ and the height of the rectangle is $1/(\beta - \alpha)$, so the area of the rectangle is 1. As such, this rectangle's top can be thought of as part of a probability density curve. Since we specify that x must lie between α and β, the probability of a point occurring outside the interval $[\alpha, \beta]$ is, by definition, 0. From a geometric point of view, x chosen at random from α to β means we are equally likely to land anywhere in the interval from α to β. For this reason, the top of the (rectangle's) density curve is flat or uniform.

Now suppose that a and b are numbers such that $\alpha \le a < b \le \beta$. What is the probability that a number x chosen at random from α to β will fall in the interval $[a, b]$? Consider the graph



Because x is chosen at random from $[\alpha, \beta]$, the area of the rectangle that lies above $[a, b]$ is the probability that x lies in $[a, b]$. This area is

$$P(a < x < b) = \frac{b - a}{\beta - \alpha}$$

In this way we can assign a probability to any interval inside $[\alpha, \beta]$. This probability distribution is called the *continuous uniform distribution* (also called the rectangular distribution). Using some extra mathematics, it can be shown that if x is a random variable with this distribution, then the mean and standard deviation of x are

$$\mu = \frac{\alpha + \beta}{2} \quad \text{and} \quad \sigma = \frac{\beta - \alpha}{\sqrt{12}}$$

Sedimentation experiments are very important in the study of biology, medicine, hydrodynamics, petroleum engineering, civil engineering, and so on. The size (diameter) of approximately spherical particles is important since larger particles hinder and sometimes block the movement of smaller particles. Usually the size of sediment particles follows a uniform distribution (Reference: Y. Zimmels, "Theory of Kindred Sedimentation of Polydisperse Mixtures," *AIChE Journal*, Vol. 29, No. 4, pp. 669–676).

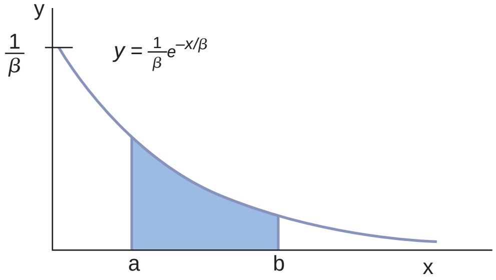
Suppose a veterinary science experiment injects very small, spherical pellets of low-level radiation directly into an animal's bloodstream. The purpose is to attempt to cure a form of recurring cancer. The pellets eventually dissolve and pass through the animal's system. Diameters of the pellets are uniformly distributed from 0.015 mm to 0.065 mm. If a pellet enters an artery, what is the probability that it will be the following sizes?

(a) 0.050 mm or larger. *Hint*: All particles are between 0.015 mm and 0.065 mm, so larger than 0.050 means $0.050 \le x \le 0.065$.
(b) 0.040 mm or smaller
(c) between 0.035 mm and 0.055 mm
(d) Compute the mean size of the particles.
(e) Compute the standard deviation of particle size.

17. ***Uniform Distribution: Measurement Errors*** Measurement errors from instruments are often modeled using the uniform distribution (see Problem 16). To determine the range of a large public address system, acoustical engineers use a method of triangulation to measure the shock waves sent out by the speakers. The time at which the waves arrive at the sensors must be measured accurately. In this context, a negative error means the signal arrived too early. A positive error means the signal arrived too late. Measurement errors in reading these times have a uniform distribution from -0.05 to $+0.05$ microseconds. (Reference: J. Perruzzi, and E. Hilliard, "Modeling Time Delay Measurement Errors," *Journal of the Acoustical Society of America*, Vol. 75, No. 1, pp. 197–201.) What is the probability that such measurements will be in error by
(a) less than $+0.03$ microsecond (i.e., $-0.05 \le x < 0.03$)?
(b) more than -0.02 microsecond?
(c) between -0.04 and $+0.01$ microsecond?
(d) Find the mean and standard deviation of measurement errors. Measurements from an instrument are called *unbiased* if the mean of the measurement errors is zero. Would you say the measurements for these acoustical sensors are unbiased? Explain.

18. ***Expand Your Knowledge: Exponential Distribution*** The Poisson distribution (Section 5.4) gives the probability for the *number of occurrences* for a "rare" event. Now, let x be a random variable that represents the *waiting time* between rare events. Using some mathematics, it can be shown that x has an *exponential distribution*. Let $x > 0$ be a random variable and let $\beta > 0$ be a constant. Then

$y = \frac{1}{\beta} e^{-x/\beta}$ is a curve representing the exponential distribution. Areas under this curve give us exponential probabilities.



If a and b are any numbers such that $0 < a < b$, then using some extra mathematics, it can be shown that the area under the curve above the interval $[a, b]$ is

$$P(a < x < b) = e^{-a/\beta} - e^{-b/\beta}$$

Notice that by definition, x cannot be negative, so, $P(x < 0) = 0$. The random variable x is called an *exponential random variable*. Using some more mathematics, it can be shown that the mean and standard deviation of x are

$$\mu = \beta \quad \text{and} \quad \sigma = \beta$$

Note: The number $e = 2.71828 \ldots$ is used throughout probability, statistics, and mathematics. The key e^x is conveniently located on most calculators.

Comment: The Poisson and exponential distributions have a special relationship. Specifically, it can be shown that the *waiting time* between successive Poisson arrivals (i.e., successes or rare events) has an exponential distribution with $\beta = 1/\lambda$, where λ is the average number of Poisson successes (rare events) per unit of time. For more on this topic, please see Problem 20.)

Fatal accidents on scheduled domestic passenger flights are rare events. In fact, airlines do all they possibly can to prevent such accidents. However, around the world such fatal accidents do occur. Let x be a random variable representing the waiting time between fatal airline accidents. Research has shown that x has an exponential distribution with a mean of approximately 44 days (Reference: R. Pyke, "Spacings," *Journal of the Royal Statistical Society* B, Vol. 27, No. 3, p. 426.)

We take the point of view that x (measured in days as units) is a continuous random variable. Suppose a fatal airline accident has just been reported on the news. What is the probability that the waiting time to the next reported fatal airline accident is

(a) less than 30 days (i.e., $0 \le x < 30$)?
(b) more than 50 days (i.e., $50 < x < \infty$)? *Hint:* $e^{-\infty} = 0$.
(c) between 20 and 60 days?
(d) What are the mean and the standard deviation of the waiting times x?

19. ***Exponential Distribution: Supply and Demand*** Another application for exponential distributions (see Problem 18) is supply/demand problems. The operator of a pumping station in a small Wyoming town has observed that demand for water on a typical summer afternoon is exponentially distributed with a mean of 75 cfs (cubic feet per second). Let x be a random variable that represents the town's demand for water (in cfs). What is the probability that on a typical summer afternoon, this town will have a water demand x

blueflames/iStockphoto.com

(a) more than 60 cfs (i.e., $60 < x < \infty$)? *Hint:* $e^{-\infty} = 0$.
(b) less than 140 cfs (i.e., $0 < x < 140$)?
(c) between 60 and 100 cfs?
(d) ***Brain teaser*** How much water c (in cfs) should the station pump to be 80% sure that the town demand x (in cfs) will not exceed the supply c? *Hint:* First explain why the equation $P(0 < x < c) = 0.80$ represents the problem as stated. Then solve for c.

20. *Exponential Distribution: Waiting Time Between Poisson Events* A common application of exponential distributions (see Problem 18) is waiting time between Poisson events (e.g., successes in the Poisson distribution; see Section 5.4). In our study of the Poisson distribution (Example 11, Section 5.4), we saw that the mean success rate per hour of catching a fish at Pyramid Lake is $\lambda = 0.667$ fish/hour. From this we see that the mean waiting time between fish can be thought of as $\beta = 1/\lambda = 1/0.667 \approx 1.5$ hours/fish. Remember, the fish at Pyramid Lake tend to be large. Suppose you have just caught a fish. Let x be a random variable representing the waiting time (in hours) to catch the next fish. Use the exponential distribution to determine the probability that the waiting time is
(a) less than half an hour (i.e., $0 < x < 0.5$).
(b) more than 3 hours (i.e., $3 < x < \infty$). *Hint:* $e^{-\infty} = 0$.
(c) between 1 and 3 hours (i.e., $1 < x < 3$).
(d) What are the mean and the standard deviation of the x distribution?

SECTION 6.2

Standard Units and Areas Under the Standard Normal Distribution

FOCUS POINTS

- Given μ and σ, convert raw data to z scores.
- Given μ and σ, convert z scores to raw data.
- Graph the standard normal distribution, and find areas under the standard normal curve.

z Scores and Raw Scores

Normal distributions vary from one another in two ways: The mean μ may be located anywhere on the x axis, and the bell shape may be more or less spread according to the size of the standard deviation σ. The differences among the normal distributions cause difficulties when we try to compute the area under the curve in a specified interval of x values and, hence, the probability that a measurement will fall into that interval.

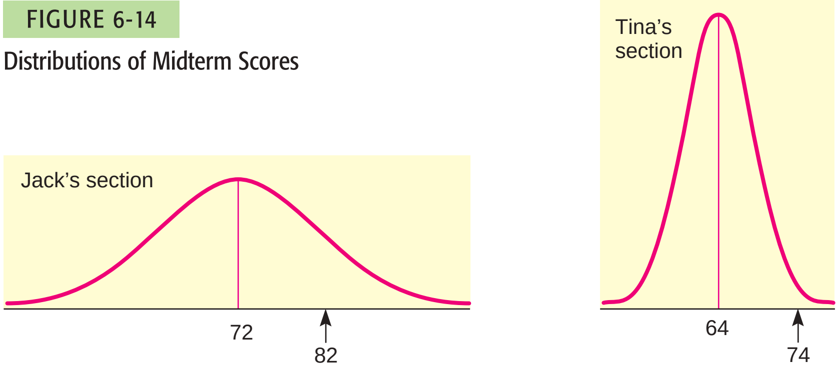
It would be a futile task to try to set up a table of areas under the normal curve for each different μ and σ combination. We need a way to standardize the distributions so that we can use *one* table of areas for *all* normal distributions. We achieve this standardization by considering how many standard deviations a measurement lies from the mean. In this way, we can compare a value in one normal distribution with a value in another, different normal distribution. The next situation shows how this is done.

Suppose Tina and Jack are in two different sections of the same course. Each section is quite large, and the scores on the midterm exams of each section follow a normal distribution. In Tina's section, the average (mean) was 64 and her score was 74. In Jack's section, the mean was 72 and his score was 82. Both Tina and Jack were pleased that their scores were each 10 points above the average of each respective section. However, the fact that each was 10 points above average does not really tell us how each did *with respect to the other students in the section*. In Figure 6-14, we see the normal distribution of grades for each section.

Tina's 74 was higher than most of the other scores in her section, while Jack's 82 is only an upper-middle score in his section. Tina's score is far better with respect to her class than Jack's score is with respect to his class.

FIGURE 6-14

Distributions of Midterm Scores



The preceding situation demonstrates that it is not sufficient to know the difference between a measurement (x value) and the mean of a distribution. We need also to consider the spread of the curve, or the standard deviation. What we really want to know is the number of standard deviations between a measurement and the mean. This "distance" takes both μ and σ into account.

We can use a simple formula to compute the number z of standard deviations between a measurement x and the mean μ of a normal distribution with standard deviation σ:

$$\left(\begin{array}{c}\text{Number of standard deviations}\\ \text{between the measurement and}\\ \text{the mean}\end{array}\right) = \left(\frac{\begin{array}{c}\text{Difference between the}\\ \text{measurement and the mean}\end{array}}{\text{Standared deviation}}\right)$$

z score or standard score

The z **value** or z **score** gives the number of standard deviations between the original measurement x and the mean μ of the x distribution.

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

TABLE 6-2

***x* Values and Corresponding *z* Values**

x Value in Original Distribution	Corresponding z Value or Standard Unit
$x = \mu$	$z = 0$
$x > \mu$	$z > 0$
$x < \mu$	$z < 0$

Standard units

The mean is a special value of a distribution. Let's see what happens when we convert $x = \mu$ to a z value:

$$z = \frac{x - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

The mean of the original distribution is always zero, in *standard units*. This makes sense because the mean is zero standard variations from itself.

An x value in the original distribution that is *above* the mean μ has a corresponding z value that is *positive*. Again, this makes sense because a measurement above the mean would be a positive number of standard deviations from the mean. Likewise, an x value *below* the mean has a *negative* z value. (See Table 6-2.)

Note

Unless otherwise stated, in the remainder of the book we will take the word *average* to be either the sample arithmetic mean $\bar{x}$ or the population mean μ.

EXAMPLE 4 STANDARD SCORE

A pizza parlor franchise specifies that the average (mean) amount of cheese on a large pizza should be 8 ounces and the standard deviation only 0.5 ounce. An inspector picks out a large pizza at random in one of the pizza parlors and finds that it is made with 6.9 ounces of cheese. Assume that the amount of cheese on a pizza follows a normal distribution. If the amount of cheese is below the mean by more than 3 standard deviations, the parlor will be in danger of losing its franchise.

Lois Ellen Frank/CORBIS

How many standard deviations from the mean is 6.9? Is the pizza parlor in danger of losing its franchise?

SOLUTION: Since we want to know the number of standard deviations from the mean, we want to convert 6.9 to standard z units.

$$z = \frac{x - \mu}{\sigma} = \frac{6.9 - 8}{0.5} = -2.20$$

Interpretation The amount of cheese on the selected pizza is only 2.20 standard deviations below the mean. The fact that z is negative indicates that the amount of cheese is 2.20 standard deviations *below* the mean. The parlor will not lose its franchise based on this sample.

Raw score, x

We have seen how to convert from x measurements to standard units z. We can easily reverse the process to find the original *raw score* x if we know the mean and standard deviation of the original x distribution. Simply solve the z score formula for x.

Given an x distribution with mean μ and standard deviation σ, the **raw score x** corresponding to a z score is

$$x = z\sigma + \mu$$

GUIDED EXERCISE 4 Standard score and raw score

Rod figures that it takes an average (mean) of 17 minutes with a standard deviation of 3 minutes to drive from home, park the car, and walk to an early morning class.

(a) One day it took Rod 21 minutes to get to class. How many standard deviations from the average is that? Is the z value positive or negative? Explain why it should be either positive or negative.

→ The number of standard deviations from the mean is given by the z value:

$$z = \frac{x - \mu}{\sigma} = \frac{21 - 17}{3} \approx 1.33$$

The z value is positive. We should expect a positive z value, since 21 minutes is *more* than the mean of 17.

(b) What commuting time corresponds to a standard score of $z = -2.5$? *Interpretation* Could Rod count on making it to class in this amount of time or less?

→ $$x = z\sigma + \mu$$
$$= (-2.5)(3) + 17$$
$$= 9.5 \text{ minutes}$$

No, commute times at or less than 2.5 standard deviations below the mean are rare.

LOOKING FORWARD

The basic structure of the formula for the standard score of a distribution is very general. When we verbalize the formula, we see it is

$$z = \frac{\text{measurement} - \text{mean of the distribution}}{\text{standard deviation of the distribution}}$$

We will see this general formula used again and again. In particular, when we look at sampling distributions for the mean (Section 6.4) and when we use the normal approximation of the binomial distribution (Section 6.6), we'll see this formula. We'll also see it when we discuss the sampling distribution for proportions (Section 6.6). Further uses occur in computations for confidence intervals (Chapter 7) and hypothesis testing (Chapter 8).

Standard Normal Distribution

If the original distribution of x *values is normal,* then the corresponding z *values have a normal distribution as well.* The z distribution has a mean of 0 and a standard deviation of 1. The normal curve with these properties has a special name.

Standard normal distribution

The **standard normal distribution** is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ (Figure 6-15).

Any normal distribution of x values can be converted to the standard normal distribution by converting all x values to their corresponding z values. The resulting standard distribution will always have mean $\mu = 0$ and standard deviation $\sigma = 1$.

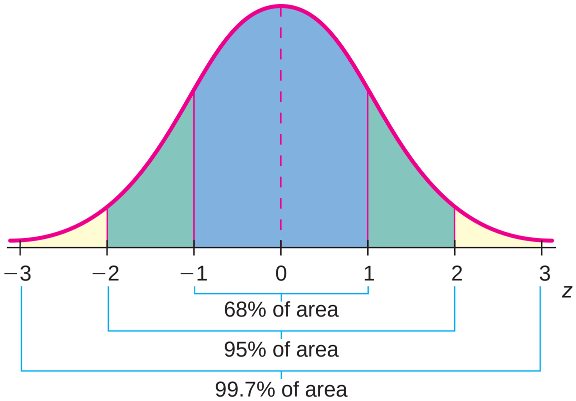
Areas Under the Standard Normal Curve

Area under the standard normal curve

We have seen how to convert *any* normal distribution to the *standard* normal distribution. We can change any x value to a z value and back again. But what is the advantage of all this work? The advantage is that there are extensive tables that show the *area under the standard normal curve* for almost any interval along the z axis. The areas are important because each area is equal to the *probability* that the measurement of an item selected at random falls in this interval. Thus, the *standard* normal distribution can be a tremendously helpful tool.

FIGURE 6-15

The Standard Normal Distribution ($\mu = 0, \sigma = 1$)



Using a Standard Normal Distribution Table

Using a table to find areas and probabilities associated with the standard normal distribution is a fairly straightforward activity. However, it is important to first observe the range of z values for which areas are given. This range is usually depicted in a picture that accompanies the table.

Left-tail style table

In this text, *we will use the left-tail style table*. This style table gives cumulative areas to the left of a specified z. Determining other areas under the curve utilizes the fact that the area under the entire curve is 1. Taking advantage of the symmetry of the normal distribution is also useful. The procedures you learn for using the left-tail style normal distribution table apply directly to cumulative normal distribution areas found on calculators and in computer software packages such as Excel 2007 and Minitab.

EXAMPLE 5 STANDARD NORMAL DISTRIBUTION TABLE

Use Table 5 of Appendix II to find the described areas under the standard normal curve.

(a) Find the area under the standard normal curve to the left of $z = -1.00$.

SOLUTION: First, shade the area to be found on the standard normal distribution curve, as shown in Figure 6-16. Notice that the z value we are using is negative. This means that we will look at the portion of Table 5 of Appendix II for which the z values are negative. In the upper-left corner of the table we see the letter z. The column under z gives us the units value and tenths value for z. The other column headings indicate the hundredths value of z. Table entries give areas under the standard normal curve to the left of the listed z values. To find the area to the left of $z = -1.00$, we use the row headed by -1.0 and then move to the column headed by the hundredths position, .00. This entry is shaded in Table 6-3. We see that the area is 0.1587.

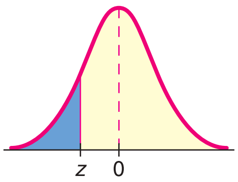


TABLE 6-3 Excerpt from Table 5 of Appendix II Showing Negative z Values

z	.00	.01	...	.07	.08	.09
−3.4	.0003	.0003	...	.0003	.0003	.0002
⋮						
−1.1	.1357	.1335	...	.1210	.1190	.1170
−1.0	.1587	.1562	...	.1423	.1401	.1379
−0.9	.1841	.1814	...	.1660	.1635	.1611
⋮						
−0.0	.5000	.4960	...	.4721	.4681	.4641

(b) Find the area to the left of $z = 1.18$, as illustrated in Figure 6-17.

FIGURE 6-16

Area to the Left of $z = -1.00$

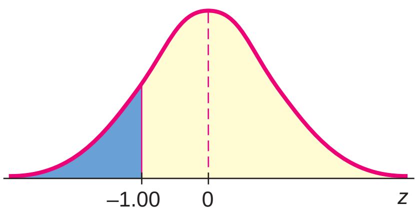
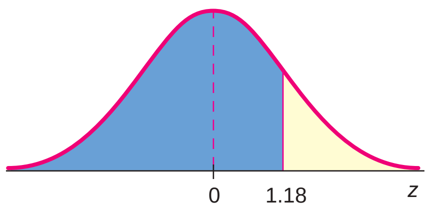


FIGURE 6-17

Area to the Left of $z = 1.18$



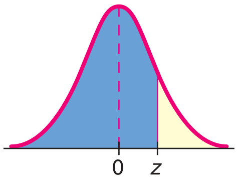


TABLE 6-4 Excerpt from Table 5 of Appendix II Showing Positive z Values

z	.00	.01	.02	. . .	.08	.09
0.0	.5000	.5040	.5080	. . .	.5319	.5359
:						
0.9	.8159	.8186	.8212	. . .	.8365	.8359
1.0	.8413	.8438	.8461	. . .	.8599	.8621
1.1	.8643	.8665	.8686	. . .	.8810	.8830
:						
3.4	.9997	.9997	.9997	. . .	.9997	.9998

SOLUTION: In this case, we are looking for an area to the left of a positive z value, so we look in the portion of Table 5 that shows positive z values. Again, we first sketch the area to be found on a standard normal curve, as shown in Figure 6-17. Look in the row headed by 1.1 and move to the column headed by .08. The desired area is shaded (see Table 6-4). We see that the area to the left of 1.18 is 0.8810.

GUIDED EXERCISE 5 *Using the standard normal distribution table*

Table 5, Areas of a Standard Normal Distribution, is located in Appendix II as well as in the end-papers of the text. Spend a little time studying the table, and then answer these questions.

(a) As z values increase, do the areas to the left of z increase?

→ Yes. As z values increase, we move to the right on the normal curve, and the areas increase.

(b) If a z value is negative, is the area to the left of z less than 0.5000?

→ Yes. Remember that a negative z value is on the left side of the standard normal distribution. The entire left half of the normal distribution has area 0.5, so any area to the left of $z = 0$ will be less than 0.5.

(c) If a z value is positive, is the area to the left of z greater than 0.5000?

→ Yes. Positive z values are on the right side of the standard normal distribution, and any area to the left of a positive z value includes the entire left half of the normal distribution.

Using Table 5 to find other areas

Table 5 gives areas under the standard normal distribution that are to the *left of a z value*. How do we find other areas under the standard normal curve?

PROCEDURE

How to use a left-tail style standard normal distribution table

1. For areas to the left of a specified z value, use the table entry directly.
2. For areas to the right of a specified z value, look up the table entry for z and subtract the area from 1.
   Note: Another way to find the same area is to use the symmetry of the normal curve and look up the table entry for $-z$.
3. For areas between two z values, z_1 and z_2 (where $z_2 > z_1$), *subtract* the table area for z_1 from the table area for z_2.

FIGURE 6-18

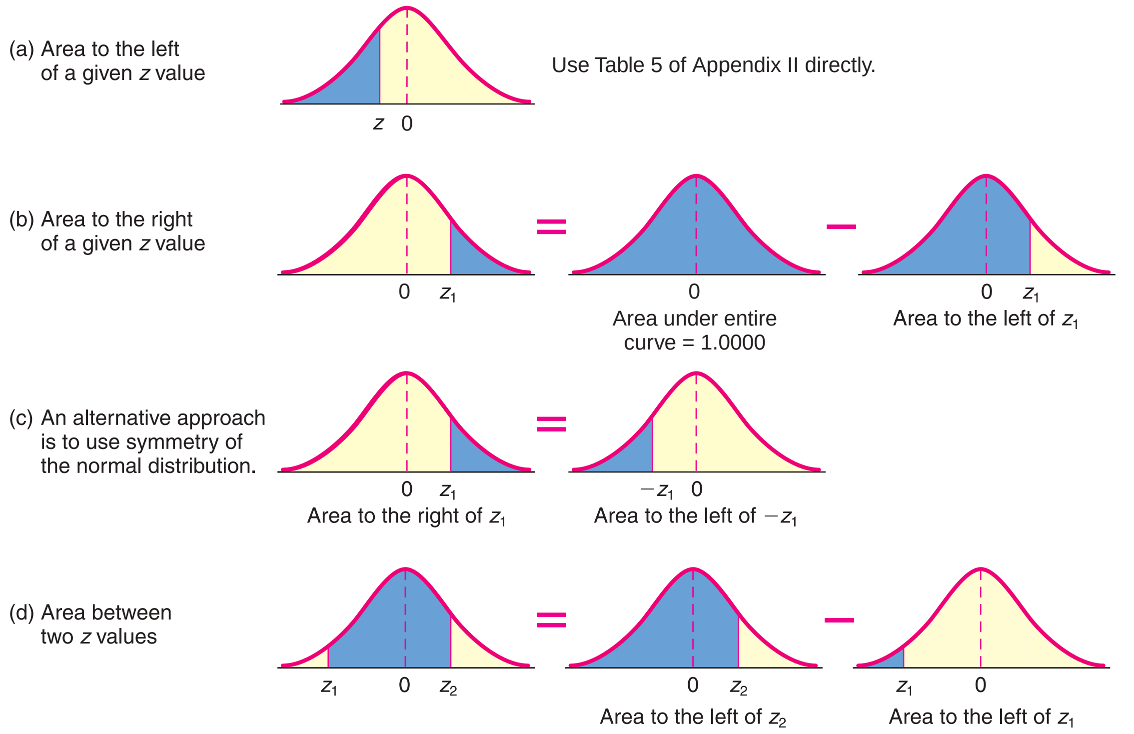


Figure 6-18 illustrates the procedure for using Table 5, Areas of a Standard Normal Distribution, to find any specified area under the standard normal distribution. Again, it is useful to sketch the area in question before you use Table 5.

COMMENT Notice that the z values shown in Table 5 of Appendix II are formatted to the hundredths position. It is convenient to *round or format z values to the hundredths position* before using the table. The areas are all given to four places after the decimal, so give your answers to four places after the decimal.

COMMENT The smallest z value shown in Table 5 is -3.49, while the largest value is 3.49. These values are, respectively, far to the left and far to the right on the standard normal distribution, with very little area beyond either value. We will follow the common convention of treating any area to the left of a z value smaller than -3.49 as 0.000. Similarly, we will consider any area to the right of a z value greater than 3.49 as 0.000. We understand that there is some area in these extreme tails. However, these areas are each less than 0.0002. Now let's get real about this! Some very specialized applications, beyond the scope of this book, do need to measure areas and corresponding probabilities in these extreme tails. But in most practical applications, *we follow the convention of treating the areas in the extreme tails as zero.*

Convention for using Table 5 of Appendix II

1. Treat any area to the left of a z value smaller than -3.49 as 0.000.
2. Treat any area to the left of a z value greater than 3.49 as 1.000.

EXAMPLE 6 USING TABLE TO FIND AREAS

Use Table 5 of Appendix II to find the specified areas.

(a) Find the area between $z = 1.00$ and $z = 2.70$.

SOLUTION: First, sketch a diagram showing the area (see Figure 6-19). Because we are finding the area between two z values, we subtract corresponding table entries.

$$\begin{aligned}(\text{Area between 1.00 and 2.70}) &= (\text{Area left of 2.70}) - (\text{Area left of 1.00})\\ &= 0.9965 - 08413\\ &= 0.1552\end{aligned}$$

(b) Find the area to the right of $z = 0.94$.

SOLUTION: First, sketch the area to be found (see Figure 6-20).

$$\begin{aligned}(\text{Area to right of 0.94}) &= (\text{Area under entire curve}) - (\text{Area to left of 0.94})\\ &= 1.000 - 0.8264\\ &= 0.1736\end{aligned}$$

Alternatively,

$$\begin{aligned}(\text{Area to right of 0.94}) &= (\text{Area to left of } -0.94)\\ &= 0.1736\end{aligned}$$

FIGURE 6-19

Area from $z = 1.00$ to $z = 2.70$

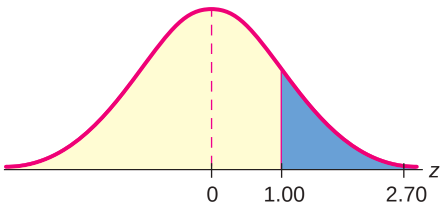
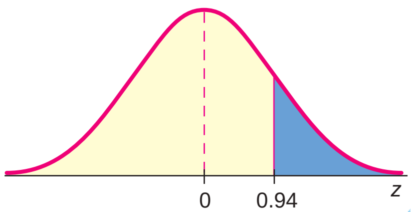


FIGURE 6-20

Area to the Right of $z = 0.94$.



Probabilities associated with the standard normal distribution

We have practiced the skill of finding areas under the standard normal curve for various intervals along the z axis. This skill is important because *the probability* that z lies in an interval *is given by the area* under the standard normal curve above that interval.

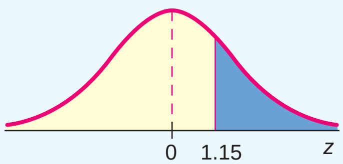
Because the normal distribution is continuous, there is no area under the curve exactly over a specific z. Therefore, probabilities such as $P(z \geq z_1)$ are the same as $P(z > z_1)$. When dealing with probabilities or areas under a normal curve that are specified with inequalities, *strict inequality* symbols can be used *interchangeably* with *inequality or equal* symbols.

GUIDED EXERCISE 6 Probabilities associated with the standard normal distribution

Let z be a random variable with a standard normal distribution.

(a) $P(z \geq 1.15)$ refers to the probability that z values lie to the right of 1.15. Shade the corresponding area under the standard normal curve (Figure 6-21) and find $P(z \geq 1.15)$.

FIGURE 6-21 Area to Be Found



$P(z \geq 1.15) = 1.000 - P(z \leq 1.15) = 1.000 - 0.8749 = 0.1251$

Alternatively,

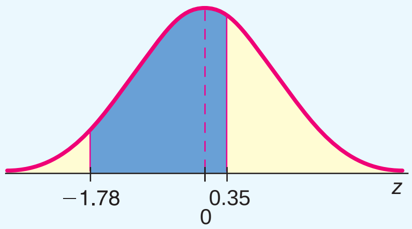
$P(z \geq 1.15) = P(z \leq -1.15) = 0.1251$

Continued

GUIDED EXERCISE 6 *continued*

(b) Find $P(-1.78 \leq z \leq 0.35)$. First, sketch the area under the standard normal curve corresponding to the area (Figure 6-22).

FIGURE 6-22 Area to Be Found



$$P(-1.78 \leq z \leq 0.35) = P(z \leq 0.35) - P(z \leq -1.78)$$
$$= 0.6368 - 0.0375 = 0.5993$$

TECH NOTES The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all provide cumulative areas under any normal distribution, including the standard normal. The Tech Note of Section 6.3 shows examples.

VIEWPOINT Mighty Oaks from Little Acorns Grow!

Just how big is that acorn? What if we compare it with other acorns? Is that oak tree taller than an average oak tree? How does it compare with other oak trees? What do you mean, this oak tree has a larger geographic range? Compared with what? Answers to questions such as these can be given only if we resort to standardized statistical units. Can you compare a single oak tree with an entire forest of oak trees? The answer is yes, if you use standardized z scores. *For more information about sizes of acorns, oak trees, and geographic locations, visit the Brase/Brase statistics site at* **http://www.cengage.com/statistics/brase** *and find the link to DASL, the Carnegie Mellon University Data and Story Library. From the DASL site, find Biology under Data Subjects, and select Acorns.*

SECTION 6.2 PROBLEMS

In these problems, assume that all distributions are *normal*. In all problems in Chapter 6, *average* is always taken to be the arithmetic mean $\bar{x}$ or μ.

1. *Statistical Literacy* What does a standard score measure?
2. *Statistical Literacy* Does a raw score less than the mean correspond to a positive or negative standard score? What about a raw score greater than the mean?
3. *Statistical Literacy* What is the value of the standard score for the mean of a distribution?
4. *Statistical Literacy* What are the values of the mean and standard deviation of a standard normal distribution?
5. *Basic Computation: z Score and Raw Score* A normal distribution has $\mu = 30$ and $\sigma = 5$.
   (a) Find the z score corresponding to $x = 25$.
   (b) Find the z score corresponding to $x = 42$.
   (c) Find the raw score corresponding to $z = -2$.
   (d) Find the raw score corresponding to $z = 1.3$.

6. ***Basic Computation: z Score and Raw Score*** A normal distribution has $\mu = 10$ and $\sigma = 2$.
(a) Find the z score corresponding to $x = 12$.
(b) Find the z score corresponding to $x = 4$.
(c) Find the raw score corresponding to $z = 1.5$.
(d) Find the raw score corresponding to $z = -1.2$.

7. ***Critical Thinking*** Consider the following scores:
(i) Score of 40 from a distribution with mean 50 and standard deviation 10
(ii) Score of 45 from a distribution with mean 50 and standard deviation 5

How do the two scores compare relative to their respective distributions?

8. ***Critical Thinking*** Raul received a score of 80 on a history test for which the class mean was 70 with standard deviation 10. He received a score of 75 on a biology test for which the class mean was 70 with standard deviation 2.5. On which test did he do better relative to the rest of the class?

9. ***z Scores: First Aid Course*** The college physical education department offered an advanced first aid course last semester. The scores on the comprehensive final exam were normally distributed, and the z scores for some of the students are shown below:

Robert, 1.10	Juan, 1.70	Susan, −2.00
Joel, 0.00	Jan, −0.80	Linda, 1.60.


Comstock/Jupiter Images

(a) Which of these students scored above the mean?
(b) Which of these students scored on the mean?
(c) Which of these students scored below the mean?
(d) If the mean score was $\mu = 150$ with standard deviation $\sigma = 20$, what was the final exam score for each student?

10. ***z Scores: Fawns*** Fawns between 1 and 5 months old in Mesa Verde National Park have a body weight that is approximately normally distributed with mean $\mu = 27.2$ kilograms and standard deviation $\sigma = 4.3$ kilograms (based on information from *The Mule Deer of Mesa Verde National Park,* by G. W. Mierau and J. L. Schmidt, Mesa Verde Museum Association). Let x be the weight of a fawn in kilograms. Convert each of the following x intervals to z intervals.
(a) $x < 30$ (b) $19 < x$ (c) $32 < x < 35$

Convert each of the following z intervals to x intervals.
(d) $-2.17 < z$ (e) $z < 1.28$ (f) $-1.99 < z < 1.44$
(g) ***Interpretation*** If a fawn weighs 14 kilograms, would you say it is an unusually small animal? Explain using z values and Figure 6-15.
(h) ***Interpretation*** If a fawn is unusually large, would you say that the z value for the weight of the fawn will be close to 0, −2, or 3? Explain.

11. ***z Scores: Red Blood Cell Count*** Let x = red blood cell (RBC) count in millions per cubic millimeter of whole blood. For healthy females, x has an approximately normal distribution with mean $\mu = 4.8$ and standard deviation $\sigma = 0.3$ (based on information from *Diagnostic Tests with Nursing Implications,* edited by S. Loeb, Springhouse Press). Convert each of the following x intervals to z intervals.
(a) $4.5 < x$ (b) $x < 4.2$ (c) $4.0 < x < 5.5$

Convert each of the following z intervals to x intervals.
(d) $z < -1.44$ (e) $1.28 < z$ (f) $-2.25 < z < -1.00$
(g) ***Interpretation*** If a female had an RBC count of 5.9 or higher, would that be considered unusually high? Explain using z values and Figure 6-15.

12. ***Normal Curve: Tree Rings*** Tree-ring dates were used extensively in archaeological studies at Burnt Mesa Pueblo (*Bandelier Archaeological Excavation Project: Summer 1989 Excavations at Burnt Mesa Pueblo,* edited by Kohler, Washington State University Department of Anthropology). At one site on the mesa, tree-ring

dates (for many samples) gave a mean date of μ_1 = year 1272 with standard deviation $\sigma_1 = 35$ years. At a second, removed site, the tree-ring dates gave a mean of μ_2 = year 1122 with standard deviation $\sigma_2 = 40$ years. Assume that both sites had dates that were approximately normally distributed. In the first area, an object was found and dated as x_1 = year 1250. In the second area, another object was found and dated as x_2 = year 1234.

(a) Convert both x_1 and x_2 to z values, and locate both of these values under the standard normal curve of Figure 6-15.

(b) *Interpretation* Which of these two items is the more unusual as an archaeological find in its location?

Basic Computation: Finding Areas Under the Standard Normal Curve In Problems 13–30, sketch the areas under the standard normal curve over the indicated intervals, and find the specified areas.

13. To the right of $z = 0$
14. To the left of $z = 0$
15. To the left of $z = -1.32$
16. To the left of $z = -0.47$
17. To the left of $z = 0.45$
18. To the left of $z = 0.72$
19. To the right of $z = 1.52$
20. To the right of $z = 0.15$
21. To the right of $z = -1.22$
22. To the right of $z = -2.17$
23. Between $z = 0$ and $z = 3.18$
24. Between $z = 0$ and $z = -1.93$
25. Between $z = -2.18$ and $z = 1.34$
26. Between $z = -1.40$ and $z = 2.03$
27. Between $z = 0.32$ and $z = 1.92$
28. Between $z = 1.42$ and $z = 2.17$
29. Between $z = -2.42$ and $z = -1.77$
30. Between $z = -1.98$ and $z = -0.03$

Basic Computation: Finding Probabilities In Problems 31–50, let z be a random variable with a standard normal distribution. Find the indicated probability, and shade the corresponding area under the standard normal curve.

31. $P(z \le 0)$
32. $P(z \ge 0)$
33. $P(z \le -0.13)$
34. $P(z \le -2.15)$
35. $P(z \le 1.20)$
36. $P(z \le 3.20)$
37. $P(z \ge 1.35)$
38. $P(z \ge 2.17)$
39. $P(z \ge -1.20)$
40. $P(z \ge -1.50)$
41. $P(-1.20 \le z \le 2.64)$
42. $P(-2.20 \le z \le 1.40)$
43. $P(-2.18 \le z \le -0.42)$
44. $P(-1.78 \le z \le -1.23)$
45. $P(0 \le z \le 1.62)$
46. $P(0 \le z \le 0.54)$
47. $P(-0.82 \le z \le 0)$
48. $P(-2.37 \le z \le 0)$
49. $P(-0.45 \le z \le 2.73)$
50. $P(-0.73 \le z \le 3.12)$

SECTION 6.3

Areas Under Any Normal Curve

FOCUS POINTS

- Compute the probability of "standardized events."
- Find a z score from a given normal probability (inverse normal).
- Use the inverse normal to solve guarantee problems.

Normal Distribution Areas

In many applied situations, the original normal curve is not the standard normal curve. Generally, there will not be a table of areas available for the original normal curve. This does not mean that we cannot find the probability that a measurement x will fall into an interval from a to b. What we must do is *convert* the original measurements x, a, and b to z values.

PROCEDURE

How to work with normal distributions

To find areas and probabilities for a random variable x that follows a normal distribution with mean μ and standard deviation σ, convert x values to z values using the formula

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

Then use Table 5 of Appendix II to find corresponding areas and probabilities.

To compute a z score, we need to know the standard deviation of the distribution. But what if we don't know σ? If we have a range of data values, we can estimate σ. Problems 31 to 34 show how this is done.

EXAMPLE 7

Normal distribution probability

Let x have a normal distribution with $\mu = 10$ and $\sigma = 2$. Find the probability that an x value selected at random from this distribution is between 11 and 14. In symbols, find $P(11 \le x \le 14)$.

SOLUTION: Since probabilities correspond to areas under the distribution curve, we want to find the area under the x curve above the interval from $x = 11$ to $x = 14$. To do so, we will convert the x values to standard z values and then use Table 5 of Appendix II to find the corresponding area under the standard curve.

We use the formula

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

to convert the given x interval to a z interval.

$$z_1 = \frac{11 - 10}{2} = 0.50 \qquad (\text{Use } x = 11, \mu = 10, \sigma = 2.)$$

$$z_2 = \frac{14 - 10}{2} = 2.00 \qquad (\text{Use } x = 14, \mu = 10, \sigma = 2.)$$

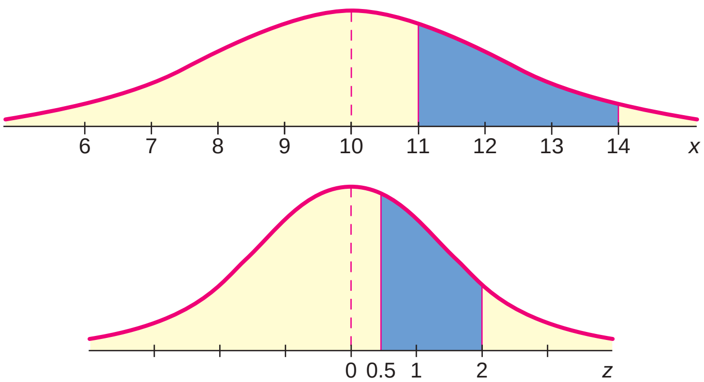
The corresponding areas under the x and z curves are shown in Figure 6-23. From Figure 6-23 we see that

$$\begin{aligned} P(11 \le x \le 14) &= P(0.50 \le z \le 2.00) \\ &= P(z \le 2.00) - P(z \le 0.50) \\ &= 0.9772 - 0.6915 \quad (\text{From Table 5, Appendix II}) \\ &= 0.2857 \end{aligned}$$

Interpretation The probability is 0.2857 that an x value selected at random from a normal distribution with mean 10 and standard deviation 2 lies between 11 and 14.

FIGURE 6-23

Corresponding Areas Under the x Curve and z Curve



GUIDED EXERCISE 7 Normal distribution probability

Sunshine Stereo cassette decks have a deck life that is normally distributed with a mean of 2.3 years and a standard deviation of 0.4 year. What is the probability that a cassette deck will break down during the guarantee period of 2 years?

(a) Let x represent the life of a cassette deck. The statement that the cassette deck breaks during the 2-year guarantee period means the life is less than 2 years, or $x \le 2$. Convert this to a statement about z.

$$z = \frac{x - \mu}{\sigma} = \frac{2 - 2.3}{0.4} = -0.75$$

So, $x \le 2$ means $z \le -0.75$.

(b) Indicate the area to be found in Figure 6-24. Does this area correspond to the probability that $z \le -0.75$?

See Figure 6-25.
Yes, the shaded area does correspond to the probability that $z \le -0.75$.

FIGURE 6-24

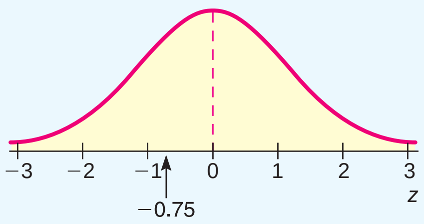
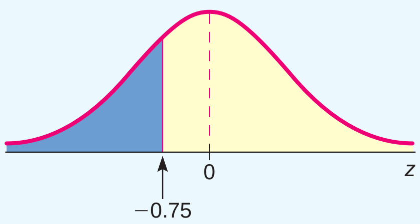


FIGURE 6-25 $z \le -0.75$



(c) Use Table 5 of Appendix II to find $P(z \le -0.75)$.

0.2266

(d) *Interpretation* What is the probability that the cassette deck will break before the end of the guarantee period? [*Hint:* $P(x \le 2) = P(z \le -0.75)$.]

The probability is

$$P(x \le 2) = P(z \le -0.75) = 0.2266$$

This means that the company will repair or replace about 23% of the cassette decks.

TECH NOTES

The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all provide areas under any normal distribution. Excel 2007 and Minitab give the left-tail area to the left of a specified x value. The TI-84Plus/TI-83Plus/TI-*n*spire has you specify an interval from a lower bound to an upper bound and provides the area under the normal curve for that interval. For example, to solve Guided Exercise 7 regarding the probability a cassette deck will break during the guarantee period, we find $P(x \le 2)$ for a normal distribution with $\mu = 2.3$ and $\sigma = 0.4$.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Press the **DISTR** key, select **2:normalcdf** (**lower bound, upper bound,** μ, σ) and press Enter. Type in the specified values. For a left-tail area, use a lower bound setting at about 4 standard deviations below the mean. Likewise, for a right-tail area, use an upper bound setting about 4 standard deviations above the mean. For our example, use a lower bound of $\mu - 4\sigma = 2.3 - 4(0.4) = 0.7$.

```
normalcdf(.7,2,2.3,
.4)
          .2265955934
```

Excel 2007 Select **Insert Function** (f_x) In the dialogue box, select **Statistical** for the Category and then for the Function, select **NORMDIST**. Fill in the dialogue box using True for cumulative.

f_x =NORMDIST(2,2.3,0.4, TRUE)

C	D	E
0.226627		

What if we need to compute conditional probabilities based on the normal distribution? Problems 39 and 40 of this section show how we can do this.

Minitab Use the menu selection **Calc ➤ Probability Distribution ➤ Normal.** Fill in the dialogue box, marking cumulative.

```
Cumulative Distribution Function
Normal with mean = 2.3 and
standard deviation = 0.4
       x     P(X <= x)
     2.0     0.2266
```

Inverse Normal Distribution

Finding z or x, given a probability

Sometimes we need to find z or x values that correspond to a given area under the normal curve. This situation arises when we want to specify a guarantee period such that a given percentage of the total products produced by a company last at least as long as the duration of the guarantee period. In such cases, we use the standard normal distribution table "in reverse." When we look up an area and find the corresponding z value, we are using the *inverse normal probability distribution*.

Inverse normal probability distribution

EXAMPLE 8 Find x, given probability

Magic Video Games, Inc., sells an expensive video games package. Because the package is so expensive, the company wants to advertise an impressive guarantee for the life expectancy of its computer control system. The guarantee policy will refund the full purchase price if the computer fails during the guarantee period. The research department has done tests that show that the mean life for the computer is 30 months, with standard deviation of 4 months. The computer life is normally distributed. How long can the guarantee period be if management does not want to refund the purchase price on more than 7% of the Magic Video packages?

SOLUTION: Let us look at the distribution of lifetimes for the computer control system, and shade the portion of the distribution in which the computer lasts fewer months than the guarantee period. (See Figure 6-26.)

FIGURE 6-26

7% of the Computers Have a Lifetime Less Than the Guarantee Period

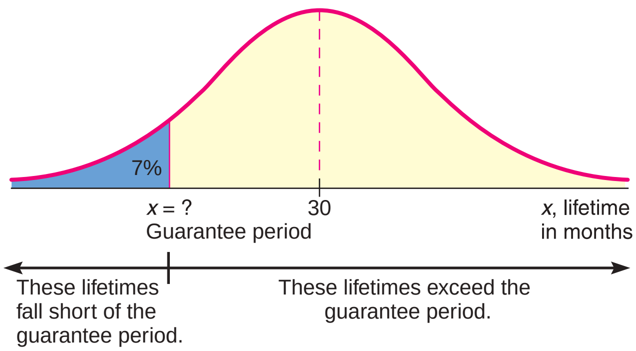


TABLE 6-5 **Excerpt from Table 5 of Appendix II**

z	.00	. . .	.07		.08	.09
⋮						
−1.4	.0808		.0708		.0694	.0681
				↑ 0.0700		

Pierre Arsenault/Alamy

If a computer system lasts fewer months than the guarantee period, a full-price refund will have to be made. The lifetimes requiring a refund are in the shaded region in Figure 6-26. This region represents 7% of the total area under the curve.

We can use Table 5 of Appendix II to find the z value such that 7% of the total area under the *standard* normal curve lies to the left of the z value. Then we convert the z value to its corresponding x value to find the guarantee period.

We want to find the z value with 7% of the area under the standard normal curve to the left of z. Since we are given the area in a left tail, we can use Table 5 of Appendix II directly to find z. The area value is 0.0700. However, this area is not in our table, so we use the closest area, which is 0.0694, and the corresponding z value of $z = -1.48$ (see Table 6-5).

To translate this value back to an x value (in months), we use the formula

$$\begin{aligned} x &= z\sigma + \mu \\ &= -1.48(4) + 30 \quad \text{(Use } \sigma = 4 \text{ months and } \mu = 30 \text{ months.)} \\ &= 24.08 \text{ months} \end{aligned}$$

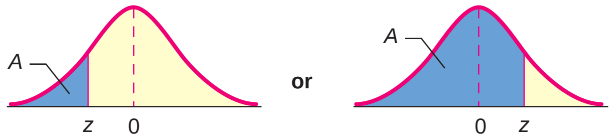
Interpretation The company can guarantee the Magic Video Games package for $x = 24$ months. For this guarantee period, they expect to refund the purchase price of no more than 7% of the video games packages.

Example 8 had us find a z value corresponding to a given area to the left of z. What if the specified area is to the right of z or between $-z$ and z? Figure 6-27 shows us how to proceed.

FIGURE 6-27

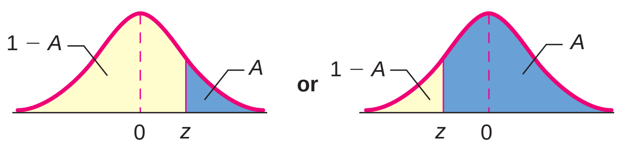
Inverse Normal: Use Table 5 of Appendix II to Find z Corresponding to a Given Area A ($0 < A < 1$)

(a) **Left-tail case:**
The given area A is to the left of z.



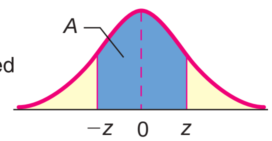
For the left-tail case, look up the number A in the body of the table and use the corresponding z value.

(b) **Right-tail case:**
The given area A is to the right of z.



For the right-tail case, look up the number $1 - A$ in the body of the table and use the corresponding z value.

(c) **Center case:**
The given area A is symmetric and centered above $z = 0$. Half of A lies to the left and half lies to the right of $z = 0$.



For the center case, look up the number $\frac{1 - A}{2}$ in the body of the table and use the corresponding $\pm z$ value.

COMMENT When we use Table 5 of Appendix II to find a z value corresponding to a given area, we usually use the nearest area value rather than interpolating between values. However, when the area value given is exactly halfway between two area values of the table, we use the z value halfway between the z values of the corresponding table areas. Example 9 demonstrates this procedure. However, this interpolation convention is not always used, especially if the area is changing slowly, as it does in the tail ends of the distribution. *When the z value corresponding to an area is smaller than -2, the standard convention is to use the z value corresponding to the smaller area. Likewise, when the z value is larger than 2, the standard convention is to use the z value corresponding to the larger area.* We will see an example of this special case in Guided Exercise 9.

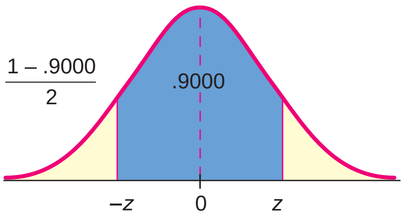
EXAMPLE 9 Find z

Find the z value such that 90% of the area under the standard normal curve lies between $-z$ and z.

SOLUTION: Sketch a picture showing the described area (see Figure 6-28).

FIGURE 6-28

Area Between $-z$ and z Is 90%



We find the corresponding area in the left tail.

$$(\text{Area left of } -z) = \frac{1 - 0.9000}{2}$$
$$= 0.0500$$

Looking in Table 6-6, we see that 0.0500 lies exactly between areas 0.0495 and 0.0505. The halfway value between $z = -1.65$ and $z = -1.64$ is $z = -1.645$. Therefore, we conclude that 90% of the area under the standard normal curve lies between the z values -1.645 and 1.645.

TABLE 6-6 **Excerpt from Table 5 of Appendix II**

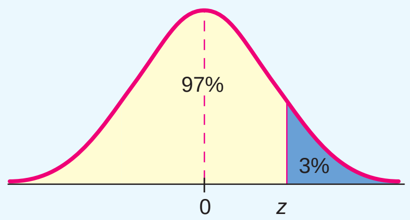
z	. . .	.04		.05
:				
−1.6		.0505		.0495
			↑ 0.0500	

GUIDED EXERCISE 8 Find z

Find the z value such that 3% of the area under the standard normal curve lies to the right of z.

(a) Draw a sketch of the standard normal distribution showing the described area (Figure 6-29).

FIGURE 6-29 3% of Total Area Lies to the Right of z



(b) Find the area to the left of z.

Area to the left of $z = 1 - 0.0300 = 0.9700$.

(c) Look up the area in Table 6-7 and find the corresponding z.

The closest area is 0.9699. This area is to the left of $z = 1.88$.

TABLE 6-7 **Excerpt from Table 5 of Appendix II**

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

(d) Suppose the time to complete a test is normally distributed with $\mu = 40$ minutes and $\sigma = 5$ minutes. After how many minutes can we expect all but about 3% of the tests to be completed?

We are looking for an x value such that 3% of the normal distribution lies to the right of x. In part (c), we found that 3% of the standard normal curve lies to the right of $z = 1.88$. We convert $z = 1.88$ to an x value.

$$x = z\sigma + \mu$$
$$= 1.88(5) + 40 = 49.4 \text{ minutes}$$

All but about 3% of the tests will be complete after 50 minutes.

(e) Use Table 6-8 to find a z value such that 3% of the area under the standard normal curve lies to the left of z.

The closest area is 0.0301. This is the area to the left of $z = -1.88$.

TABLE 6-8 **Excerpt from Table 5 of Appendix II**

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294

(f) Compare the z value of part (c) with the z value of part (e). Is there any relationship between the z values?

One z value is the negative of the other. This result is expected because of the symmetry of the normal distribution.

TECH NOTES

When we are given a z value and we find an area to the left of z, we are using a normal distribution function. When we are given an area to the left of z and we find the corresponding z, we are using an inverse normal distribution function. The

TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all have inverse normal distribution functions for any normal distribution. For instance, to find an x value from a normal distribution with mean 40 and standard deviation 5 such that 97% of the area lies to the left of x, use the described instructions.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Press the **DISTR** key and select **3:invNorm(area,μ,σ)**.

```
invNorm(.97,40,5)
        49.40396805
```

Excel 2007 Select **Insert Function** f_x In the dialogue box, select **Statistical** for the Category and then for the Function, select **NORMINV**. Fill in the dialogue box.

f_x	=NORMINV(0.97,40,5)	
	C	D
	49.40395	

Minitab Use the menu selection **Calc ➤ Probability Distribution ➤ Normal.** Fill in the dialogue box, marking Inverse Cumulative.

```
Inverse Cumulative Distribution Function
Normal with mean = 40.000 and
  standard deviation = 5.00000
         P(X <= x)      x
            0.9700   49.4040
```

LOOKING FORWARD

In our work with confidence intervals (Chapter 7), we will use inverse probability distributions for the normal distribution and for the Student's t distribution (a similar distribution introduced in Chapter 7). Just as in Example 9, we'll use inverse probability distributions to identify values such that 90%, 95%, or 99% of the area under the distribution graph centered over the mean falls between the values.

CRITICAL THINKING

Checking for Normality

How can we tell if data follow a normal distribution? There are several checks we can make. The following procedure lists some guidelines.

PROCEDURE

How to determine whether data have a normal distribution

The following guidelines represent some useful devices for determining whether or not data follow a normal distribution.

1. **Histogram:** Make a histogram. For a normal distribution, the histogram should be roughly bell-shaped.

Continued

2. **Outliers:** For a normal distribution, there should not be more than one outlier. One way to check for outliers is to use a box-and-whisker plot. Recall that outliers are those data values that are

   above Q_3 by an amount greater than 1.5 × interquartile range

   below Q_1 by an amount greater than 1.5 × interquartile range

3. **Skewness:** Normal distributions are symmetric. One measure of skewness for sample data is given by Pearson's index:

$$\text{Pearson's index} = \frac{3(\bar{x} - \text{median})}{s}$$

   An index value greater than 1 or less than −1 indicates skewness. Skewed distributions are not normal.

4. **Normal quantile plot (or normal probability plot):** This plot is provided through statistical software on a computer or graphing calculator. The Using Technology feature, Application 1, at the end of this chapter gives a brief description of how such plots are constructed. The section also gives commands for producing such plots on the TI-84Plus/TI-83Plus/TI-*n*spire calculators, Minitab, or SPSS.

   Examine a normal quantile plot of the data.

   If the points lie close to a straight line, the data come from a distribution that is approximately normal.

   If the points do not lie close to a straight line or if they show a pattern that is not a straight line, the data are likely to come from a distribution that is not normal.

EXAMPLE 10 ASSESSING NORMALITY

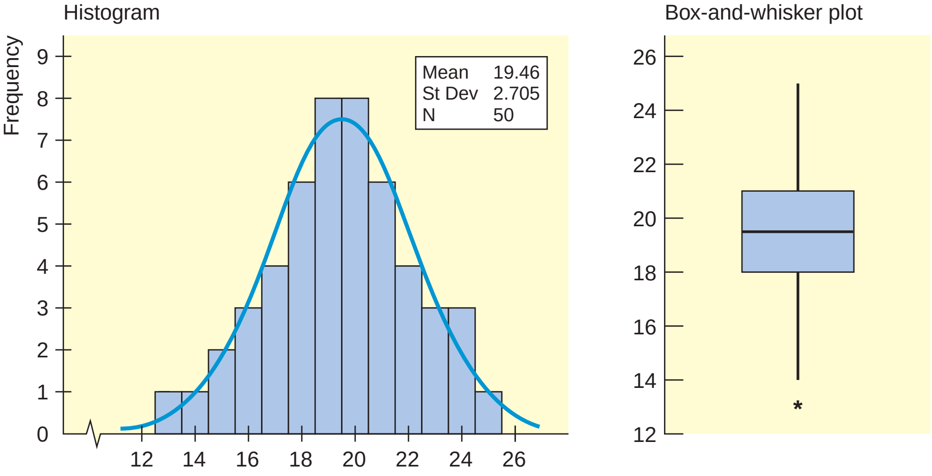
Consider the following data, which are rounded to the nearest integer.

19	19	19	16	21	14	23	17	19	20	18	24	20	13	16
17	19	18	19	17	21	24	18	23	19	21	22	20	20	20
24	17	20	22	19	22	21	18	20	22	16	15	21	23	21
18	18	20	15	25										

(a) Look at the histogram and box-and-whisker plot generated by Minitab in Figure 6-30 and comment about normality of the data from these indicators.

FIGURE 6-30

Histogram and Box-and-Whisker Plot



SOLUTION: Note that the histogram is approximately normal. The box-and-whisker plot shows just one outlier. Both of these graphs indicate normality.

(b) Use Pearson's index to check for skewness.

SOLUTION: Summary statistics from Minitab:

```
Variable    N  N*   Mean  Se Mean  StDev  Minimum      Q1  Median      Q3
C2         50   0  19.460   0.382  2.705   13.000  18.000  19.500  21.000
Variable        Maximum
C2               25.000
```

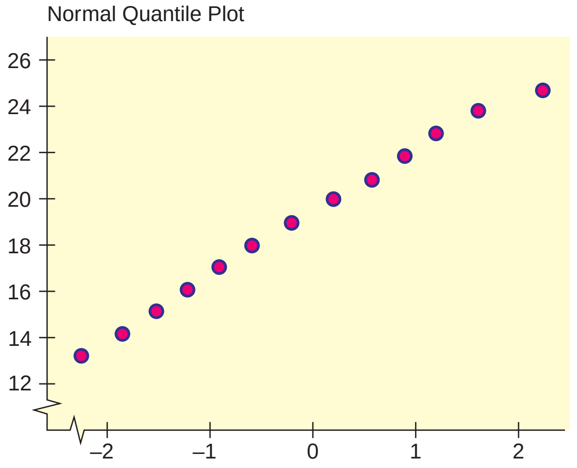
We see that $\bar{x} = 19.46$, median $= 19.5$, and $s = 2.705$.

$$\text{Pearson's index} = \frac{3(19.46 - 19.5)}{2.705} \approx -0.04$$

Since the index is between -1 and 1, we detect no skewness. The data appear to be symmetric.

(c) Look at the normal quantile plot in Figure 6-31 and comment on normality.

FIGURE 6-31
Normal Quantile Plot



SOLUTION: The data fall close to a straight line, so the data appear to come from a normal distribution.

(d) *Interpretation* Interpret the results.

SOLUTION: The histogram is roughly bell-shaped, there is only one outlier, Pearson's index does not indicate skewness, and the points on the normal quantile plot lie fairly close to a straight line. It appears that the data are likely from a distribution that is approximately normal.

VIEWPOINT Want to Be an Archaeologist?

Each year about 4500 students work with professional archaeologists in scientific research at the Crow Canyon Archaeological Center, Cortez, Colorado. In fact, Crow Canyon was included in The Princeton Review Guide to America's Top 100 Internships. *The nonprofit, multidisciplinary program at Crow Canyon enables students and laypeople with little or no archaeology background to get started in archaeological research. The only requirement is that you be interested in Native American culture and history. By the way, a knowledge of introductory statistics could come in handy for this internship. For more information about the program, visit the Brase/Brase statistics site at* **http://www.cengage.com/statistics/brase** *and find the link to Crow Canyon.*

SECTION 6.3 PROBLEMS

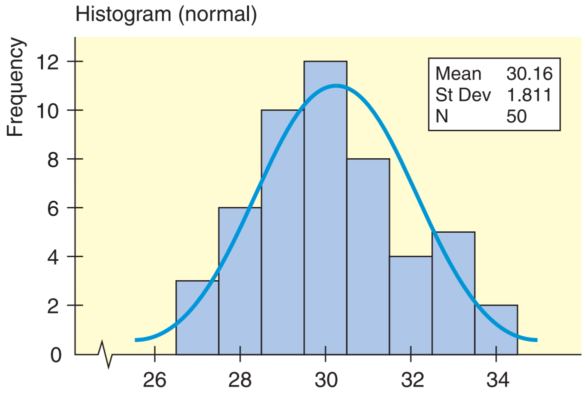
1. *Statistical Literacy* Consider a normal distribution with mean 30 and standard deviation 2. What is the probability a value selected at random from this distribution is greater than 30?

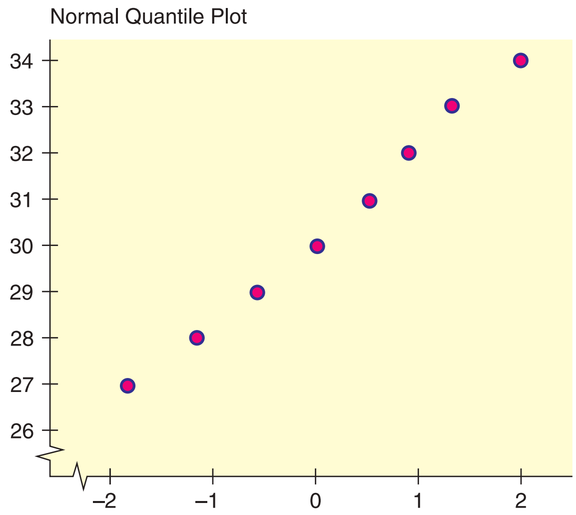
2. *Statistical Literacy* Suppose 5% of the area under the standard normal curve lies to the right of z. Is z positive or negative?

3. *Statistical Literacy* Suppose 5% of the area under the standard normal curve lies to the left of z. Is z positive or negative?

4. *Critical Thinking: Normality* Consider the following data. The summary statistics, histogram, and normal quantile plot were generated by Minitab.

27	27	27	28	28	28	28	28	28	29	29	29	29	29	29
29	29	29	29	30	30	30	30	30	30	30	30	30	30	30
30	31	31	31	31	31	31	31	31	32	32	32	32	33	33
33	33	33	34	34										






```
Variable    N  N*   Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Data       50   0  30.160   0.256  1.811   27.000  29.000  30.000  31.000
Variable    Maximum
Data         34.000
```

(a) Does the histogram indicate normality for the data distribution? Explain.
(b) Does the normal quantile plot indicate normality for the data distribution? Explain.
(c) Compute the interquartile range and check for outliers.
(d) Compute Pearson's index. Does the index value indicate skewness?
(e) Using parts (a) through (d), would you say the data are from a normal distribution?

Basic Computation: Find Probabilities In Problems 5–14, assume that x has a normal distribution with the specified mean and standard deviation. Find the indicated probabilities.

5. $P(3 \le x \le 6)$; $\mu = 4$; $\sigma = 2$
6. $P(10 \le x \le 26)$; $\mu = 15$; $\sigma = 4$
7. $P(50 \le x \le 70)$; $\mu = 40$; $\sigma = 15$
8. $P(7 \le x \le 9)$; $\mu = 5$; $\sigma = 1.2$
9. $P(8 \le x \le 12)$; $\mu = 15$; $\sigma = 3.2$
10. $P(40 \le x \le 47)$; $\mu = 50$; $\sigma = 15$
11. $P(x \ge 30)$; $\mu = 20$; $\sigma = 3.4$
12. $P(x \ge 120)$; $\mu = 100$; $\sigma = 15$
13. $P(x \ge 90)$; $\mu = 100$; $\sigma = 15$
14. $P(x \ge 2)$; $\mu = 3$; $\sigma = 0.25$

Basic Computation: Find z Values In Problems 15–24, find the z value described and sketch the area described.

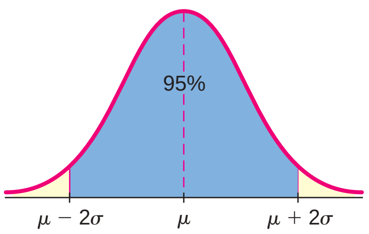
15. Find z such that 6% of the standard normal curve lies to the left of z.
16. Find z such that 5.2% of the standard normal curve lies to the left of z.
17. Find z such that 55% of the standard normal curve lies to the left of z.
18. Find z such that 97.5% of the standard normal curve lies to the left of z.
19. Find z such that 8% of the standard normal curve lies to the right of z.
20. Find z such that 5% of the standard normal curve lies to the right of z.
21. Find z such that 82% of the standard normal curve lies to the right of z.
22. Find z such that 95% of the standard normal curve lies to the right of z.
23. Find the z value such that 98% of the standard normal curve lies between $-z$ and z.
24. Find the z value such that 95% of the standard normal curve lies between $-z$ and z.
25. ***Medical: Blood Glucose*** A person's blood glucose level and diabetes are closely related. Let x be a random variable measured in milligrams of glucose per deciliter (1/10 of a liter) of blood. After a 12-hour fast, the random variable x will have a distribution that is approximately normal with mean $\mu = 85$ and standard deviation $\sigma = 25$ (Source: *Diagnostic Tests with Nursing Implications,* edited by S. Loeb, Springhouse Press). *Note:* After 50 years of age, both the mean and standard deviation tend to increase. What is the probability that, for an adult (under 50 years old) after a 12-hour fast,
    (a) x is more than 60?
    (b) x is less than 110?
    (c) x is between 60 and 110?
    (d) x is greater than 140 (borderline diabetes starts at 140)?
26. ***Medical: Blood Protoplasm*** Porphyrin is a pigment in blood protoplasm and other body fluids that is significant in body energy and storage. Let x be a random variable that represents the number of milligrams of porphyrin per deciliter of blood. In healthy adults, x is approximately normally distributed with mean $\mu = 38$ and standard deviation $\sigma = 12$ (see reference in Problem 25). What is the probability that
    (a) x is less than 60?
    (b) x is greater than 16?
    (c) x is between 16 and 60?
    (d) x is more than 60? (This may indicate an infection, anemia, or another type of illness.)
27. ***Archaeology: Hopi Village*** Thickness measurements of ancient prehistoric Native American pot shards discovered in a Hopi village are approximately normally distributed, with a mean of 5.1 millimeters (mm) and a standard deviation of 0.9 mm (Source: *Homol'ovi II: Archaeology of an Ancestral Hopi Village, Arizona,* edited by E. C. Adams and K. A. Hays, University of Arizona Press). For a randomly found shard, what is the probability that the thickness is
    (a) less than 3.0 mm?
    (b) more than 7.0 mm?
    (c) between 3.0 mm and 7.0 mm?
28. ***Law Enforcement: Police Response Time*** Police response time to an emergency call is the difference between the time the call is first received by the dispatcher and the time a patrol car radios that it has arrived at the scene (based on information from *The Denver Post*). Over a long period of time, it has been determined that the police response time has a normal distribution with a mean of 8.4 minutes and a standard deviation of 1.7 minutes. For a

randomly received emergency call, what is the probability that the response time will be
(a) between 5 and 10 minutes?
(b) less than 5 minutes?
(c) more than 10 minutes?

29. *Guarantee: Batteries* Quick Start Company makes 12-volt car batteries. After many years of product testing, the company knows that the average life of a Quick Start battery is normally distributed, with a mean of 45 months and a standard deviation of 8 months.
(a) If Quick Start guarantees a full refund on any battery that fails within the 36-month period after purchase, what percentage of its batteries will the company expect to replace?
(b) *Inverse Normal Distribution* If Quick Start does not want to make refunds for more than 10% of its batteries under the full-refund guarantee policy, for how long should the company guarantee the batteries (to the nearest month)?

30. *Guarantee: Watches* Accrotime is a manufacturer of quartz crystal watches. Accrotime researchers have shown that the watches have an average life of 28 months before certain electronic components deteriorate, causing the watch to become unreliable. The standard deviation of watch lifetimes is 5 months, and the distribution of lifetimes is normal.
(a) If Accrotime guarantees a full refund on any defective watch for 2 years after purchase, what percentage of total production should the company expect to replace?
(b) *Inverse Normal Distribution* If Accrotime does not want to make refunds on more than 12% of the watches it makes, how long should the guarantee period be (to the nearest month)?

31. *Expand Your Knowledge: Estimating the Standard Deviation* *Consumer Reports* gave information about the ages at which various household products are replaced. For example, color TVs are replaced at an average age of $\mu = 8$ years after purchase, and the (95% of data) range was from 5 to 11 years. Thus, the range was $11 - 5 = 6$ years. Let x be the age (in years) at which a color TV is replaced. Assume that x has a distribution that is approximately normal.
(a) The empirical rule (Section 6.1) indicates that for a symmetrical and bell-shaped distribution, approximately 95% of the data lies within two standard deviations of the mean. Therefore, a 95% range of data values extending from $\mu - 2\sigma$ to $\mu + 2\sigma$ is often used for "commonly occurring" data values. Note that the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$ is 4σ in length. This leads to a "rule of thumb" for estimating the standard deviation from a 95% range of data values.



Estimating the standard deviation

For a symmetric, bell-shaped distribution,

$$\textbf{standard deviation} \approx \frac{\text{range}}{4} \approx \frac{\text{high value} - \text{low value}}{4}$$

where it is estimated that about 95% of the commonly occurring data values fall into this range.

Use this "rule of thumb" to approximate the standard deviation of x values, where x is the age (in years) at which a color TV is replaced.
(b) What is the probability that someone will keep a color TV more than 5 years before replacement?

(c) What is the probability that someone will keep a color TV fewer than 10 years before replacement?

(d) ***Inverse Normal Distribution*** Assume that the average life of a color TV is 8 years with a standard deviation of 1.5 years before it breaks. Suppose that a company guarantees color TVs and will replace a TV that breaks while under guarantee with a new one. However, the company does not want to replace more than 10% of the TVs under guarantee. For how long should the guarantee be made (rounded to the nearest tenth of a year)?

32. ***Estimating the Standard Deviation: Refrigerator Replacement*** *Consumer Reports* indicated that the average life of a refrigerator before replacement is $\mu = 14$ years with a (95% of data) range from 9 to 19 years. Let x = age at which a refrigerator is replaced. Assume that x has a distribution that is approximately normal.

(a) Find a good approximation for the standard deviation of x values. *Hint:* See Problem 31.

(b) What is the probability that someone will keep a refrigerator fewer than 11 years before replacement?

(c) What is the probability that someone will keep a refrigerator more than 18 years before replacement?

(d) ***Inverse Normal Distribution*** Assume that the average life of a refrigerator is 14 years, with the standard deviation given in part (a) before it breaks. Suppose that a company guarantees refrigerators and will replace a refrigerator that breaks while under guarantee with a new one. However, the company does not want to replace more than 5% of the refrigerators under guarantee. For how long should the guarantee be made (rounded to the nearest tenth of a year)?

33. ***Estimating the Standard Deviation: Veterinary Science*** The resting heart rate for an adult horse should average about $\mu = 46$ beats per minute with a (95% of data) range from 22 to 70 beats per minute, based on information from the *Merck Veterinary Manual* (a classic reference used in most veterinary colleges). Let x be a random variable that represents the resting heart rate for an adult horse. Assume that x has a distribution that is approximately normal.

(a) Estimate the standard deviation of the x distribution. *Hint:* See Problem 31.

(b) What is the probability that the heart rate is fewer than 25 beats per minute?

(c) What is the probability that the heart rate is greater than 60 beats per minute?

(d) What is the probability that the heart rate is between 25 and 60 beats per minute?

(e) ***Inverse Normal Distribution*** A horse whose resting heart rate is in the upper 10% of the probability distribution of heart rates may have a secondary infection or illness that needs to be treated. What is the heart rate corresponding to the upper 10% cutoff point of the probability distribution?

34. ***Estimating the Standard Deviation: Veterinary Science*** How much should a healthy kitten weigh? A healthy 10-week-old (domestic) kitten should weigh an average of $\mu = 24.5$ ounces with a (95% of data) range from 14 to 35 ounces. (See reference in Problem 33.) Let x be a random variable that represents the weight (in ounces) of a healthy 10-week-old kitten. Assume that x has a distribution that is approximately normal.

(a) Estimate the standard deviation of the x distribution. *Hint:* See Problem 31.

(b) What is the probability that a healthy 10-week-old kitten will weigh less than 14 ounces?

(c) What is the probability that a healthy 10-week-old kitten will weigh more than 33 ounces?
(d) What is the probability that a healthy 10-week-old kitten will weigh between 14 and 33 ounces?
(e) ***Inverse Normal Distribution*** A kitten whose weight is in the bottom 10% of the probability distribution of weights is called *undernourished*. What is the cutoff point for the weight of an undernourished kitten?

35. ***Insurance: Satellites*** A relay microchip in a telecommunications satellite has a life expectancy that follows a normal distribution with a mean of 90 months and a standard deviation of 3.7 months. When this computer-relay microchip malfunctions, the entire satellite is useless. A large London insurance company is going to insure the satellite for $50 million. Assume that the only part of the satellite in question is the microchip. All other components will work indefinitely.


Max Dannenbaum/Stone/Getty Images

(a) ***Inverse Normal Distribution*** For how many months should the satellite be insured to be 99% confident that it will last beyond the insurance date?
(b) If the satellite is insured for 84 months, what is the probability that it will malfunction before the insurance coverage ends?
(c) If the satellite is insured for 84 months, what is the expected loss to the insurance company?
(d) If the insurance company charges $3 million for 84 months of insurance, how much profit does the company expect to make?

36. ***Convertion Center: Exhibition Show Attendance*** Attendance at large exhibition shows in Denver averages about 8000 people per day, with standard deviation of about 500. Assume that the daily attendance figures follow a normal distribution.
(a) What is the probability that the daily attendance will be fewer than 7200 people?
(b) What is the probability that the daily attendance will be more than 8900 people?
(c) What is the probability that the daily attendance will be between 7200 and 8900 people?

37. ***Exhibition Shows: Inverse Normal Distribution*** Most exhibition shows open in the morning and close in the late evening. A study of Saturday arrival times showed that the average arrival time was 3 hours and 48 minutes after the doors opened, and the standard deviation was estimated at about 52 minutes. Assume that the arrival times follow a normal distribution.
(a) At what time after the doors open will 90% of the people who are coming to the Saturday show have arrived?
(b) At what time after the doors open will only 15% of the people who are coming to the Saturday show have arrived?
(c) Do you think the probability distribution of arrival times for Friday might be different from the distribution of arrival times for Saturday? Explain.

38. ***Budget: Maintenance*** The amount of money spent weekly on cleaning, maintenance, and repairs at a large restaurant was observed over a long period of time to be approximately normally distributed, with mean $\mu = \$615$ and standard deviation $\sigma = \$42$.
(a) If $646 is budgeted for next week, what is the probability that the actual costs will exceed the budgeted amount?
(b) ***Inverse Normal Distribution*** How much should be budgeted for weekly repairs, cleaning, and maintenance so that the probability that the budgeted amount will be exceeded in a given week is only 0.10?

39. ***Expand Your Knowledge: Conditional Probability*** Suppose you want to eat lunch at a popular restaurant. The restaurant does not take reservations, so there is usually a waiting time before you can be seated. Let x represent the length of time waiting to be seated. From past experience, you know that the mean waiting time is $\mu = 18$ minutes with $\sigma = 4$ minutes. You assume that the x distribution is approximately normal.
(a) What is the probability that the waiting time will *exceed* 20 minutes, given that it has exceeded 15 minutes? *Hint:* Compute $P(x > 20 \mid x > 15)$.
(b) What is the probability that the waiting time will exceed 25 minutes, given that it has exceeded 18 minutes? *Hint:* Compute $P(x > 25 \mid x > 18)$.
(c) *Hint for solution:* Review item 6, conditional probability, in the summary of basic probability rules at the end of Section 4.2. Note that
$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$
and show that in part (a),
$$P(x > 20 \mid x > 15) = \frac{P((x > 20) \text{ and } (x > 15))}{P(x > 15)} = \frac{P(x > 20)}{P(x > 15)}$$


AP Photo/Judi Bottoni

40. ***Conditional Probability: Cycle Time*** A cement truck delivers mixed cement to a large construction site. Let x represent the cycle time in minutes for the truck to leave the construction site, go back to the cement plant, fill up, and return to the construction site with another load of cement. From past experience, it is known that the mean cycle time is $\mu = 45$ minutes with $\sigma = 12$ minutes. The x distribution is approximately normal.
(a) What is the probability that the cycle time will *exceed* 60 minutes, given that it has exceeded 50 minutes? *Hint:* See Problem 39, part (c).
(b) What is the probability that the cycle time will exceed 55 minutes, given that it has exceeded 40 minutes?

SECTION 6.4

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 complete 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.*

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.

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

Population parameter

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 11 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 11 SAMPLING DISTRIBUTION FOR $\bar{x}$


Philip James Corwin/Documentary/Corbis

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 6-9, 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 6-9. But just looking at 100 of the $\bar{x}$ values doesn't tell us much.

TABLE 6-9 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 6-10 Frequency Table for 100 Values of $\bar{x}$

Class	Class Limits 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

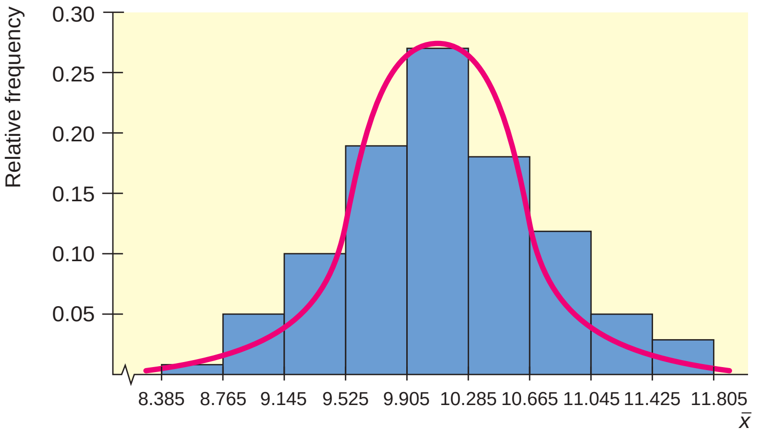
Let's organize our $\bar{x}$ values into a frequency table. We used a class width of 0.38 to make Table 6-10.

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 6-10 contains relative frequencies f/100. Recall that 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 (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 6-32 is a relative-frequency or probability distribution of the $\bar{x}$ values.

FIGURE 6-32

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



The bars of Figure 6-32 represent our estimated probabilities of $\bar{x}$ values based on the data of Table 6-9. 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 6-32 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 6-32, 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*.

There are other sampling distributions besides the $\bar{x}$ distribution. Section 6.6 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 7 and 8 on estimation and testing.

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

GUIDED EXERCISE 9 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 based on all possible samples of the same size.

(d) In Table 6-9, 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 6-32. 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 7 will concentrate on a method of inference called *estimation*. Chapter 8 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, styles, and insights. Ansel Adams became famous for photographing lyrical, unforgettable land scapes such as "Moonrise, Hernandez, New Mexico." Adams claimed that he was a strong believer in the quote by Pasteur. In fact, he claimed 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 6.4 PROBLEMS

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. *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 6.5

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 6.4, 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 6.1 **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 $\sigma/\sqrt{n}$.

We conclude from Theorem 6.1 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.

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

where n is the sample size,
μ is the mean of the x distribution, and
σ is the standard deviation of the x distribution.

Theorem 6.1 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 other (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 6.1.

EXAMPLE 12 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 Section 6.3, with $\mu = 10.2$ and $\sigma = 1.4$, to get

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

Therefore,

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

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

appletat/iStockphoto.com

(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 6.1, part (b), tells us that

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

If $\sigma_{\bar{x}}$ represents the standard deviation of the $\sigma_{\bar{x}}$ distribution, then Theorem 6.1, 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 6.1, 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 6.1, 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

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

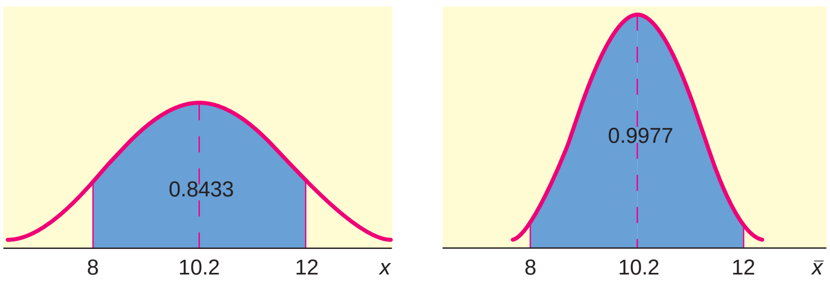
The standard deviation of the $\bar{x}$ distribution is less than half the standard deviation of the x distribution. Figure 6-33 shows the distributions of x and $\bar{x}$.

FIGURE 6-33

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

(a) The x distribution with $\mu = 10.2$ and $\sigma = 1.4$

(b) The $\bar{x}$ distribution with $\mu_{\bar{x}} = 10.2$ and $\sigma_{\bar{x}} = 0.63$ for samples of size $n = 5$



Looking at Figure 6-33(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 6.1 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.

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 6.1 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.

THEOREM 6.2 **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.

The central limit theorem is indeed surprising! It says that x can have *any* distribution whatsoever, but that 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.

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

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

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

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

Continued

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 10 shows how to standardize when appropriate. Then, Example 13 demonstrates the use of the central limit theorem in a decision-making process.

GUIDED EXERCISE 10 Central limit theorem

(a) Suppose x has a *normal* distribution with mean $\mu = 18$ and standard deviation $\sigma = 13$. 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 13 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

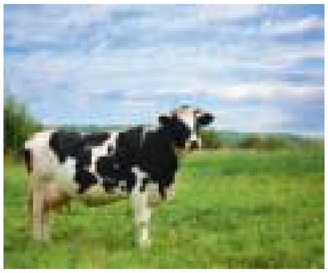
$$\mu_{\bar{x}} = \mu = 2500$$
$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 300/\sqrt{42} \approx 46.3$$

(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.



Aleksas Kvedoras/iStockphoto.com

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

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

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

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 for milk that is not contaminated 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.

In Problems 21, 22, and 23, we'll apply the central limit theorem to solve problems involving a *sum* of random variables.

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

Jupiter Images

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 $z \approx \frac{11 - 12.1}{0.54} = -2.04$

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

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 6.6, 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.

LOOKING FORWARD

Sampling distributions for the mean $\bar{x}$ and for proportions $\hat{p}$ (Section 6.6) will form the basis of our work with estimation (Chapter 7) and hypothesis testing (Chapter 8). With these inferential statistics methods, we will be able to use information from a sample to make statements regarding the population. These statements will be made in terms of probabilities derived from the underlying sampling distributions.

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 Create Havoc Around Our System 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 some thing 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 6.5 PROBLEMS

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. *Basic Computation: Central Limit Theorem* Suppose x has a distribution with a mean of 8 and a standard deviation of 16. Random samples of size $n = 64$ are drawn.

(a) Describe the $\bar{x}$ distribution and compute the mean and standard deviation of the distribution.
(b) Find the z value corresponding to $\bar{x} = 9$.
(c) Find $P(\bar{x} > 9)$.
(d) *Interpretation* Would it be unusual for a random sample of size 64 from the x distribution to have a sample mean greater than 9? Explain.

6. *Basic Computation: Central Limit Theorem* Suppose x has a distribution with a mean of 20 and a standard deviation of 3. Random samples of size $n = 36$ are drawn.
(a) Describe the $\bar{x}$ distribution and compute the mean and standard deviation of the distribution.
(b) Find the z value corresponding to $\bar{x} = 19$.
(c) Find $P(\bar{x} < 19)$.
(d) *Interpretation* Would it be unusual for a random sample of size 36 from the x distribution to have a sample mean less than 19? Explain.

7. *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?

8. *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 $P(68 \le \bar{x} \le 73)$.

9. *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.

10. *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?

11. *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).

12. *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.

13. *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?

14. ***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?

15. ***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 6.1. 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.

16. ***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 15). 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?

17. ***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*

Steve Krull/iStockphoto.com

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.

18. ***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?

19. ***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.

20. ***Finance: European Growth Fund*** A european growth mutual fund 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 19), x has mean $\mu = 1.4\%$ and standard deviation $\sigma = 0.8\%$.

Mark L. Stephenson/CORBIS

(a) Let's consider the monthly return of the stocks in the European growth 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 European growth fund) has a distribution that is approximately normal? Explain. *Hint:* See Problem 19, 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 6.1 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.

21. ***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 $P(w < 90)$?

22. ***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 another, 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 21.

23. ***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.)

Vladimir Wrangel,2009/ Used under license from Shutterstock.com

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 21 and Theorem 6.1.

SECTION 6.6 Normal Approximation to Binomial Distribution and to $\hat{p}$ Distribution

FOCUS POINTS

- State the assumptions needed to use the normal approximation to the binomial distribution.
- Compute μ and σ for the normal approximation.
- Use the continuity correction to convert a range of r values to a corresponding range of normal x values.
- Convert the x values to a range of standardized z scores and find desired probabilities.
- Describe the sampling distribution for proportions $\hat{p}$.

The probability that a new vaccine will protect adults from cholera is known to be 0.85. The vaccine is administered to 300 adults who must enter an area where the disease is prevalent. What is the probability that more than 280 of these adults will be protected from cholera by the vaccine?

This question falls into the category of a binomial experiment, with the number of trials n equal to 300, the probability of success p equal to 0.85, and the number of successes r greater than 280. It is possible to use the formula for the binomial distribution to compute the probability that r is greater than 280. However, this approach would involve a number of tedious and long calculations. There is an easier way to do this problem, for under the conditions stated below, the normal distribution can be used to approximate the binomial distribution.

Criteria $np > 5$ and $nq > 5$

Normal approximation to the binomial distribution

Consider a binomial distribution where

n = number of trials

r = number of successes

p = probability of success on a single trial

$q = 1 - p$ = probability of failure on a single trail

If $np > 5$ and $nq > 5$, then r has a binomial distribution that is approximated by a normal distribution with

$\mu = np$ and $\sigma = \sqrt{npq}$

Note: As n increases, the approximation becomes better.

Example 14 demonstrates that as n increases, the normal approximation to the binomial distribution improves.

EXAMPLE 14 Binomial distribution graphs

Graph the binomial distributions for which $p = 0.25$, $q = 0.75$, and the number of trials is first $n = 3$, then $n = 10$, then $n = 25$, and finally $n = 50$.

SOLUTION: The authors used a computer program to obtain the binomial distributions for the given values of p, q, and n. The results have been organized and graphed in Figures 6-34, 6-35, 6-36, and 6-37.

FIGURE 6-34

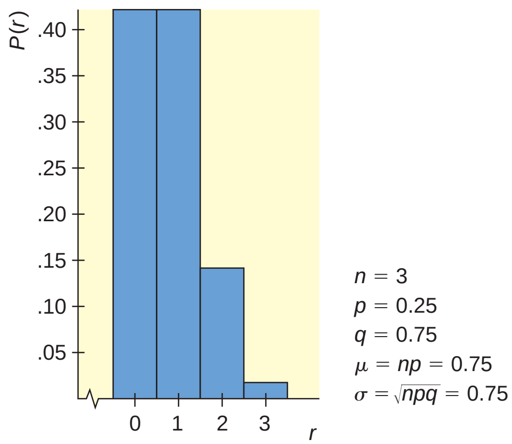


FIGURE 6-35

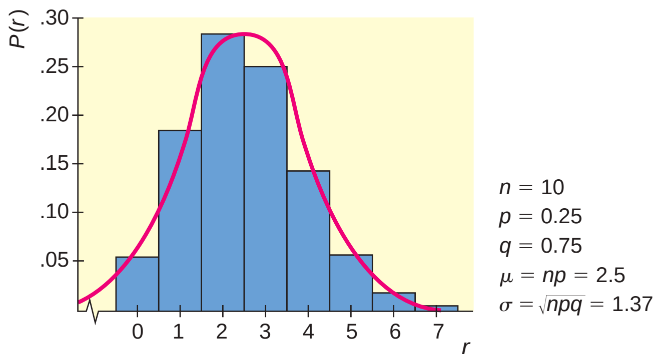


FIGURE 6-36

Good Normal Approximation; $np > 5$ and $nq > 5$

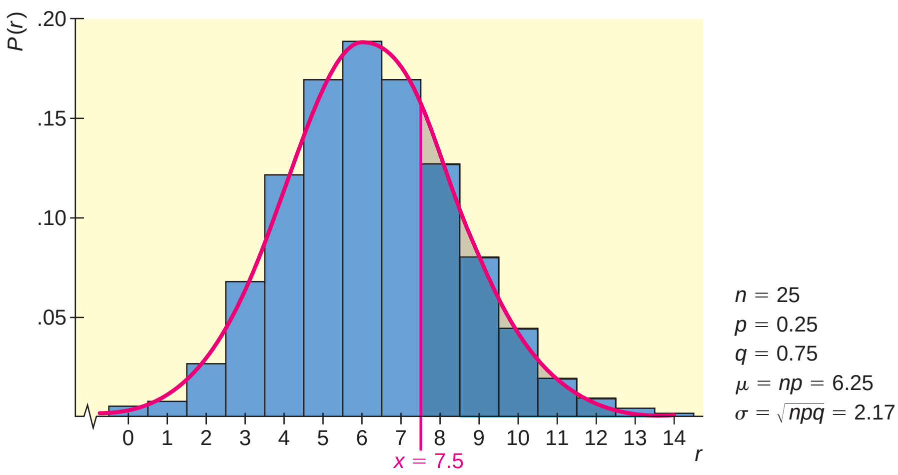
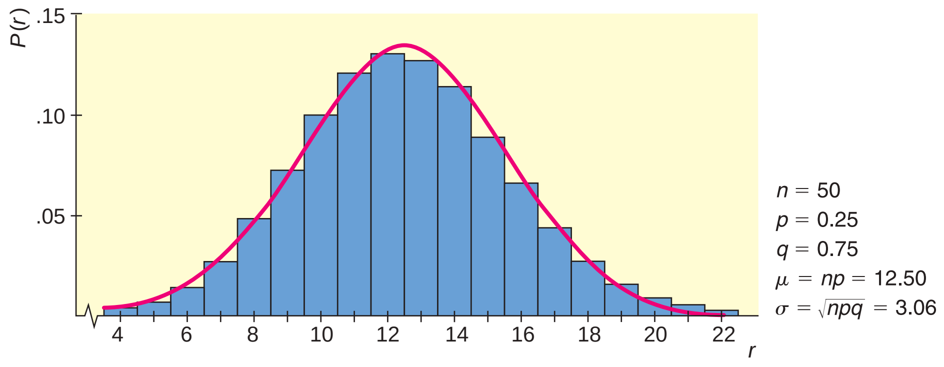


FIGURE 6-37

Good Normal Approximation; $np > 5$ and $nq > 5$



When $n = 3$, the outline of the histogram does not even begin to take the shape of a normal curve. But when $n = 10$, 25, or 50, it does begin to take a normal shape, indicated by the red curve. From a theoretical point of view, the histograms in Figures 6-35, 6-36, and 6-37 would have bars for all values of r from $r = 0$ to $r = n$. However, in the construction of these histograms, the bars of height less than 0.001 unit have been omitted—that is, in this example, probabilities less than 0.001 have been rounded to 0.

EXAMPLE 15 NORMAL APPROXIMATION

The owner of a new apartment building must install 25 water heaters. From past experience in other apartment buildings, she knows that Quick Hot is a good brand. A Quick Hot heater is guaranteed for 5 years only, but from the owner's past experience, she knows that the probability it will last 10 years is 0.25.

(a) What is the probability that 8 or more of the 25 water heaters will last at least 10 years? Define success to mean a water heater that lasts at least 10 years.

SOLUTION: In this example, $n = 25$ and $p = 0.25$, so Figure 6-36 (on the preceding page) represents the probability distribution we will use. Let r be the binomial random variable corresponding to the number of successes out of $n = 25$ trials. We want to find $P(r \geq 8)$ by using the normal approximation. This probability is represented graphically (Figure 6-36) by the area of the bar over 8 plus the areas of all bars to the right of the bar over 8.

Let x be a normal random variable corresponding to a normal distribution, with $\mu = np = 25(0.25) = 6.25$ and $\sigma = \sqrt{npq} = \sqrt{25(0.25)(0.75)} \approx 2.17$. This normal curve is represented by the red line in Figure 6-36. The area under the normal curve from $x = 7.5$ to the right is approximately the same as the areas of the bars from the bar over $r = 8$ to the right. It is important to notice that we start with $x = 7.5$ because the bar over $r = 8$ really starts at $x = 7.5$.

Marcus Lindstrøm/ iStockphoto.com

The areas of the bars and the area under the corresponding red (normal) curve are approximately equal, so we conclude that $P(r \geq 8)$ is approximately equal to $P(x \geq 7.5)$.

When we convert $x = 7.5$ to standard units, we get

$$z = \frac{x - \mu}{\sigma} = \frac{7.5 - 6.25}{2.17} \approx 0.58 \qquad \text{(Use } \mu = 6.25 \text{ and } \sigma = 2.17.\text{)}$$

The probability we want is

$$P(x \geq 7.5) = P(z \geq 0.58) = 1 - P(z \leq 0.58) = 1 - 0.7190 = 0.2810$$

(b) How does this result compare with the result we can obtain by using the formula for the binomial probability distribution with $n = 25$ and $p = 0.25$?

SOLUTION: Using the binomial distribution function on the TI-84Plus/TI-83Plus/TI-*n*spire model calculators, the authors computed that

$P(r \geq 8) \approx 0.2735$. This means that the probability is approximately 0.27 that 8 or more water heaters will last at least 10 years.

(c) How do the results of parts (a) and (b) compare?

SOLUTION: The error of approximation is the difference between the approximate normal value (0.2810) and the binomial value (0.2735). The error is only $0.2810 - 0.2735 = 0.0075$, which is negligible for most practical purposes.

We knew in advance that the normal approximation to the binomial probability would be good, since $np = 25(0.25) = 6.25$ and $nq = 25(0.75) = 18.75$ are both greater than 5. These are the conditions that assure us that the normal approximation will be sufficiently close to the binomial probability for most practical purposes.

Continuity correction: Converting r values to x values

Remember that when we use the normal distribution to approximate the binomial, we are computing the areas under bars. The bar over the discrete variable r extends from $r - 0.5$ to $r + 0.5$. This means that the corresponding continuous normal variable x extends from $r - 0.5$ to $r + 0.5$. Adjusting the values of discrete random variables to obtain a corresponding range for a continuous random variable is called making a *continuity correction.*

PROCEDURE

How to make the continuity correction

Convert the discrete random variable r (number of successes) to the continuous normal random variable x by doing the following:

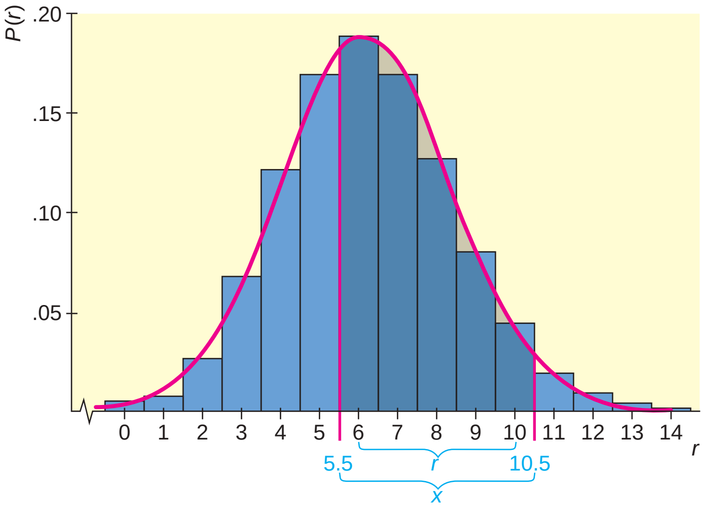
1. If r is a **left point** of an interval, subtract 0.5 to obtain the corresponding normal variable x; that is, $x = r - 0.5$.
2. If r is a **right point** of an interval, add 0.5 to obtain the corresponding normal variable x; that is, $x = r + 0.5$.

For instance, $P(6 \leq r \leq 10)$, where r is a binomial random variable, is approximated by $P(5.5 \leq x \leq 10.5)$, where x is the corresponding normal random variable (see Figure 6-38).

COMMENT Both the binomial and Poisson distributions are for *discrete* random variables. Therefore, adding or subtracting 0.5 to r was not necessary when we approximated the binomial distribution by the Poisson distribution (Section 5.4). However, the normal distribution is for a *continuous* random variable. In this case, adding or subtracting 0.5 to or from (as appropriate) r will improve the approximation of the normal to the binomial distribution.

FIGURE 6-38

$P(6 \leq r \leq 10)$ Is Approximately Equal to $P(5.5 \leq x \leq 10.5)$



GUIDED EXERCISE 12 ***Continuity correction***

From many years of observation, a biologist knows that the probability is only 0.65 that any given Arctic tern will survive the migration from its summer nesting area to its winter feeding grounds. A random sample of 500 Arctic terns were banded at their summer nesting area. Use the normal approximation to the binomial and the following steps to find the probability that between 310 and 340 of the banded Arctic terns will survive the migration. Let r be the number of surviving terns.

William Mullins/Photo Researchers

Arctic tern

(a) To approximate $P(310 \leq r \leq 340)$, we use the normal curve with $\mu =$ ________ and $\sigma =$ ________.

➡ We use the normal curve with

$\mu = np = 500(0.65) = 325$ and

$\sigma = \sqrt{npq} = \sqrt{500(0.65)(0.35)} \approx 10.67$

(b) $P(310 \leq r \leq 340)$ is approximately equal to $P($ ________ $\leq x \leq$ ________$)$, where x is a variable from the normal distribution described in part (a).

➡ Since 310 is the left endpoint, we subtract 0.5, and since 340 is the right endpoint, we add 0.5. Consequently,

$P(310 \leq r \leq 340) \approx P(309.5 \leq x \leq 340.5)$

(c) Convert the condition $309.5 \leq x \leq 340.5$ to a condition in standard units.

➡ Since $\mu = 325$ and $\sigma \approx 10.67$, the condition $309.5 \leq x \leq 340.5$ becomes

$$\frac{309.5 - 325}{10.67} \leq z \leq \frac{340.5 - 325}{10.67}$$

or

$$-1.45 \leq z \leq 1.45$$

(d) $P(310 \leq r \leq 340) = P(309.5 \leq x \leq 340.5)$
$= P(-1.45 \leq z \leq 1.45)$
$=$ ________

➡ $P(-1.45 \leq z \leq 1.45) = P(z \leq 1.45) - P(z \leq -1.45)$
$= 0.9265 - 0.0735$
$= 0.8530$

(e) Will the normal distribution make a good approximation to the binomial for this problem? Explain your answer.

➡ Since

$np = 500(0.65) = 35$

and

$nq = 500(0.35) = 175$

are both greater than 5, the normal distribution will be a good approximation to the binomial.

Sampling Distributions for Proportions

In Sections 6.4 and 6.5, we studied the sampling distribution for the mean. Now we have the tools to look at sampling distributions for proportions. Suppose we repeat a binomial experiment with n trials again and again and, for each n trials, record the sample proportion of successes $\hat{p} = r/n$. The $\hat{p}$ values form a sampling distribution for proportions.

Sampling distribution for $\hat{p}$

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

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

Standard error of a proportion

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. Earlier 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 $\mu_{\hat{p}} = u_r/n = np/n = p$ and $\sigma_{\hat{p}} = \sigma_r/n = \sqrt{npq}/n = \sqrt{pq/n}$. Although the values r/n are discrete for a fixed n, we do not use a continuity correction for the $\hat{p}$ distribution. This is the accepted standard practice for applications in inferential statistics, especially considering the requirements that $np > 5$ and $nq >$ are met.

We see from the sampling distribution for proportions that the mean of the $\hat{p}$ distribution is p, the population proportion of successes. This means that $\hat{p}$ is an *unbiased* estimator for p.

LOOKING FORWARD

We will use the sampling distribution for proportions in our work with estimation (Chapter 7) and hypothesis testing (Chapter 8).

EXAMPLE 16 SAMPLING DISTRIBUTION OF $\hat{p}$

The annual crime rate in the Capital 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 **http://www.cengage.com/statistics/brase** 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.

Beyza Sultan Durna/iStockphoto.com

(a) What is the population probability p that a resident in the Capital 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

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

(b) Consider the random variable

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

Can we approximate the $\hat{p}$ distribution 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 the $\hat{p}$ distribution with a normal distribution.

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

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

VIEWPOINT Sunspots, Tree Rings, and Statistics

Ancient Chinese astronomers recorded extreme sunspot activity, with a peak around 1200 A.D. Mesa Verde tree rings in the period between 1276 and 1299 were unusually narrow, indicating a drought and/or a severe cold spell in the region at that time. A cooling trend could have narrowed the window of frost-free days below the approximately 80 days needed for cultivation of aboriginal corn and beans. Is this the reason the ancient Anasazi dwellings in Mesa Verde were abandoned? Is there a connection to the extreme sunspot activity? Much research and statistical work continue to be done on this topic.

Reference: *Prehistoric Astronomy in the Southwest,* by J. McKim Malville and C. Putnam, Department of Astronomy, University of Colorado.

SECTION 6.6 PROBLEMS

Note: When we say *between a and b,* we mean every value from a to b, *including a* and b. Due to rounding, your answers might vary slightly from answers given in the text.

1. *Statistical Literacy* Binomial probability distributions depend on the number of trials n of a binomial experiment and the probability of success p on each trial. Under what conditions is it appropriate to use a normal approximation to the binomial?

2. *Statistical Literacy* When we use a normal distribution to approximate a binomial distribution, why do we make a continuity correction?

3. *Basic Computation: Normal Approximation to a Binomial Distribution* Suppose we have a binomial experiment with $n = 40$ trials and a probability of success $p = 0.50$.

(a) Is it appropriate to use a normal approximation to this binomial distribution? Why?
(b) Compute μ and σ of the approximating normal distribution.
(c) Use a continuity correction factor to convert the statement $r \geq 23$ successes to a statement about the corresponding normal variable x.
(d) Estimate $P(r \geq 23)$.
(e) *Interpretation* Is it unusual for a binomial experiment with 40 trials and probability of success 0.50 to have 23 or more successes? Explain.

4. *Basic Computation: Normal Approximation to a Binomial Distribution* Suppose we have a binomial experiment with $n = 40$ trials and probability of success $p = 0.85$.
(a) Is it appropriate to use a normal approximation to this binomial distribution? Why?
(b) Compute μ and σ of the approximating normal distribution.
(c) Use a continuity correction factor to convert the statement $r < 30$ successes to a statement about the corresponding normal variable x.
(d) Estimate $P(r < 30)$.
(e) *Interpretation* Is it unusual for a binomial experiment with 40 trials and probability of success 0.85 to have fewer than 30 successes? Explain.

5. *Critical Thinking* You need to compute the probability of 5 or fewer successes for a binomial experiment with 10 trials. The probability of success on a single trial is 0.43. Since this probability of success is not in the table, you decide to use the normal approximation to the binomial. Is this an appropriate strategy? Explain.

6. *Critical Thinking* Consider a binomial experiment with 20 trials and probability 0.45 of success on a single trial.
(a) Use the binomial distribution to find the probability of exactly 10 successes.
(b) Use the normal distribution to approximate the probability of exactly 10 successes.
(c) Compare the results of parts (a) and (b).

In the following problems, check that it is appropriate to use the normal approximation to the binomial. Then use the normal distribution to estimate the requested probabilities.

7. *Health: Lead Contamination* More than a decade ago, high levels of lead in the blood put 88% of children at risk. A concerted effort was made to remove lead from the environment. Now, according to the *Third National Health and Nutrition Examination Survey* (*NHANES III*) conducted by the Centers for Disease Control, only 9% of children in the United States are at risk of high blood-lead levels.
(a) In a random sample of 200 children taken more than a decade ago, what is the probability that 50 or more had high blood-lead levels?
(b) In a random sample of 200 children taken now, what is the probability that 50 or more have high blood-lead levels?

8. *Insurance: Claims* Do you try to *pad* an insurance claim to cover your deductible? About 40% of all U.S. adults will try to pad their insurance claims! (Source: *Are You Normal*?, by Bernice Kanner, St. Martin's Press.) Suppose that you are the director of an insurance adjustment office. Your office has just received 128 insurance claims to be processed in the next few days. What is the probability that
(a) half or more of the claims have been padded?
(b) fewer than 45 of the claims have been padded?
(c) from 40 to 64 of the claims have been padded?
(d) more than 80 of the claims have *not* been padded?

9. *Longevity: 90th Birthday* It is estimated that 3.5% of the general population will live past their 90th birthday (*Statistical Abstract of the United States*, 112th Edition). In a graduating class of 753 high school seniors, what is the probability that

(a) 15 or more will live beyond their 90th birthday?
(b) 30 or more will live beyond their 90th birthday?
(c) between 25 and 35 will live beyond their 90th birthday?
(d) more than 40 will live beyond their 90th birthday?

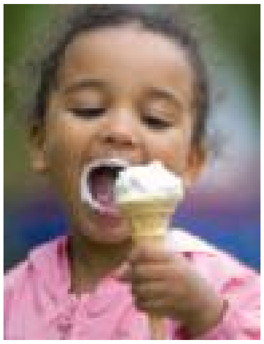
10. *Fishing: Billfish* Ocean fishing for billfish is very popular in the Cozumel region of Mexico. In *World Record Game Fishes* (published by the International Game Fish Association), it was stated that in the Cozumel region, about 44% of strikes (while trolling) result in a catch. Suppose that on a given day a fleet of fishing boats got a total of 24 strikes. What is the probability that the number of fish caught was
(a) 12 or fewer?
(b) 5 or more?
(c) between 5 and 12?

11. *Grocery Stores: New Products* *The Denver Post* stated that 80% of all new products introduced in grocery stores fail (are taken off the market) within 2 years. If a grocery store chain introduces 66 new products, what is the probability that within 2 years
(a) 47 or more fail?
(b) 58 or fewer fail?
(c) 15 or more succeed?
(d) fewer than 10 succeed?

12. *Crime: Murder* What are the chances that a person who is murdered actually knew the murderer? The answer to this question explains why a lot of police detective work begins with relatives and friends of the victim! About 64% of people who are murdered actually knew the person who committed the murder (*Chances: Risk and Odds in Everyday Life,* by James Burke). Suppose that a detective file in New Orleans has 63 current unsolved murders. What is the probability that
(a) at least 35 of the victims knew their murderers?
(b) at most 48 of the victims knew their murderers?
(c) fewer than 30 victims did *not* know their murderers?
(d) more than 20 victims did *not* know their murderers?

13. *Supermarkets: Free Samples* Do you take the free samples offered in supermarkets? About 60% of all customers will take free samples. Furthermore, of those who take the free samples, about 37% will buy what they have sampled. (See reference in Problem 8.) Suppose you set up a counter in a supermarket offering free samples of a new product. The day you are offering free samples, 317 customers pass by your counter.
(a) What is the probability that more than 180 take your free sample?
(b) What is the probability that fewer than 200 take your free sample?
(c) What is the probability that a customer takes a free sample *and* buys the product? *Hint:* Use the multiplication rule for *dependent* events. Notice that we are given the conditional probability $P(\text{buy} \mid \text{sample}) = 0.37$, while $P(\text{sample}) = 0.60$.
(d) What is the probability that between 60 and 80 customers will take the free sample *and* buy the product? *Hint:* Use the probability of success calculated in part (c).

14. *Ice Cream: Flavors* What's your favorite ice cream flavor? For people who buy ice cream, the all-time favorite is still vanilla. About 25% of ice cream sales are vanilla. Chocolate accounts for only 9% of ice cream sales. (See reference in Problem 8.) Suppose that 175 customers go to a grocery store in Cheyenne, Wyoming, today to buy ice cream.
(a) What is the probability that 50 or more will buy vanilla?
(b) What is the probability that 12 or more will buy chocolate?



Darren Baker, 2009. Used under permission from Shutterstock.com

(c) A customer who buys ice cream is not limited to one container or one flavor. What is the probability that someone who is buying ice cream will buy chocolate or vanilla? *Hint:* Chocolate flavor and vanilla flavor are not mutually exclusive events. Assume that the choice to buy one flavor is independent of the choice to buy another flavor. Then use the multiplication rule for independent events, together with the addition rule for events that are not mutually exclusive, to compute the requested probability. (See Section 4.2.)

(d) What is the probability that between 50 and 60 customers will buy chocolate or vanilla ice cream? *Hint:* Use the probability of success computed in part (c).

15. ***Airline Flights: No-Shows*** Based on long experience, an airline has found that about 6% of the people making reservations on a flight from Miami to Denver do not show up for the flight. Suppose the airline overbooks this flight by selling 267 ticket reservations for an airplane with only 255 seats.

(a) What is the probability that a person holding a reservation will show up for the flight?

(b) Let $n = 267$ represent the number of ticket reservations. Let r represent the number of people with reservations who show up for the flight. Which expression represents the probability that a seat will be available for everyone who shows up holding a reservation?

$$P(255 \le r); \quad P(r \le 255); \quad P(r \le 267); \quad P(r = 255)$$

(c) Use the normal approximation to the binomial distribution and part (b) to answer the following question: What is the probability that a seat will be available for every person who shows up holding a reservation?

16. ***General: Approximations*** We have studied *two* approximations to the binomial, the normal approximation and the Poisson approximation (Section 5.4). Write a brief but complete essay in which you discuss and summarize the *conditions* under which each approximation would be used, the *formulas* involved, and the *assumptions* made for each approximation. Give details and examples in your essay. How could you apply these statistical methods in your everyday life?

17. ***Statistical Literacy*** Under what conditions is it appropriate to use a normal distribution to approximate the $\hat{p}$ distribution?

18. ***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?

19. ***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.

20. ***Basic Computation: $\hat{p}$ Distribution*** 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 the $\hat{p}$ distribution by a normal distribution? Why? What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

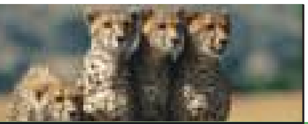
(b) Suppose $n = 25$ and $p = 0.15$. Can we safely approximate the $\hat{p}$ distribution by a normal distribution? Why or why not?

(c) Suppose $n = 48$ and $p = 0.15$. Can we approximate the $\hat{p}$ distribution by a normal distribution? Why? What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

21. ***Basic Computation: $\hat{p}$ Distribution*** Suppose we have a binomial experiment in which success is defined to be a particular quality or attribute that interests us.

(a) Suppose $n = 100$ and $p = 0.23$. Can we safely approximate the $\hat{p}$ distribution by a normal distribution? Why? Compute $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.

(b) Suppose $n = 20$ and $p = 0.23$. Can we safely approximate the $\hat{p}$ distribution by a normal distribution? Why or why not?



Chapter Review

SUMMARY

In this chapter, we examined properties and applications of the normal probability distribution.

- A normal probability distribution is a distribution of a continuous random variable. Normal distributions are bell-shaped and symmetric around the mean. The high point occurs over the mean, and most of the area occurs within 3 standard deviations of the mean. The mean and median are equal.
- The empirical rule for normal distributions gives areas within 1, 2, and 3 standard deviations of the mean. Approximately

  68% of the data lie within the interval $\mu \pm \sigma$

  95% of the data lie within the interval $\mu \pm 2\sigma$

  99.7% of the data lie within the interval $\mu \pm 3\sigma$
- For symmetric, bell-shaped distributions,

  $$\text{standard deviation} \approx \frac{\text{range of data}}{4}$$
- A z score measures the number of standard deviations a raw score x lies from the mean.

  $$z = \frac{x - \mu}{\sigma} \quad \text{and} \quad x = z\sigma + \mu$$
- For the standard normal distribution, $\mu = 0$ and $\sigma = 1$.
- Table 5 of Appendix II gives areas under a standard normal distribution that are to the left of a specified value of z.
- After raw scores x have been converted to z scores, the standard normal distribution table can be used to find probabilities associated with intervals of x values from any normal distribution.
- The inverse normal distribution is used to find z values associated with areas to the left of z. Table 5 of Appendix II can be used to find approximate z values associated with specific probabilities.
- Tools for assessing the normality of a data distribution include:

  Histogram of the data. A roughly bell-shaped histogram indicates normality.

  Presence of outliers. A limited number indicates normality.

  Skewness. For normality, Pearson's index is between -1 and 1.

  Normal quantile plot. For normality, points lie close to a straight line.
- Control charts are an important application of normal distributions.
- 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

  $$\mu_{\hat{p}} = p$$
  $$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$
- The binomial distribution can be approximated by a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$ provided

  $$np > 5 \text{ and } nq > 5, \text{ with } q = 1 - p$$

  and a continuity correction is made.

Data from many applications follow distributions that are approximately normal. We will see normal distributions used extensively in later chapters.

IMPORTANT WORDS & SYMBOLS

Section 6.1
Normal distributions 250
Normal curves 251
Downward cup and upward cup on normal curves 251
Symmetry of normal curves 251
Normal density function 252
Empirical rule 252
Control chart 255
Out-of-control signals 256
Uniform distribution 263
Exponential distribution 264

Section 6.2
z value, z score or standard score 267
Standard units 267
Raw score, x 268
Standard normal distribution ($\mu = 0$ and $\sigma = 1$) 269
Area under the standard normal curve 269

Section 6.3
Areas under any normal curve 277
Inverse normal distribution 279
Normality indicators 283

Section 6.4
Statistic 291
Population parameter 291
Sampling distribution 292

Section 6.5
$\mu_{\bar{x}}$ 297
$\sigma_{\bar{x}}$ 297
Standard error of the mean 298
Central limit theorem 299
Large sample 299

Section 6.6
Normal approximation to the binomial distribution 309
Continuity correction 311
Sampling distribution for $\hat{p}$ 313
$\mu_{\hat{p}}$ 313
$\sigma_{\hat{p}}$ 313
Standard error of a proportion 313

VIEWPOINT Nenana Ice Classic

The Nenana Ice Classic is a betting pool offering a large cash prize to the lucky winner who can guess the time, to the nearest minute, of the ice breakup on the Tanana River in the town of Nenana, Alaska. Official breakup time is defined as the time when the surging river dislodges a tripod on the ice. This breaks an attached line and stops a clock set to Yukon Standard Time. The event is so popular that the first state legislature of Alaska (1959) made the Nenana Ice Classic an official statewide lottery. Since 1918, the earliest breakup has been April 20, 1940, at 3:27 P.M., and the latest recorded breakup was May 20, 1964, at 11:41 A.M. Want to make a statistical guess predicting when the ice will break up? Breakup times from the years 1918 to 1996 are recorded in The Alaska Almanac, *published by Alaska Northwest Books, Anchorage.*

CHAPTER REVIEW PROBLEMS

1. *Statistical Literacy* Describe a normal probability distribution.
2. *Statistical Literacy* According to the empirical rule, approximately what percentage of the area under a normal distribution lies within 1 standard deviation of the mean? within 2 standard deviations? within 3 standard deviations?
3. *Statistical Literacy* Is a process in control if the corresponding control chart for data having a normal distribution shows a value beyond 3 standard deviations of the mean?
4. *Statistical Literacy* Can a normal distribution always be used to approximate a binomial distribution? Explain.
5. *Statistical Literacy* What characteristic of a normal quantile plot indicates that the data follow a distribution that is approximately normal?

6. *Statistical Literacy* For a normal distribution, is it likely that a data value selected at random is more than 2 standard deviations above the mean?

7. *Statistical Literacy* Give the formula for the *standard error* of the sample mean $\bar{x}$ distribution, based on samples of size n from a distribution with standard deviation σ.

8. *Statistical Literacy* Give the formula for the *standard error* of the sample proportion $\hat{p}$ distribution, based on n binomial trials with probability of success p on each trial.

9. *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?

10. *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?

11. *Basic Computation: Probability* Given that x is a normal variable with mean $\mu = 47$ and standard deviation $\sigma = 6.2$, find
    (a) $P(x \leq 60)$ (b) $P(x \geq 50)$ (c) $P(50 \leq x \leq 60)$

12. *Basic Computation: Probability* Given that x is a normal variable with mean $\mu = 110$ and standard deviation $\sigma = 12$, find
    (a) $P(x \leq 120)$ (b) $P(x \geq 80)$ (c) $P(108 \leq x \leq 117)$

13. *Basic Computation: Inverse Normal* Find z such that 5% of the area under the standard normal curve lies to the right of z.

14. *Basic Computation: Inverse Normal* Find z such that 99% of the area under the standard normal curve lies between $-z$ and z.

15. *Medical: Blood Type* Blood type AB is found in only 3% of the population (*Textbook of Medical Physiology,* by A. Guyton, M.D.). If 250 people are chosen at random, what is the probability that
    (a) 5 or more will have this blood type?
    (b) between 5 and 10 will have this blood type?

16. *Customer Complaints: Time* The Customer Service Center in a large New York department store has determined that the amount of time spent with a customer about a complaint is normally distributed, with a mean of 9.3 minutes and a standard deviation of 2.5 minutes. What is the probability that for a randomly chosen customer with a complaint, the amount of time spent resolving the complaint will be
    (a) less than 10 minutes?
    (b) longer than 5 minutes?
    (c) between 8 and 15 minutes?

17. *Recycling: Aluminum Cans* One environmental group did a study of recycling habits in a California community. It found that 70% of the aluminum cans sold in the area were recycled.
    (a) If 400 cans are sold today, what is the probability that 300 or more will be recycled?

(b) Of the 400 cans sold, what is the probability that between 260 and 300 will be recycled?

18. ***Guarantee: Disc Players*** Future Electronics makes compact disc players. Its research department found that the life of the laser beam device is normally distributed, with mean 5000 hours and standard deviation 450 hours.
(a) Find the probability that the laser beam device will wear out in 5000 hours or less.
(b) ***Inverse Normal Distribution*** Future Electronics wants to place a guarantee on the players so that no more than 5% fail during the guarantee period. Because the laser pickup is the part most likely to wear out first, the guarantee period will be based on the life of the laser beam device. How many playing hours should the guarantee cover? (Round to the next playing hour.)

19. ***Guarantee: Package Delivery*** Express Courier Service has found that the delivery time for packages is normally distributed, with mean 14 hours and standard deviation 2 hours.
(a) For a package selected at random, what is the probability that it will be delivered in 18 hours or less?
(b) ***Inverse Normal Distribution*** What should be the guaranteed delivery time on all packages in order to be 95% sure that the package will be delivered before this time? *Hint:* Note that 5% of the packages will be delivered at a time beyond the guaranteed time period.

20. ***Control Chart: Landing Gear*** Hydraulic pressure in the main cylinder of the landing gear of a commercial jet is very important for a safe landing. If the pressure is not high enough, the landing gear may not lower properly. If it is too high, the connectors in the hydraulic line may spring a leak.

In-flight landing tests show that the actual pressure in the main cylinders is a variable with mean 819 pounds per square inch and standard deviation 23 pounds per square inch. Assume that these values for the mean and standard deviation are considered safe values by engineers.


Jeff Greenberg/PhotoEdit

(a) For nine consecutive test landings, the pressure in the main cylinder is recorded as follows:

Landing number	1	2	3	4	5	6	7	8	9
Pressure	870	855	830	815	847	836	825	810	792

Make a control chart for the pressure in the main cylinder of the hydraulic landing gear, and plot the data on the control chart. Looking at the control chart, would you say the pressure is "in control" or "out of control"? Explain your answer. Identify any out-of-control signals by type (I, II, or III).
(b) For 10 consecutive test landings, the pressure was recorded on another plane as follows:

Landing number	1	2	3	4	5	6	7	8	9	10
Pressure	865	850	841	820	815	789	801	765	730	725

Make a control chart and plot the data on the chart. Would you say the pressure is "in control" or not? Explain your answer. Identify any out-of-control signals by type (I, II, or III).

21. ***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?

22. ***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).

23. ***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?

24. ***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?

25. ***Basic Computation: $\hat{p}$ Distribution*** Suppose we have a binomial distribution with $n = 24$ trials and probability of success $p = 0.4$ on each trial. The sample proportion of successes is $\hat{p} = r/n$.
   (a) Is it appropriate to approximate the $\hat{p}$ distribution with a normal distribution? Explain.
   (b) What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

26. ***Green Behavior: Purchasing Habits*** A recent Harris Poll on green behavior showed that 25% of adults often purchase used items instead of new ones. Consider a random sample of 75 adults. Let $\hat{p}$ be the sample proportion of adults who often purchase used instead of new items.
   (a) Is it appropriate to approximate the $\hat{p}$ distribution with a normal distribution? Explain.
   (b) What are the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

DATA HIGHLIGHTS: GROUP PROJECTS

Mike Mazzaschi/Stock Boston

Wild iris

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:

> 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. [E. Anderson, "The Irises of the Gaspé Peninsula," *Bulletin, American Iris Society*, Vol. 59 pp. 2–5, 1935.]

The data in Table 6-11 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, pp. 179–188, 1936). To find these data, visit the Brase/Brase statistics site at **http://www.cengage.com/statistics/brase** 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/TI-*n*spire 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 6-39 shows a histogram for the given data generated on a TI-84Plus/TI-83Plus/TI-*n*spire 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 into 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 fall 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.

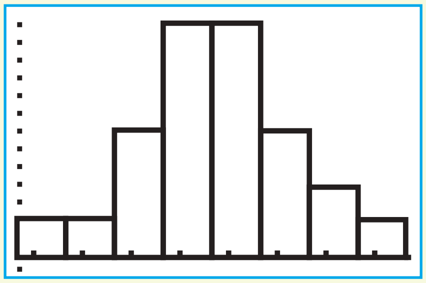
(e) Compare your answers to parts (c) and (d). Do you notice any differences? Why would these differences occur?

TABLE 6-11 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 6-39

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



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. If you look up the word *empirical* in a dictionary, you will find that it means "relying on experiment and observation rather than on theory." Discuss the empirical rule in this context. The empirical rule certainly applies to the normal distribution, but does it also apply to a wide variety of other distributions that are not *exactly* (theoretically) normal? Discuss the terms *mound-shaped* and *symmetrical*. Draw several sketches of distributions that are mound-shaped *and* symmetrical. Draw sketches of distributions that are not mound-shaped or symmetrical. To which distributions will the empirical rule apply?

2. Why are standard z values so important? Is it true that z values have no units of measurement? Why would this be desirable for comparing data sets with *different* units of measurement? How can we assess differences in quality or performance by simply comparing z values under a standard normal curve? Examine the formula for computing standard z values. Notice that it involves *both* the mean and the standard deviation. Recall that in Chapter 3 we commented that the mean of a data collection is not entirely adequate to describe the data; you need the standard deviation as well. Discuss this topic again in light of what you now know about normal distributions and standard z values.

3. Most companies that manufacture a product have a division responsible for quality control or quality assurance. The purpose of the quality-control division is to make reasonably certain that the products manufactured are up to company standards. Write a brief essay in which you describe how the statistics you have learned so far could be applied to an industrial application (such as control charts and the Antlers Lodge example).

4. 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 6.1 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 size of $n = 100$, 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.

5. 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 7, 8, and 9. 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 *not* to 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

Application 1

How can we determine if data originated from a normal distribution? We can look at a stem-and-leaf plot or histogram of the data to check for general symmetry, skewness, clusters of data, or outliers. However, a more sensitive way to check that a distribution is normal is to look at a special graph called a *normal quantile plot* (or a variation of this plot called a *normal probability plot* in some software packages). It really is not feasible to make a normal quantile plot by hand, but statistical software packages provide such plots. A simple version of the basic idea behind normal quantile plots involves the following process:

(a) Arrange the observed data values in order from smallest to largest, and determine the percentile occupied by each value. For instance, if there are 20 data values, the smallest datum is at the 5% point, the next smallest is at the 10% point, and so on.

(b) Find the z values that correspond to the percentile points. For instance, the z value that corresponds to the percentile 5% (i.e., percent in the left tail of the distribution) is $z = -1.645$.

(c) Plot each data value x against the corresponding percentile z score. If the data are close to a normal distribution, the plotted points will lie close to a straight line. (If the data are close to a standard normal distribution, the points will lie close to the line $x = z$.)

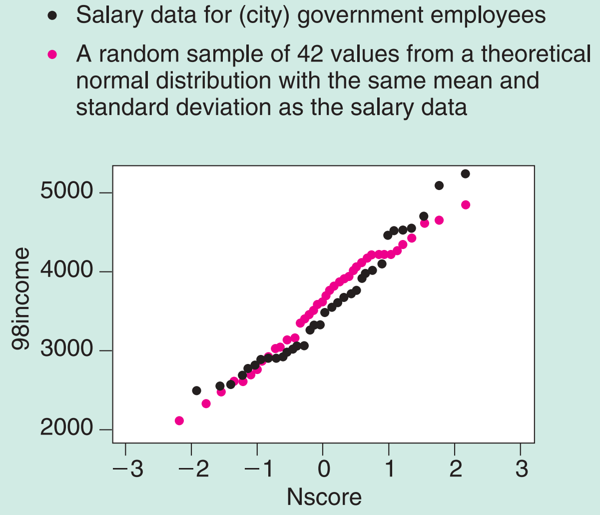
The actual process that statistical software packages use to produce the z scores for the data is more complicated.

Interpreting normal quantile plots

If the points of a normal quantile plot lie close to a straight line, the plot indicates that the data follow a normal distribution. Systematic deviations from a straight line or bulges in the plot indicate that the data distribution is not normal. Individual points off the line may be outliers.

Consider Figure 6-40. This figure shows Minitab-generated quantile plots for two data sets. The black dots show the normal quantile plot for the salary data of the first application. The red dots show the normal quantile plot for a random sample of 42 data values drawn from a theoretical normal distribution with the same mean and standard deviation as the salary data ($\mu \approx 3421$, $\sigma \approx 709$).

FIGURE 6-40 Normal Quantile Plots



(a) Do the black dots lie close to a straight line? Do the salaries appear to follow a normal distribution? Are there any outliers on the low or high side? Would you say that any of the salaries are "out of line" for a normal distribution?

(b) Do the red dots lie close to a straight line? We know the red dots represent a sample drawn from a normal distribution. Is the normal quantile plot for the red dots consistent with this fact? Are there any outliers shown?

Technology Hints

TI-84Plus/TI-83Plus/TI-*n*spire

Enter the data. Press **STATPLOT** and select one of the plots. Highlight **ON**. Then highlight the sixth plot option. To get a plot similar to that of Figure 6-40, choose Y as the data axis.

Minitab

Minitab has several types of normal quantile plots that use different types of scales. To create a normal quantile plot similar to that of Figure 6-40, enter the data in column C1. Then use the menu choices **Calc ➤ Calculator.** In the dialogue box listing the functions, scroll to **Normal Scores.** Use **NSCOR(C1)** and store the results in column C2. Finally, use the menu choices **Graph ➤ Plot.** In the dialogue box, use C1 for variable y and C2 for variable x.

SPSS

Enter the data. Use the menu choices **Analyze ➤ Descriptive Statistics ➤ Explore.** In the dialogue box, move your data variable to the dependent list. Click **Plots . . . .** Check "Normality plots with tests." The graph appears in the output window.

Application 2

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 6-12.

(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.

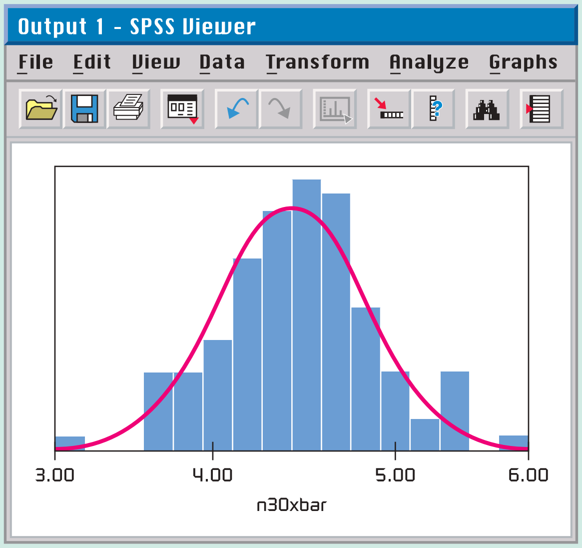
TABLE 6-12 **Frequency Table of Sample Means**

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

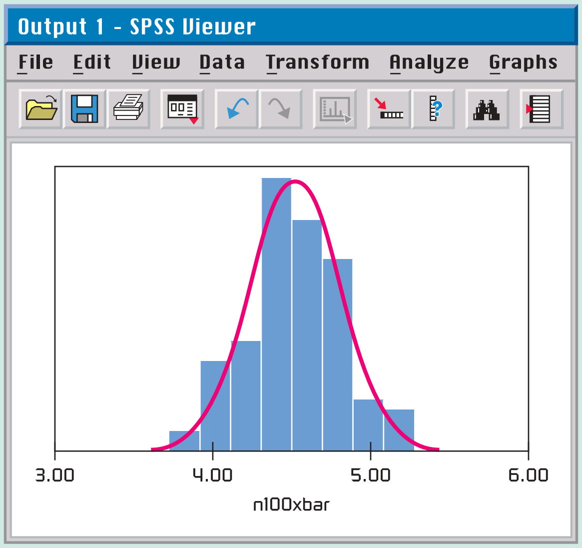
FIGURE 6-41

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

(a) $n = 30$



(b) $n = 100$



Technology Hints

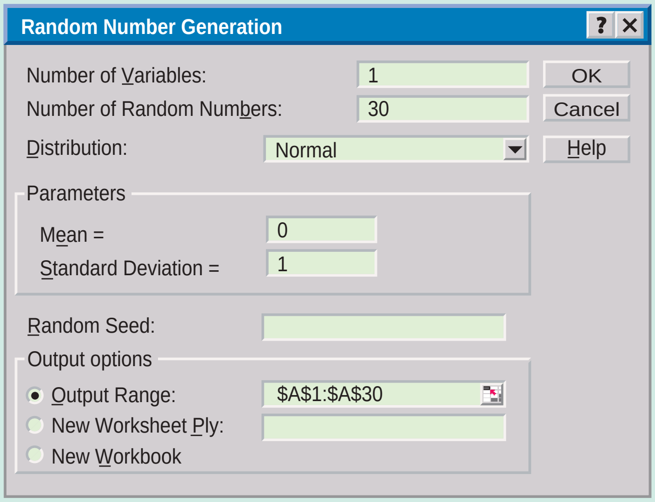
The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, Minitab, and SPSS all support the process of drawing random samples from a variety of distributions. Macros can be written in Excel 2007, Minitab, and the professional version of SPSS to repeat the six steps of the project. Figure 6-41 shows 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/TI-*n*spire

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 2007

On the Home screen, click on the **Data** tab. In the **Analysis** group, select **Data Analysis.** In the dialogue box, select **Random Number Generator.** The next 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.



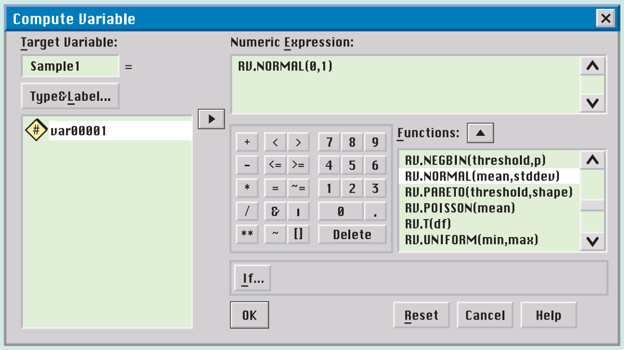
Minitab

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.

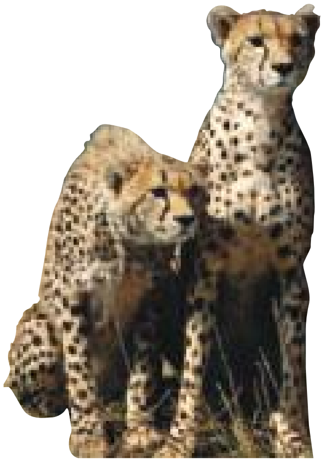
SPSS

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 sample1 as the target variable for the first sample, and so forth.

In the function group, select **Random Numbers** and in the functions and special variables group, select **Rv.Uniform** for samples from a uniform distribution. Functions **Rv.Normal** and **Rv.Binom** provide random samples from normal and binomial distributions, respectively. For each function, the necessary parameters are described.



Hubert Stadler/Corbis



Cumulative Review Problems

CHAPTERS 4–6

The Hill of Tara is located in south-central Meath, not far from Dublin, Ireland. Tara is of great cultural and archaeological importance, since it is by legend the seat of the ancient high kings of Ireland. For more information, see *Tara: An Archaeological Survey*, by Conor Newman, Royal Irish Academy, Dublin.

Hubert Stadler/Corbis

Magnetic surveying is one technique used by archaeologists to determine anomalies arising from variations in magnetic susceptibility. Unusual changes in magnetic susceptibility might (or might not) indicate an important archaeological discovery. Let x be a random variable that represents a magnetic susceptibility (MS) reading for a randomly chosen site on the Hill of Tara. A random sample of 120 sites gave the readings shown in Table A below.

TABLE A **Magnetic Susceptibility Readings, centimeter-gram-second $\times 10^{-6}$ (cmg $\times 10^{-6}$)**

Comment	Magnetic Susceptibility	Number of Readings	Estimated Probability
"cool"	$0 \le x < 10$	30	30/120 = 0.25
"neutral"	$10 \le x < 20$	54	54/120 = 0.45
"warm"	$20 \le x < 30$	18	18/20 = 0.15
"very interesting"	$30 \le x < 40$	12	12/120 = 0.10
"hot spot"	$40 \le x$	6	6/120 = 0.05

Answers may vary slightly due to rounding.

1. *Statistical Literacy: Sample Space* What is a statistical experiment? How could the magnetic susceptibility intervals $0 \le x < 10$, $10 \le x < 20$, and so on, be considered events in the sample space of all possible readings?

2. *Statistical Literacy: Probability* What is probability? What do we mean by relative frequency as a probability estimate for events? What is the law of large numbers? How would the law of large numbers apply in this context?

3. *Statistical Literacy: Probability Distribution* Do the probabilities shown in Table A add up to 1? Why should they total to 1?

4. *Probability Rules* For a site chosen at random, estimate the following probabilities.
   (a) $P(0 \le x < 30)$
   (b) $P(10 \le x < 40)$
   (c) $P(x < 20)$
   (d) $P(x \ge 20)$
   (e) $P(30 \le x)$
   (f) $P(x$ *not* less than 10)
   (g) $P(0 \le x < 10$ or $40 \le x)$
   (h) $P(40 \le x$ *and* $20 \le x)$

5. *Conditional Probability* Suppose you are working in a "warm" region in which all MS readings are 20 or higher. In this same region, what is the probability that you will find a "hot spot" in which the readings are 40 or higher? Use conditional probability to estimate $P(40 \le x \mid 20 \le x)$. *Hint*: See Problem 39 of Section 6.3.

6. *Discrete Probability Distribution* Consider the midpoint of each interval. Assign the value 45 as the midpoint for the interval $40 \le x$. The midpoints constitute the sample space for a discrete random variable. Using Table A, compute the expected value μ and the standard deviation σ.

Midpoint x	5	15	25	35	45
$P(x)$					

7. *Binomial Distribution* Suppose a reading between 30 and 40 is called "very interesting" from an archaeological point of view. Let us say you take readings at $n = 12$ sites chosen at random. Let r be a binomial random variable that represents the number of "very interesting" readings from these 12 sites.
   (a) Let us call "very interesting" a binomial success. Use Table A to find p, the probability of success on a single trial, where $p = P(\text{success}) = P(30 \leq x < 40)$.
   (b) What is the expected value μ and standard deviation σ for the random variable r?
   (c) What is the probability that you will find *at least* one "very interesting" reading in the 12 sites?
   (d) What is the probability that you will find *fewer than* three "very interesting" readings in the 12 sites?

8. *Geometric Distribution* Suppose a "hot spot" is a site with a reading of 40 or higher.
   (a) In a binomial setting, let us call success a "hot spot." Use Table A to find $p = P(\text{success}) = P(40 \leq x)$ for a single trial.
   (b) Suppose you decide to take readings at random until you get your *first* "hot spot." Let n be a random variable representing the trial on which you get your first "hot spot." Use the geometric probability distribution to write out a formula for $P(n)$.
   (c) What is the probability that you will need more than four readings to find the first "hot spot"? Compute $P(n > 4)$.

9. *Poisson Approximation to the Binomial* Suppose an archaeologist is looking for geomagnetic "hot spots" in an unexplored region of Tara. As in Problem 8, we have a binomial setting where success is a "hot spot." In this case, the probability of success is $p = P(40 \leq x)$. The archaeologist takes $n = 100$ magnetic susceptibility readings in the new, unexplored area. Let r be a binomial random variable representing the number of "hot spots" in the 100 readings.
   (a) We want to approximate the binomial random variable r by a Poisson distribution. Is this appropriate? What requirements must be satisfied before we can do this? Do you think these requirements are satisfied in this case? Explain. What is the value of λ?
   (b) What is the probability that the archaeologists will find six or fewer "hot spots?" *Hint*: Use Table 4 of Appendix II.
   (c) What is the probability that the archaeologists will find more than eight "hot spots"?

10. *Normal Approximation to the Binomial* Consider a binomial setting in which "neutral" is defined to be a success. So, $p = P(\text{success}) = P(10 \leq x < 20)$. Suppose $n = 65$ geomagnetic readings are taken. Let r be a binomial random variable that represents the number of "neutral" geomagnetic readings.
   (a) We want to approximate the binomial random variable r by a normal variable x. Is this appropriate? What requirements must be satisfied before we can do this? Do you think these requirements are satisfied in this case? Explain.
   (b) What is the probability that there will be at least 20 "neutral" readings out of these 65 trials?
   (c) Why would the Poisson approximation to the binomial *not* be appropriate in this case? Explain.

11. *Normal Distribution* *Oxygen demand* is a term biologists use to describe the oxygen needed by fish and other aquatic organisms for survival. The Environmental Protection Agency conducted a study of a wetland area in Marin County, California. In this wetland environment, the mean oxygen demand was $\mu = 9.9$ mg/L with 95% of the data ranging from 6.5 mg/L to 13.3 mg/L (Reference: EPA Report 832-R-93-005). Let x be a random variable that represents oxygen demand in this wetland environment. Assume x has a probability distribution that is approximately normal.
   (a) Use the 95% data range to estimate the standard deviation for oxygen demand. *Hint:* See Problem 31 of Section 6.3.
   (b) An oxygen demand below 8 indicates that some organisms in the wetland environment may be dying. What is the probability that the oxygen demand will fall below 8 mg/L?
   (c) A high oxygen demand can also indicate trouble. An oxygen demand above 12 may indicate an overabundance of organisms that endanger some types of plant life. What is the probability that the oxygen demand will exceed 12 mg/L?

12. *Statistical Literacy* Please give a careful but brief answer to each of the following questions.
   (a) What is a population? How do you get a simple random sample? Give examples.
   (b) What is a sample statistic? What is a sampling distribution? Give examples.
   (c) Give a careful and complete statement of the central limit theorem.
   (d) List at least three areas of everyday life to which the above concepts can be applied. Be specific.

13. ***Sampling Distribution*** Workers at a large toxic cleanup project are concerned that their white blood cell counts may have been reduced. Let x be a random variable that represents white blood cell count per cubic millimeter of whole blood in a healthy adult. Then $\mu = 7500$ and $\sigma \approx 1750$ (Reference: *Diagnostic Tests with Nursing Applications*, S. Loeb). A random sample of $n = 50$ workers from the toxic cleanup site were given a blood test that showed $\bar{x} = 6820$. What is the probability that, for healthy adults, $\bar{x}$ will be this low or lower?
   (a) How does the central limit theorem apply? Explain.
   (b) Compute $P(\bar{x} \le 6820)$.
   (c) ***Interpretation*** Based on your answer to part (b), would you recommend that additional facts be obtained, or would you recommend that the workers' concerns be dismissed? Explain.

14. ***Sampling Distribution*** Do you have a great deal of confidence in the advice given to you by your medical doctor? About 45% of all adult Americans claim they do have a great deal of confidence in thier M.D.s (Reference: *National Opinion Research Center*, University of Chicago). Suppose a random sample of $n = 32$ adults in a health insurance program are asked about their confidence in the medical advice their doctors give.
   (a) Is the normal approximation to the proportion $\hat{p} = r/n$ valid?
   (b) Find the values of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.

15. ***Summary*** Write a brief but complete essay in which you describe the probability distributions you have studied so far. Which apply to discrete random variables? Which apply to continuous random variables? Under what conditions can the binomial distribution be approximated by the normal? by the Poisson?