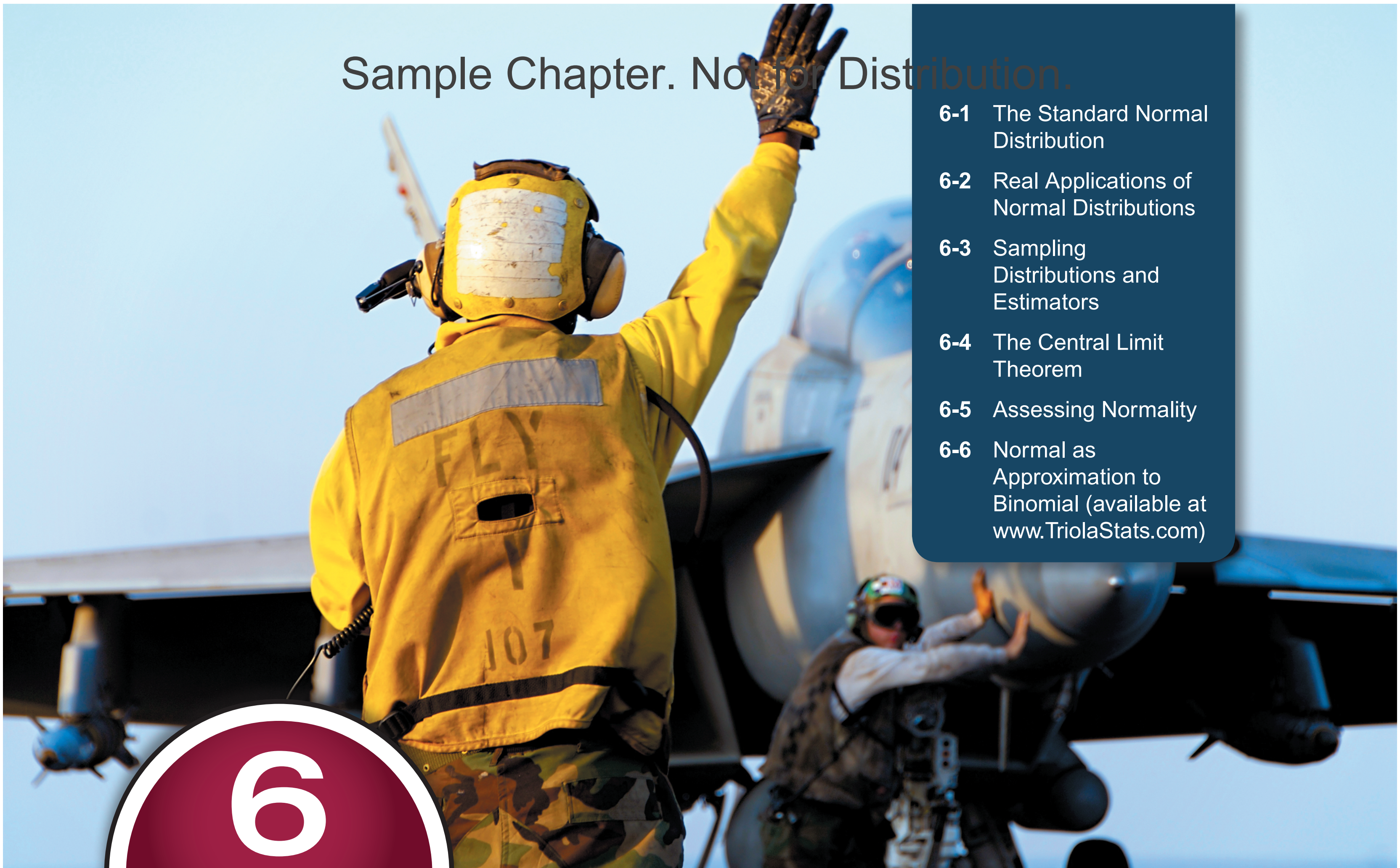


6

NORMAL PROBABILITY DISTRIBUTIONS

CHAPTER PROBLEM

So, You Want to Fly a U.S. Air Force Jet?

Do you satisfy the U.S. Air Force requirement of having a height between 64 inches and 77 inches? What percentage of adult males satisfy this height requirement? What percentage of adult females satisfy this height requirement?

Ergonomics 101 *Ergonomics* is a discipline focused on the design of tools and equipment so that they can be used safely, comfortably, and efficiently. The design of civilian and military aircraft makes extensive use of ergonomics. Some of the basic tools introduced in this chapter will enable us to solve many problems related to ergonomics, such as those involving the following requirements.

- The U.S. Air Force requires that its pilots must have heights between 64 inches and 77 inches. (This requirement was being changed at the time of this writing.)
- The U.S. Army requires that women must be between 58 in. and 80 in. tall.

- Section 4.4.2 of the Americans with Disabilities Act relates to height clearances with this statement: "Walks, halls, corridors, passageways, aisles, or other circulation spaces shall have 80 in. (2030 mm) minimum clear head room."
- The elevator in the San Francisco Airport rental car facility has a placard indicating a maximum load of 4000 lb or 27 passengers.
- A Disney requirement for someone wanting to be employed as the Tinkerbell character is that they must have a height between 58 inches and 62 inches.
- Radio City Music Hall Rockette dancers must be females with heights between 66 inches and 70.5 inches.
- The Bombardier Dash 8 aircraft can carry 37 passengers, and the fuel and baggage allow for a total passenger load of 6200 lb.
- When women were finally allowed to become pilots of fighter jets, engineers needed to redesign the ejection seats because they had been originally designed for men weighing between 140 lb and 211 lb.

Ergonomic problems often involve extremely important safety issues, and here are real cases that proved to be fatal:

- "We have an emergency for Air Midwest fifty-four eighty," said pilot Katie Leslie, just before her Beech plane crashed in Charlotte, North Carolina, resulting in the death of all 21 crew and passengers. Excessive total weight of the passengers was suspected as a factor that contributed to the crash.
- After 20 passengers perished when the *Ethan Allen* tour boat capsized on New York's Lake George, an investigation showed that although the number of passengers was below the maximum allowed, the boat should have been certified for a much smaller number of passengers.
- A water taxi sank in Baltimore's Inner Harbor, killing 5 of the 25 people on board. The boat was certified to carry 25 passengers, but their total weight exceeded the safe load of 3500 lb, so the number of passengers should have been limited to 20.

CHAPTER OBJECTIVES

Chapter 5 introduced *discrete* probability distributions, but in this chapter we introduce *continuous* probability distributions. Most of this chapter focuses on *normal distributions*, which are the most important distributions in the field of statistics. Here are the chapter objectives:

6-1 The Standard Normal Distribution

- Describe the characteristics of a standard normal distribution.
- Find the probability of some range of *z* scores in a standard normal distribution.
- Find *z* scores corresponding to regions under the curve representing a standard normal distribution.

6-2 Real Applications of Normal Distributions

- Develop the ability to describe a normal distribution (not necessarily a standard normal distribution).
- Find the probability of some range of values in a normal distribution.
- Find *x* scores corresponding to regions under the curve representing a normal distribution, and solve real problems using this skill.

6-3 Sampling Distributions and Estimators

- Develop the ability to describe a *sampling distribution of a statistic.*
- Determine whether a sample statistic serves as a good estimator of the corresponding population parameter.

6-4 The Central Limit Theorem

- Describe the central limit theorem.
- Apply the central limit theorem by finding the probability that a sample mean falls within some specified range of values.
- Identify conditions for which it is appropriate to use a normal distribution for the distribution of sample means.

6-5 Assessing Normality

- Develop the ability to examine histograms, outliers, and normal quantile plots to determine whether sample data appear to be from a population having a distribution that is approximately normal.

6-6 Normal as Approximation to Binomial (available at www.TriolaStats.com)

- Identify conditions for which it is appropriate to use a normal distribution as an approximation to a binomial probability distribution.
- Use the normal distribution for approximating probabilities for a binomial distribution.

6-1 The Standard Normal Distribution

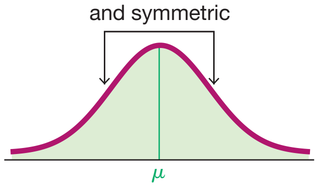


FIGURE 6-1 The Normal Distribution

Key Concept In this section we present the *standard normal distribution*, which is a specific normal distribution having the following three properties:

1. Bell-shaped: The graph of the standard normal distribution is bell-shaped (as in Figure 6-1).
2. $\mu = 0$: The standard normal distribution has a mean equal to 0.
3. $\sigma = 1$: The standard normal distribution has a standard deviation equal to 1.

In this section we develop the skill to find areas (or probabilities or relative frequencies) corresponding to various regions under the graph of the standard normal distribution. In addition, we find z scores that correspond to areas under the graph. These skills become important in the next section as we study nonstandard normal distributions and the real and important applications that they involve.

Normal Distributions

There are infinitely many different normal distributions, depending on the values used for the mean and standard deviation. We begin with a brief introduction to this general family of normal distributions.

DEFINITION

If a continuous random variable has a distribution with a graph that can be described by the equation given as Formula 6-1 (shown on the next page), we say that it has a **normal distribution.** A normal distribution is bell-shaped and symmetric, as shown in Figure 6-1.

FORMULA 6-1

$$y = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Fortunately, we won't actually use Formula 6-1, but examining the right side of the equation reveals that any particular normal distribution is determined by two parameters: the population mean, μ, and population standard deviation, σ. (In Formula 6-1, x is a variable, $\pi = 3.14159\ldots$ and $e = 2.71828\ldots$.) Once specific values are selected for μ and σ, Formula 6-1 is an equation relating x and y, and we can graph that equation to get a result that will look like Figure 6-1. And that's about all we need to know about Formula 6-1!

Uniform Distributions

The major focus of this chapter is the concept of a normal probability distribution, but we begin with a *uniform distribution* so that we can see the following two very important properties:

1. The area under the graph of a continuous probability distribution is equal to 1.
2. There is a correspondence between area and probability, so *probabilities* can be found by identifying the corresponding *areas* in the graph using this formula for the area of a rectangle:

$$\text{Area} = \text{height} \times \text{width}$$

DEFINITION

A continuous random variable has a **uniform distribution** if its values are *equally* spread over the range of possible values. The graph of a uniform distribution results in a rectangular shape.

Density Curve The graph of any continuous probability distribution is called a **density curve,** and any density curve must satisfy the requirement that the total area under the curve is exactly 1. This requirement that the area must equal 1 simplifies probability problems, so the following statement is really important:

Because the total area under any density curve is equal to 1, there is a correspondence between *area* and *probability*.

EXAMPLE 1 Waiting Times for Airport Security

During certain time periods at JFK airport in New York City, passengers arriving at the security checkpoint have waiting times that are uniformly distributed between 0 minutes and 5 minutes, as illustrated in Figure 6-2 on the next page.

Refer to Figure 6-2 to see these properties:

- All of the different possible waiting times are *equally likely*.
- Waiting times can be *any* value between 0 min and 5 min, so it is possible to have a waiting time of 1.234567 min.
- By assigning the probability of 0.2 to the height of the vertical line in Figure 6-2, the *enclosed area is exactly 1*. (In general, we should make the height of the vertical line in a uniform distribution equal to 1/range.)

continued

Power of Small Samples

The Environmental Protection Agency (EPA) had discovered that Chrysler automobiles had malfunctioning carburetors, with the result that carbon monoxide emissions were too high. Chryslers with 360- and 400-cubic-inch displacements and two-barrel carburetors were involved. The EPA ordered Chrysler to fix the problem, but Chrysler refused, and the case of *Chrysler Corporation vs. The Environmental Protection Agency* followed. That case led to the conclusion that there was "substantial evidence" that the Chryslers produced excessive levels of carbon monoxide. The EPA won the case, and Chrysler was forced to recall and repair 208,000 vehicles. In discussing this case in an article in *AMSTAT News,* Chief Statistician for the EPA Barry Nussbaum wrote this: "Sampling is expensive, and environmental sampling is usually quite expensive. At the EPA, we have to do the best we can with small samples or develop models. . . . What was the sample size required to affect such a recall (of the 208,000 Chryslers)? The answer is a mere 10. It is both an affirmation of the power of inferential statistics and a challenge to explain how such a (small) sample could possibly suffice."

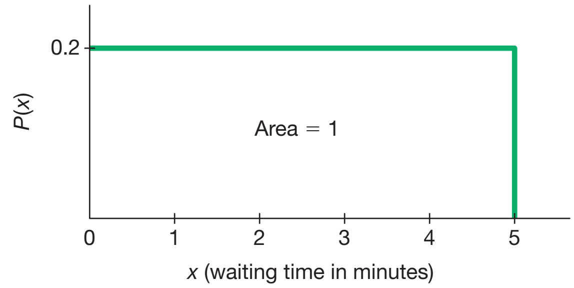


FIGURE 6-2 Uniform Distribution of Waiting Time

YOUR TURN. Do Exercise 5 "Continuous Uniform Distribution."

EXAMPLE 2 Waiting Times for Airport Security

Given the uniform distribution illustrated in Figure 6-2, find the probability that a randomly selected passenger has a waiting time of at least 2 minutes.

SOLUTION

The shaded area in Figure 6-3 represents waiting times of at least 2 minutes. Because the total area under the density curve is equal to 1, there is a correspondence between area and probability. We can easily find the desired *probability* by using *areas* as follows:

$$\begin{aligned} P(\text{wait time of at least 2 min}) &= \text{height} \times \text{width of shaded area in Figure 6-3} \\ &= 0.2 \times 3 \\ &= 0.6 \end{aligned}$$

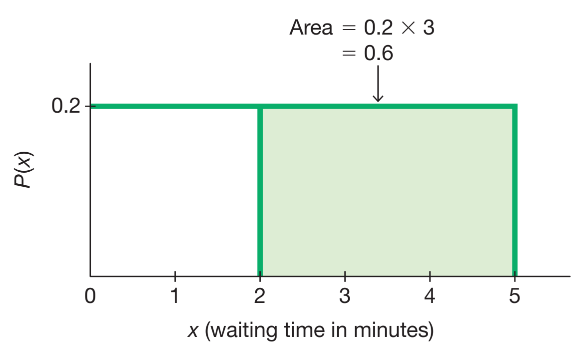


FIGURE 6-3 Using Area to Find Probability

INTERPRETATION

The probability of randomly selecting a passenger with a waiting time of at least 2 minutes is 0.6.

YOUR TURN. Do Exercise 7 "Continuous Uniform Distribution."

Standard Normal Distribution

The density curve of a uniform distribution is a horizontal straight line, so we can find the area of any rectangular region by applying this formula:

$$\text{Area} = \text{height} \times \text{width}.$$

Because the density curve of a normal distribution has a more complicated bell shape, as shown in Figure 6-1, it is more difficult to find areas. However, the basic principle is the same: *There is a correspondence between area and probability*. In Figure 6-4 we show that for a standard normal distribution, the area under the density curve is equal to 1. In Figure 6-4, we use "*z* Score" as a label for the horizontal axis, and this is common for the standard normal distribution, defined as follows.

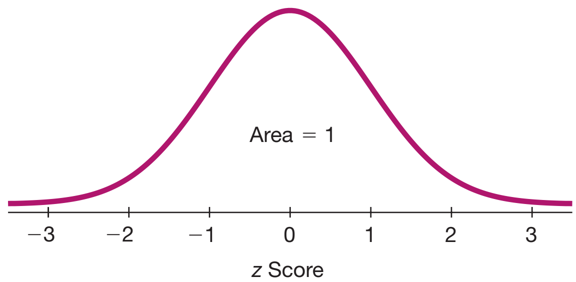


FIGURE 6-4 Standard Normal Distribution

DEFINITION

The **standard normal distribution** is a probability distribution with these properties:

- The distribution is a normal distribution, so it is bell-shaped as in Figure 6-4.
- The population parameter of the mean has the specific value of $\mu = 0$.
- The population parameter of the standard deviation has the specific value of $\sigma = 1$.
- The total area under its density curve is equal to 1 (as in Figure 6-4).

Finding Probabilities When Given *z* Scores

It is not easy to manually find areas in Figure 6-4, but we can find areas (or probabilities) for many different regions in Figure 6-4 by using technology, or we can also use Table A-2 (in Appendix A). Key features of the different methods are summarized in Table 6-1, which follows. (StatCrunch provides options for a cumulative left region, a cumulative right region, or the region between two boundaries.) Because calculators and software generally give more accurate results than Table A-2, we *strongly* recommend using technology. (When there are discrepancies, answers in Appendix D will generally include results based on technology as well as answers based on Table A-2.)

If using Table A-2, it is essential to understand these points:

1. Table A-2 is designed only for the *standard* normal distribution, which is a normal distribution with a mean of 0 and a standard deviation of 1.
2. Table A-2 is on two pages, with the left page for *negative z* scores and the right page for *positive z* scores.
3. Each value in the body of the table is a *cumulative area from the left* up to a vertical boundary above a specific *z* score.
4. When working with a graph, avoid confusion between *z* scores and areas.

   z* score:** ***Distance **along the horizontal scale of the standard normal distribution (corresponding to the number of standard deviations above or below the mean); refer to the leftmost column and top row of Table A-2.**

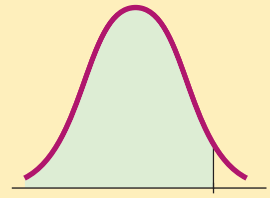
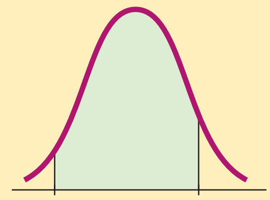
   Area: ***Region*** **under the curve; refer to the values in the** ***body*** **of Table A-2.**

5. The part of the *z* score denoting hundredths is found across the top row of Table A-2.

Go Figure

134: The number of times people check their smartphones every day, according to a Dignity Health survey of 2000 smartphone users.

TABLE 6-1 Formats Used for Finding Normal Distribution Areas

Cumulative Area from the Left The following provide the *cumulative area from the left* up to a vertical line above a specific value of *z*: • **Table A-2** • **Statdisk** • **Minitab** • **Excel** • **StatCrunch**	 **Cumulative Left Region**
Area Between Two Boundaries The following provide the area bounded on the left and bounded on the right by vertical lines above specific values. • **TI-83/84 Plus calculator** • **StatCrunch**	 **Area Between Two Boundaries**

CAUTION When working with a normal distribution, be careful to avoid confusion between *z* scores and areas.

ROUND-OFF RULE FOR *z* SCORES Round *z* scores to two decimal places, such as 2.31. (Table A-2 includes *z* scores rounded to two decimal places.)

The following examples illustrate procedures that can be used with real and important applications introduced in the following sections.

EXAMPLE 3 Bone Density Test

A bone mineral density test can be helpful in identifying the presence or susceptibility to osteoporosis, a disease that causes bones to become more fragile and more likely to break. The result of a bone density test is commonly measured as a z score. The population of z scores is normally distributed with a mean of 0 and a standard deviation of 1, so these test results meet the requirements of a standard normal distribution, and the graph of the bone density test scores is as shown in Figure 6-5.

A randomly selected adult undergoes a bone density test. Find the probability that this person has a bone density test score less than 1.27.

SOLUTION

Note that the following are the *same* (because of the aforementioned correspondence between probability and area):

- *Probability* that the bone density test score is less than 1.27
- Shaded *area* shown in Figure 6-5

So we need to find the area in Figure 6-5 below $z = 1.27$. If using technology, see the Tech Center instructions included at the end of this section. If using Table A-2, begin with the z score of 1.27 by locating 1.2 in the left column; next find the value in the adjoining row of probabilities that is directly below 0.07, as shown in the accompanying excerpt. Table A-2 shows that there is an area of 0.8980 corresponding to $z = 1.27$. We want the area *below* 1.27, and Table A-2 gives the cumulative area from the left, so the desired area is 0.8980. Because of the correspondence between area and probability, we know that the probability of a z score below 1.27 is 0.8980.

INTERPRETATION

The *probability* that a randomly selected person has a bone density test result below 1.27 is 0.8980, shown as the shaded region in Figure 6-5. Another way to interpret this result is to conclude that 89.80% of people have bone density levels below 1.27.

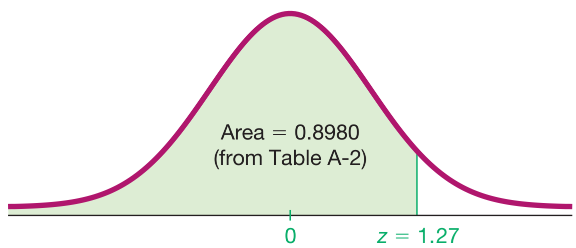


FIGURE 6-5 Finding Area to the Left of $z = 1.27$

TABLE A-2 *(continued)* Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

YOUR TURN. Do Exercise 9 "Standard Normal Distribution."

EXAMPLE 4 Bone Density Test: Finding the Area to the *Right* of a Value

Using the same bone density test from Example 3, find the probability that a randomly selected person has a result above −1.00. A value above −1.00 is considered to be in the "normal" range of bone density readings.

SOLUTION

We again find the desired *probability* by finding a corresponding *area*. We are looking for the area of the region to the right of $z = -1.00$ that is shaded in Figure 6-6 on the next page. The Statdisk display on the next page shows that the area to the right of $z = -1.00$ is 0.841345.

If we use Table A-2, we should know that it is designed to apply only to cumulative areas from the *left*. Referring to the page with *negative z* scores, we find that the cumulative area from the left up to $z = -1.00$ is 0.1587, as shown in Figure 6-6. Because the total area under the curve is 1, we can find the shaded area by

continued

subtracting 0.1587 from 1. The result is 0.8413. Even though Table A-2 is designed only for cumulative areas from the left, we can use it to find cumulative areas from the right, as shown in Figure 6-6.

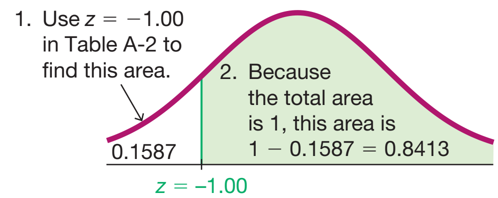
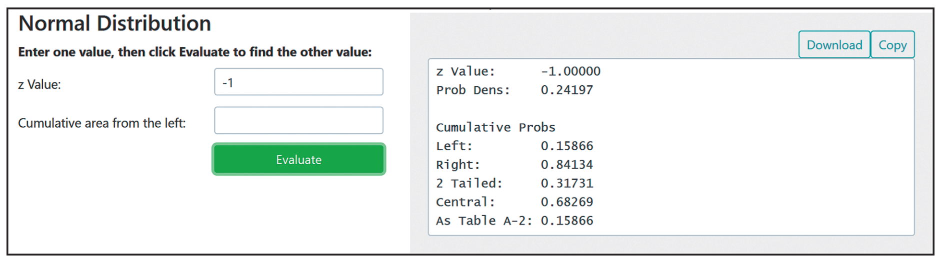


FIGURE 6-6 **Finding the Area to the Right of $z = -1$**

Statdisk



INTERPRETATION

Because of the correspondence between probability and area, we conclude that the *probability* of randomly selecting someone with a bone density reading above -1 is 0.8413 (which is the *area* to the right of $z = -1.00$). We could also say that 84.13% of people have bone density levels above -1.00.

YOUR TURN. Do Exercise 10 "Standard Normal Distribution."

Example 4 illustrates a way that Table A-2 can be used indirectly to find a cumulative area from the right. The following example illustrates another way that we can find an area indirectly by using Table A-2.

EXAMPLE 5 Bone Density Test: Finding the Area *Between* Two Values

A bone density test reading between -1.00 and -2.50 indicates that the subject has osteopenia, which is some bone loss. Find the probability that a randomly selected subject has a reading between -1.00 and -2.50.

SOLUTION

We are again dealing with normally distributed values having a mean of 0 and a standard deviation of 1. The values between -1.00 and -2.50 correspond to the shaded region in the third graph included in Figure 6-7. Table A-2 cannot be used to find that area directly, but we can use it to find the following:

- The area to the left of $z = -1.00$ is 0.1587.
- The area to the left of $z = -2.50$ is 0.0062.
- The area *between* $z = -2.50$ and $z = -1.00$ (the shaded area at the far right in Figure 6-7) is the difference between the areas found in the preceding two steps:

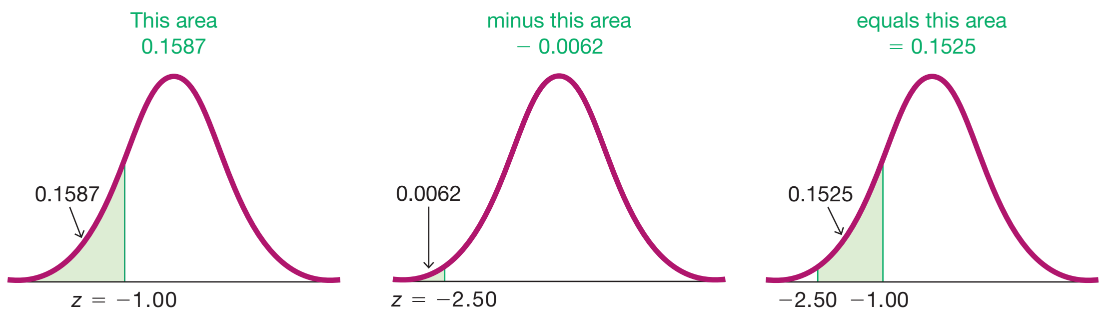


FIGURE 6-7 Finding the Area Between Two *z* Scores

INTERPRETATION

Using the correspondence between probability and area, we conclude that there is a probability of 0.1525 that a randomly selected subject has a bone density reading between −1.00 and −2.50. Another way to interpret this result is to state that 15.25% of people have osteopenia, with bone density readings between −1.00 and −2.50.

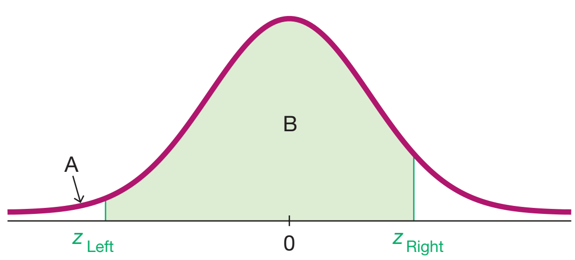
YOUR TURN. Do Exercise 11 "Standard Normal Distribution."

Example 5 can be generalized as the following rule:

The area corresponding to the region *between* two *z* scores can be found by finding the difference between the two areas found in Table A-2.

Figure 6-8 illustrates this general rule. The shaded region *B* can be found by calculating the *difference* between two areas found from Table A-2.

HINT Don't try to memorize a rule or formula for this case. Focus on *understanding* by using a graph. Draw a graph, shade the desired area, and then get creative to think of a way to find the desired area by working with cumulative areas from the left.



Shaded area B = (areas A and B combined) — (area A)

FIGURE 6-8 Finding the Area Between Two *z* Scores

Probabilities such as those in the preceding examples can also be expressed with the following notation.

Notation

$P(a < z < b)$	denotes the probability that the z score is between a and b.
$P(z > a)$	denotes the probability that the z score is greater than a.
$P(z < a)$	denotes the probability that the z score is less than a.

With this notation, $P(-2.50 < z < -1.00) = 0.1525$, states in symbols that the probability of a z score falling between −2.50 and −1.00 is 0.1525 (as in Example 5).

Finding z Scores from Known Areas

Examples 3, 4, and 5 all involved the standard normal distribution, and they were all examples with this same format: Given z scores, find areas (or probabilities). In many cases, we need a method for reversing the format: Given a known area (or probability), find the corresponding z score. In such cases, it is really important to avoid confusion between z scores and areas. Remember, z scores are *distances* along the horizontal scale, but areas (or probabilities) are regions under the density curve. (Table A-2 lists z-scores in the left column and across the top row, but areas are found in the *body* of the table.) We should also remember that z scores positioned in the left half of the curve are always negative. If we already know a probability and want to find the corresponding z score, we use the following procedure.

Procedure for Finding a z Score from a Known Area

1. Draw a bell-shaped curve and identify the region under the curve that corresponds to the given probability. If that region is not a cumulative region from the left, work instead with a known region that is a cumulative region from the left.
2. Use technology or Table A-2 to find the z score. With Table A-2, use the cumulative area from the left, locate the closest probability in the *body* of the table, and identify the corresponding z score.

Special Cases in Table A-2

z Score	Cumulative Area from the Left
1.645	0.9500
−1.645	0.0500
2.575	0.9950
−2.575	0.0050
Above 3.49	0.9999
Below −3.49	0.0001

Special Cases In the solution to Example 6 that follows, Table A-2 leads to a z score of 1.645, which is midway between 1.64 and 1.65. When using Table A-2, we can usually avoid interpolation by simply selecting the closest value. The accompanying table lists special cases that are often used in a wide variety of applications. (For one of those special cases, the value of $z = 2.576$ gives an area slightly closer to the area of 0.9950, but $z = 2.575$ has the advantage of being the value exactly midway between $z = 2.57$ and $z = 2.58$.) Except in these special cases, we can usually select the closest value in the table. (If a desired value is midway between two table values, select the larger value.) For z scores above 3.49, we can use 0.9999 as an approximation of the cumulative area from the left; for z scores below -3.49, we can use 0.0001 as an approximation of the cumulative area from the left.

EXAMPLE 6 Bone Density Test: Finding a Test Score

Use the same bone density test scores used in earlier examples. Those scores are normally distributed with a mean of 0 and a standard deviation of 1, so they meet the requirements of a standard normal distribution. Find the bone density score corresponding to P_{95}, the 95th percentile. That is, find the bone density score that separates the bottom 95% from the top 5%. See Figure 6-9.

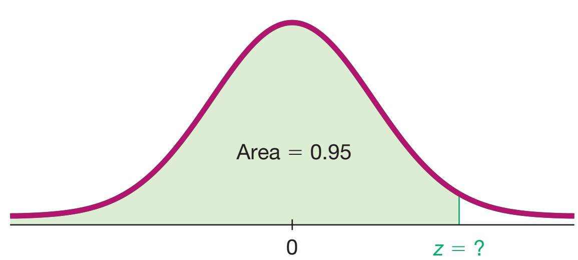


FIGURE 6-9 Finding the 95th Percentile

SOLUTION

Figure 6-9 shows the z score that is the 95th percentile, with 95% of the area (or 0.95) below it.

Technology: We could find the z score using technology. The accompanying Excel display shows that the z score with an area of 0.95 to its left is $z = 1.644853627$, or 1.645 when rounded.

Excel

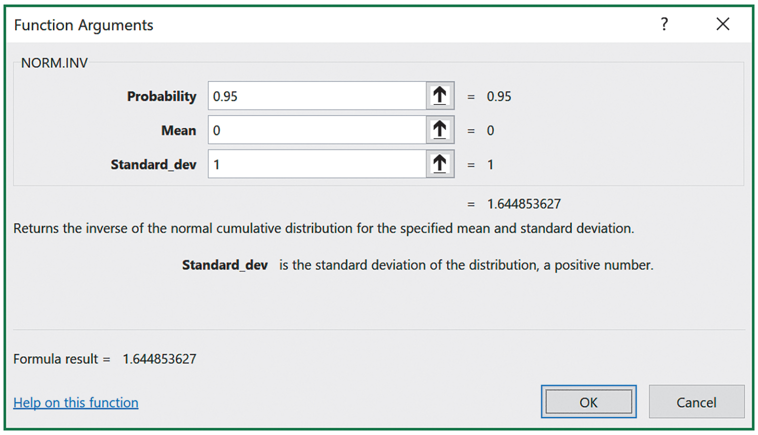


Table A-2: If using Table A-2, search for the area of 0.95 *in the body* of the table and then find the corresponding z score. In Table A-2 we find the areas of 0.9495 and 0.9505, but there's an asterisk with a special note indicating that 0.9500 corresponds to a z score of 1.645. We can now conclude that the z score in Figure 6-9 is 1.645, so the 95th percentile is $z = 1.645$.

INTERPRETATION

For bone density test scores, 95% of the scores are less than or equal to 1.645, and 5% of them are greater than or equal to 1.645.

YOUR TURN. Do Exercise 37 "Finding Bone Density Scores."

EXAMPLE 7 Bone Density Test

Using the same bone density test described in Example 3, we have a standard normal distribution with a mean of 0 and a standard deviation of 1. Find the bone density test score that separates the bottom 2.5% and find the score that separates the top 2.5%.

SOLUTION

The required z scores are shown in Figure 6-10. Those z scores can be found using technology. If using Table A-2 to find the z score located to the left, we search the *body of the table* for an area of 0.025. The result is $z = -1.96$. To find the z score located to the right, we search *the body of* Table A-2 for an area of 0.975. (Remember that Table A-2 always gives cumulative areas from the *left*.) The result is $z = 1.96$. The values of $z = -1.96$ and $z = 1.96$ separate the bottom 2.5% and the top 2.5%, as shown in Figure 6-10.

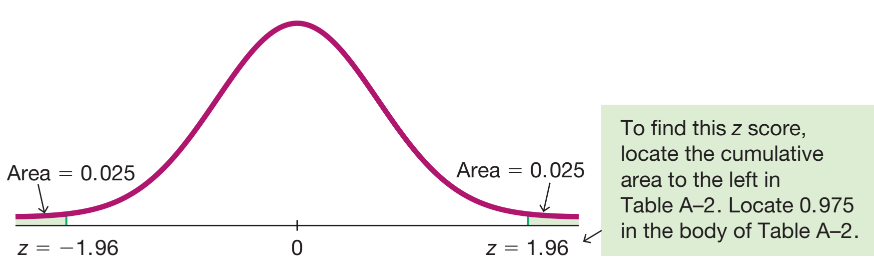


FIGURE 6-10 Finding z Scores

continued

Lefties Die Sooner?

A study by psychologists Diane Halpern and Stanley Coren received considerable media attention and generated considerable interest when it concluded that left-handed people don't live as long as right-handed people. Based on their study, it appeared that left-handed people live an average of nine years less than righties. The Halpern/Coren study has been criticized for using flawed data. They used second-hand data by surveying relatives about people who had recently died. The myth of lefties dying younger became folklore that has survived many years. However, more recent studies show that left-handed people do *not* have shorter lives than those who are right-handed.

INTERPRETATION

For the population of bone density test scores, 2.5% of the scores are equal to or less than −1.96 and 2.5% of the scores are equal to or greater than 1.96. Another interpretation is that 95% of all bone density test scores are between −1.96 and 1.96.

YOUR TURN. Do Exercise 39 "Finding Bone Density Scores."

Critical Values

DEFINITION

For the standard normal distribution, a **critical value** is a z score on the borderline separating those z scores that are *significantly low* or *significantly high*.

Common critical values are $z = -1.96$ and $z = 1.96$, and they are obtained as shown in Example 7. In Example 7, values of $z = -1.96$ or lower are significantly low because only 2.5% of the population have scores at or below −1.96, and the values at or above $z = 1.96$ are significantly high because only 2.5% of the population have scores at or above 1.96. Only 5% of all bone density scores are either −1.96 or lower or 1.96 or higher. Critical values will become extremely important in subsequent chapters. The following notation is used for critical z values found by using the standard normal distribution.

Notation

The expression z_α denotes the z score with an area of α to its right. (α is the Greek letter alpha.)

EXAMPLE 8 Finding the Critical Value z_α

Find the value of $z_{0.025}$. (Let $\alpha = 0.025$ in the expression z_α.)

SOLUTION

The notation of $z_{0.025}$ is used to represent the z score with an area of 0.025 to its *right*. Refer to Figure 6-10 and note that the value of $z = 1.96$ has an area of 0.025 to its right, so $z_{0.025} = 1.96$. Note that $z_{0.025}$ corresponds to a cumulative left area of 0.975.

YOUR TURN. Do Exercise 41 "Critical Values."

CAUTION When finding a value of z_α for a particular value of α, note that α is the area to the *right* of z_α, but Table A-2 and some technologies give cumulative areas to the *left* of a given z score. To find the value of z_α, resolve that conflict by using the value of $1 - \alpha$. For example, to find $z_{0.1}$, refer to the z score with an area of 0.9 to its left.

Examples 3 through 7 in this section are based on the real application of the bone density test, with scores that are normally distributed with a mean of 0 and standard deviation of 1, so that these scores have a standard normal distribution. Apart from the bone density test scores, it is rare to find such convenient parameters, because typical normal distributions have means different from 0 and standard deviations different from 1. In the next section we present methods for working with such normal distributions.

New Technology, New Data, New Insight

Residents of New York City believed that taxi cabs became scarce around rush hour in the late afternoon. Their complaints could not be addressed, because there were no data to support that alleged shortage. However, GPS units were installed on cabs and officials could then track their locations. After analyzing the GPS data, it was found that 20% fewer cabs were in service between 4:00 PM and 5:00 PM than in the preceding hour. Subjective beliefs and anecdotal stories were now substantiated with objective data.

Two factors were found to be responsible for the late afternoon cab shortage. First, the 12-hour shifts were scheduled to change at 5:00 PM so that drivers on both shifts would get an equal share at a rush hour. Second, rising rents in Manhattan forced many cab companies to house their cabs in Queens, so drivers had to start returning around 4:00 PM so that they could make it back in time and avoid fines for being late. In recent years, the shortage of cabs has been alleviated with the growth of companies such as Uber and Lyft.

TECH CENTER

Finding *z* Scores/Areas (Standard Normal)

Access tech supplements, videos, and data sets at **www.TriolaStats.com**

Statdisk

1. Click **Analysis** in the top menu.
2. Select **Probability Distributions** from the dropdown menu and select **Normal Distribution** from the submenu.
3. Enter the desired *z* score or cumulative area from the left of the *z* score and click **Evaluate.**

Minitab

1. Click **Calc** in the top menu.
2. Select **Probability Distributions** from the dropdown menu and select **Normal** from the submenu.

Finding Cumulative Area to the Left of a *z* Score

- Select **Cumulative probability,** enter mean of **0** and standard deviation of **1**.
- Select **Input Constant,** enter the desired *z* score, and click **OK.**

Finding *z* Score from a Known Probability

- Select **Inverse cumulative probability,** enter mean of **0** and standard deviation of **1.**
- Select **Input Constant,** enter the total area to the left of the *z* score, and click **OK.**

StatCrunch

1. Click **Stat** in the top menu.
2. Select **Calculators** from the dropdown menu and **Normal** from the submenu.
3. In the calculator box enter mean of **0** and standard deviation of **1.**
4. Enter the desired *z* score (middle box) or known probability/area (rightmost box). Select the desired inequality.
5. Click **Compute**.

TI-83/84 Plus Calculator

Unlike most other technologies, the TI-83/84 Plus bases areas on the region between two z scores, rather than cumulative regions from the left.

Finding Area Between Two *z* Scores

1. Press **2ND** then **VARS** keys to access the *DISTR* (distributions) menu.
2. Select **normalcdf** and press **ENTER**.
3. Enter the desired lower *z* score and upper *z* score. Enter **0** for μ and **1** for σ to complete the command **normalcdf(lower z,upper z,μ,σ).** Press **ENTER**.

TIP: If there is no lower *z* score, enter −99999999; if there is no upper *z* score, enter 99999999.

Finding *z* Score from a Known Probability

1. Press **2ND** then **VARS** keys to access the *DISTR* (distributions) menu.
2. Select **invNorm** and press **ENTER**.
3. Enter the area to the left or right of the *z* score, **0** for μ, and **1** for σ. For *Tail* select the tail where the area is located (*Left* or *Right*). The completed command is **InvNorm(area,0,1,TAIL)**. Press **ENTER**.

TIP: The TI-83 Plus only calculates area to the left of the *z* score. The completed command is **InvNorm(area,0,1)**.

Excel

Finding Cumulative Area to the Left of a *z* Score

1. Click **Insert Function f_x,** select the category **Statistical,** select the function **NORM.DIST,** and click **OK.**
2. For *x* enter the *z* score, enter **0** for *Mean,* enter **1** for *Standard_dev,* and enter **1** for *Cumulative*.
3. Click **OK.**

Finding *z* Score from a Known Probability

1. Click **Insert Function f_x,** select the category **Statistical,** and select the function **NORM.INV.**
2. Enter the probability, enter **0** for *Mean,* and enter **1** for *Standard_dev*.
3. Click **OK.**

R

R command not available at time of publication.

R is rapidly evolving, and an updated list of statistical commands is available at TriolaStats.com

6-1 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Normal Distribution What's wrong with the following statement? "Because the digits 0, 1, 2, . . . , 9 are the normal results from lottery drawings, such randomly selected numbers have a normal distribution."

2. Normal Distribution A normal distribution is informally described as a probability distribution that is "bell-shaped" when graphed. Draw a rough sketch of a curve having the bell shape that is characteristic of a normal distribution.

3. Standard Normal Distribution Identify the values of the mean and standard deviation for a normal distribution that is a *standard* normal distribution.

4. Notation What does the notation z_α indicate?

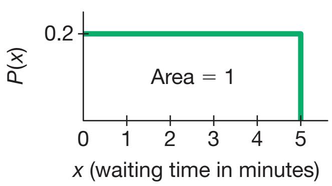


FIGURE 6-2

Continuous Uniform Distribution. ***In Exercises 5–8, refer to the continuous uniform distribution depicted in Figure 6-2 and described in Example 1. Assume that a passenger is randomly selected, and find the probability that the waiting time is within the given range.***

5. Greater than 3.00 minutes

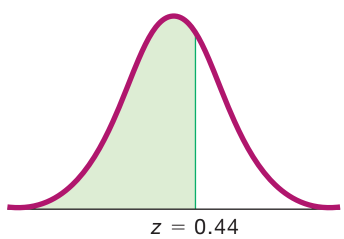
6. Less than 4.00 minutes

7. Between 2 minutes and 3 minutes

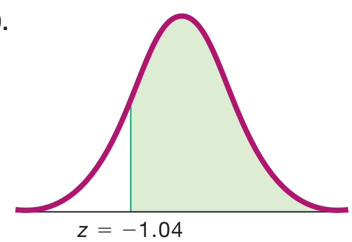
8. Between 2.5 minutes and 4.5 minutes

Standard Normal Distribution. ***In Exercises 9–12, find the area of the shaded region. The graph depicts the standard normal distribution of bone density scores with mean 0 and standard deviation 1.***

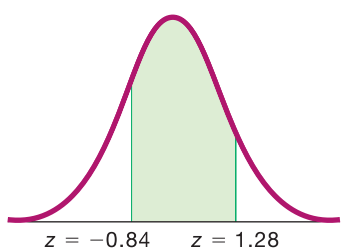
9.



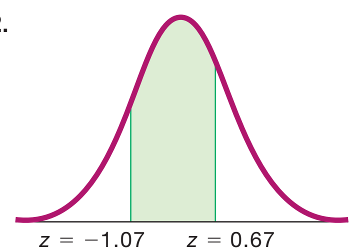
10.



11.

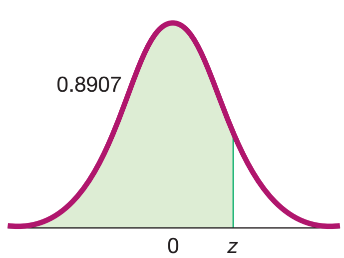


12.

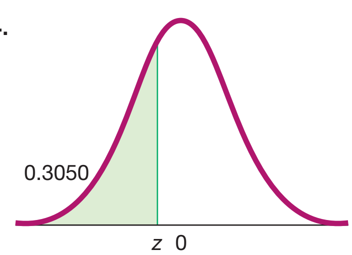


Standard Normal Distribution. ***In Exercises 13–16, find the indicated z score. The graph depicts the standard normal distribution of bone density scores with mean 0 and standard deviation 1.***

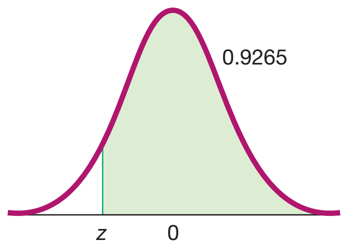
13.



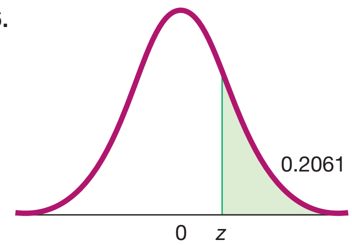
14.



15.



16.



Standard Normal Distribution. ***In Exercises 17–36, assume that a randomly selected subject is given a bone density test. Those test scores are normally distributed with a mean of 0 and a standard deviation of 1. In each case, draw a graph, then find the probability of the given bone density test scores. If using technology instead of Table A-2, round answers to four decimal places.***

17. Less than −2.00

18. Less than −0.50

19. Less than 1.33

20. Less than 2.33

21. Greater than 1.00

22. Greater than 2.33

23. Greater than −1.75

24. Greater than −2.09

25. Between 1.50 and 2.00

26. Between 1.37 and 2.25

27. Between −1.22 and −2.36

28. Between −0.45 and −2.08

29. Between −1.55 and 1.55

30. Between −0.77 and 1.42

31. Between −2.00 and 3.50

32. Between −3.52 and 2.53

33. Greater than −3.77

34. Less than −3.93

35. Between −4.00 and 4.00

36. Between −3.67 and 4.25

Finding Bone Density Scores. ***In Exercises 37–40 assume that a randomly selected subject is given a bone density test. Bone density test scores are normally distributed with a mean of 0 and a standard deviation of 1. In each case, draw a graph, then find the bone density test score corresponding to the given information. Round results to two decimal places.***

37. Find P_{99}, the 99th percentile. This is the bone density score separating the bottom 99% from the top 1%.

38. Find P_{15}, the 15th percentile. This is the bone density score separating the bottom 15% from the top 85%.

39. If bone density scores in the bottom 1% and the top 1% are used as cutoff points for levels that are too low or too high, find the two readings that are cutoff values.

40. Find the bone density scores that are the quartiles Q_1, Q_2, and Q_3.

Critical Values. ***In Exercises 41–44, find the indicated critical value. Round results to two decimal places.***

41. $z_{0.25}$

42. $z_{0.90}$

43. $z_{0.02}$

44. $z_{0.05}$

Basis for the Range Rule of Thumb and the Empirical Rule. ***In Exercises 45–48, find the indicated area under the curve of the standard normal distribution; then convert it to a percentage and fill in the blank. The results form the basis for the range rule of thumb and the empirical rule introduced in Section 3-2.***

45. About _____% of the area is between $z = -1$ and $z = 1$ (or within 1 standard deviation of the mean).

46. About _____% of the area is between $z = -2$ and $z = 2$ (or within 2 standard deviations of the mean).

47. About _____% of the area is between $z = -3$ and $z = 3$ (or within 3 standard deviations of the mean).

48. About _____% of the area is between $z = -3.5$ and $z = 3.5$ (or within 3.5 standard deviations of the mean).

6-1 Beyond the Basics

49. Significance For bone density scores that are normally distributed with a mean of 0 and a standard deviation of 1, find the *percentage* of scores that are

a. *significantly high* (or at least 2 standard deviations above the mean).

b. *significantly low* (or at least 2 standard deviations below the mean).

c. *not significant* (or less than 2 standard deviations away from the mean).

50. Distributions In a continuous uniform distribution,

$$\mu = \frac{\text{minimum} + \text{maximum}}{2} \quad \text{and} \quad \sigma = \frac{\text{range}}{\sqrt{12}}$$

a. Find the mean and standard deviation for the distribution of the waiting times represented in Figure 6-2, which accompanies Exercises 5–8.

b. For a continuous uniform distribution with $\mu = 0$ and $\sigma = 1$, the minimum is $-\sqrt{3}$ and the maximum is $\sqrt{3}$. For this continuous uniform distribution, find the probability of randomly selecting a value between –1 and 1, and compare it to the value that would be obtained by incorrectly treating the distribution as a standard normal distribution. Does the distribution affect the results very much?

6-2 Real Applications of Normal Distributions

Key Concept Most of the preceding section dealt with the real-world application of bone density scores, which have a normal distribution with $\mu = 0$ and $\sigma = 1$. However, it is rare to find other real applications of a standard normal distribution. We now extend the methods of the previous section so that we can work with any *nonstandard normal distribution* (with a mean different from 0 and/or a standard deviation different from 1). The key is a simple conversion (Formula 6-2) that allows us to "standardize" any normal distribution so that x values can be transformed to z scores; then the methods of the preceding section can be used.

FORMULA 6-2

$$z = \frac{x - \mu}{\sigma} \quad \text{(round } z \text{ scores to 2 decimal places)}$$

Figure 6-11 illustrates the conversion from a nonstandard to a standard normal distribution. The area in *any* normal distribution bounded by some score x (as in Figure 6-11a) is the *same* as the area bounded by the corresponding z score in the standard normal distribution (as in Figure 6-11b).

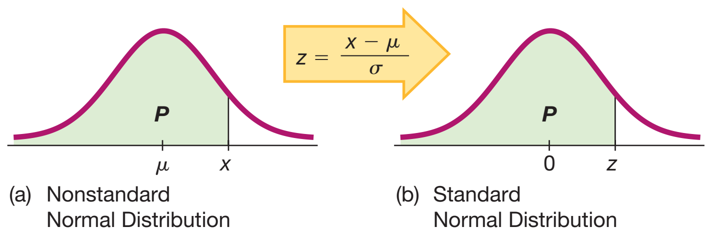


FIGURE 6-11 **Converting Distributions**

Most statistics calculators and software do not require the use of Formula 6-2 to convert to z scores because probabilities can be found directly. However, if using Table A-2, we must first convert values to standard z scores.

When finding areas with a nonstandard normal distribution, use the following procedure.

Procedure for Finding Areas with a Nonstandard Normal Distribution

1. Sketch a normal curve, label the mean and any specific x values, and then *shade* the region representing the desired probability.
2. For each relevant value x that is a boundary for the shaded region, use Formula 6-2 to convert that value to the equivalent z score. (With many technologies, this step can be skipped.)
3. Use technology (software or a calculator) or Table A-2 to find the area of the shaded region. This area is the desired probability.

The following example illustrates the above procedure.

EXAMPLE 1 What Proportion of Men Are Taller Than the 72 in. Height Requirement for Showerheads (According to Most Building Codes)?

Heights of men are normally distributed with a mean of 68.6 in. and a standard deviation of 2.8 in. (based on Data Set 1 "Body Data" in Appendix B). Find the percentage of men who are taller than a showerhead positioned at 72 in. above the floor.

SOLUTION

Step 1: See Figure 6-12, which incorporates this information: Men have heights that are normally distributed with a mean of 68.6 in. and a standard deviation of 2.8 in. The shaded region represents the men who are taller than the showerhead height of 72 in.

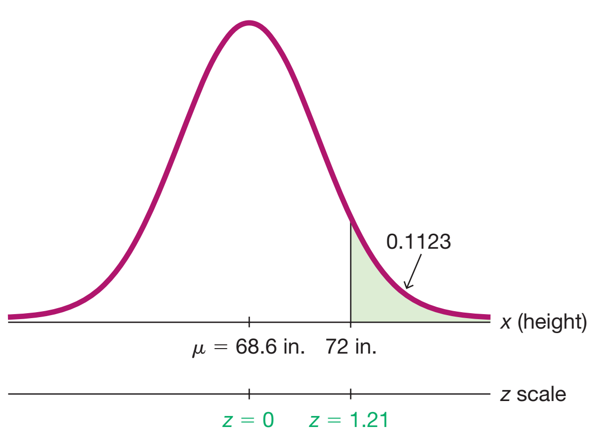


FIGURE 6-12 **Heights of Men**

continued

Go Figure

293,000,000,000: Number of e-mails sent each day.

Step 2: We can convert the showerhead height of 72 in. to the z score of 1.21 by using Formula 6-2 as follows:

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 68.6}{2.8} = 1.21 \text{ (rounded to two decimal places)}$$

Step 3: Technology: Technology can be used to find that the area to the right of 72 in. in Figure 6-12 is 0.1123 rounded. (With many technologies, Step 2 can be skipped. See technology instructions at the end of this section.) The result of 0.1123 from technology is more accurate than the result of 0.1131 found by using Table A-2.

Table A-2: Use Table A-2 to find that the cumulative area to the *left* of $z = 1.21$ is 0.8869. (Remember, Table A-2 is designed so that all areas are cumulative areas from the *left*.) Because the total area under the curve is 1, it follows that the shaded area in Figure 6-12 is $1 - 0.8869 = 0.1131$.

INTERPRETATION

The proportion of men taller than the showerhead height of 72 in. is 0.1123, or 11.23%. About 11% of men may find the design to be unsuitable. (*Note:* Some NBA teams have been known to intentionally use lower showerheads in the locker rooms of visiting basketball teams.)

YOUR TURN. Do Exercise 13 "Pulse Rates."

CP EXAMPLE 2 Air Force Height Requirement

Until recently, the U.S. Air Force required that pilots have heights between 64 in. and 77 in. Heights of women are normally distributed with a mean of 63.7 in. and a standard deviation of 2.9 in. (based on Data Set 1 "Body Data" in Appendix B). What percentage of women meet that height requirement?

SOLUTION

Figure 6-13 shows the shaded region representing heights of women between 64 in. and 77 in.

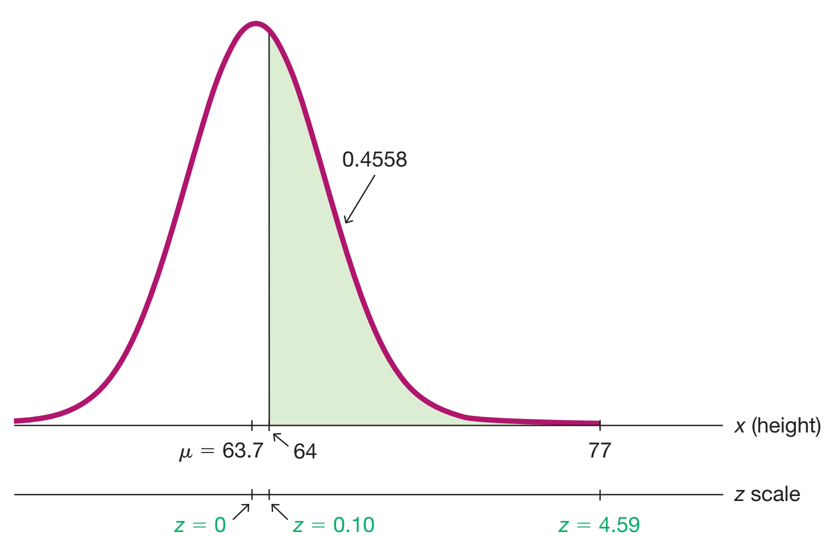


FIGURE 6-13 Heights of Women

Step 1: See Figure 6-13, which incorporates this information: Women have heights that are normally distributed with a mean of 63.7 in. and a standard deviation of 2.9 in. The shaded region represents the women with heights between 64 in. and 77 in.

Step 2: With some technologies, the shaded area in Figure 6-13 can be found directly and it is not necessary to convert the x scores of 64 in. and 77 in. to z scores. (See Step 3.)

If using Table A-2, we cannot find the shaded area directly, but we can find it indirectly by using the same procedures from Section 6-1, as follows: (1) Find the cumulative area from the left up to 77 in. (or $z = 4.59$); (2) find the cumulative area from the left up to 64 in. (or $z = 0.10$); (3) find the difference between those two areas. The heights of 77 in. and 64 in. are converted to z scores by using Formula 6-2 as follows:

$$\text{For } x = 77 \text{ in.: } z = \frac{x - \mu}{\sigma} = \frac{77 - 63.7}{2.9} = 4.59$$

$$(z = 4.59 \text{ yields an area of } 0.9999.)$$

$$\text{For } x = 64 \text{ in.: } z = \frac{x - \mu}{\sigma} = \frac{64 - 63.7}{2.9} = 0.10$$

$$(z = 0.10 \text{ yields an area of } 0.5398.)$$

Step 3: Technology: To use technology, refer to the Tech Center instructions on page 267. Technology will show that the shaded area in Figure 6-13 is 0.4588

Table A-2: Refer to Table A-2 with $z = 4.59$ and find that the cumulative area to the *left* of $z = 4.59$ is 0.9999. (Remember, Table A-2 is designed so that all areas are cumulative areas from the *left*.) Table A-2 also shows that $z = 0.10$ corresponds to an area of 0.5398. Because the areas of 0.9999 and 0.5398 are *cumulative areas from the left*, we find the shaded area in Figure 6-13 as follows:

$$\text{Shaded area in Figure 6–13} = 0.9999 - 0.5398 = 0.4601$$

There is a relatively small discrepancy between the area of 0.4588 found from technology and the area of 0.4601 found from Table A-2. The area obtained from technology is more accurate because it is based on unrounded z scores, whereas Table A-2 requires z scores rounded to two decimal places.

INTERPRETATION

Expressing the result as a percentage, we conclude that about 46% of women satisfy the requirement of having a height between 64 in. and 77 in. About 54% of women did not meet that requirement and they were not eligible to be pilots in the U.S. Air Force.

YOUR TURN. Do Exercise 15 "Pulse Rates."

Finding Values from Known Areas

Here are helpful hints for those cases in which the area (or probability or percentage) is known and we must find the relevant value(s):

1. Graphs are extremely helpful in visualizing, understanding, and successfully working with normal probability distributions, so they should always be used.
2. *Don't confuse z scores and areas.* Remember, z scores are *distances* along the horizontal scale, but areas are *regions* under the normal curve. Table A-2 lists z scores in the left columns and across the top row, but areas are found in the body of the table.

continued

3. *Choose the correct (right/left) side of the graph.* A value separating the *top* 10% from the others will be located on the right side of the graph, but a value separating the *bottom* 10% will be located on the left side of the graph.
4. A z score must be *negative* whenever it is located in the *left* half of the normal distribution.
5. Areas (or probabilities) are always between 0 and 1, and they are never negative.

Procedure for Finding Values from Known Areas or Probabilities

1. Sketch a normal distribution curve, write the given probability or percentage in the appropriate region of the graph, and identify the x value(s) being sought.
2. If using technology, refer to the Tech Center instructions at the end of this section. If using Table A-2, refer to the *body* of Table A-2 to find the area to the left of x, then identify the z score corresponding to that area.
3. If you know z and must convert to the equivalent x value, use Formula 6-2 by entering the values for μ, σ, and the z score found in Step 2, then solve for x. Based on Formula 6-2, we can solve for x as follows:

$$x = \mu + (z \cdot \sigma) \quad \text{(another form of Formula 6–2)}$$

(If z is located to the left of the mean, be sure that it is a negative number.)

4. Refer to the sketch of the curve to verify that the solution makes sense in the context of the graph and in the context of the problem.

The following example uses this procedure for finding a value from a known area.

EXAMPLE 3 Designing a Front Door for a Home

When designing equipment, one common criterion is to use a design that accommodates at least 95% of the population. What is the height of a door that would allow 95% of adults to walk through the doorway without bending or hitting their heads? Based on Data Set 1 "Body Data" in Appendix B, assume that heights of adults are normally distributed with a mean of 66.2 in. and a standard deviation of 3.8 in. How does the result compare to the door height of 80 in. required by the International Residential Code?

SOLUTION

Step 1: Figure 6-14 on the next page shows the normal distribution with the height x that we want to identify. The shaded area represents the shortest 95% of adults.

Step 2: Technology: Technology will provide the value of x in Figure 6-14. For example, see the Excel display on the next page showing that $x = 72.45044378$ in., or 72.5 in. when rounded.

Table A-2: If using Table A-2, search for an area of 0.9500 *in the body* of the table. (The area of 0.9500 shown in Figure 6-14 is a cumulative area from the left, and that is exactly the type of area listed in Table A-2.) The area of 0.9500 is between the Table A-2 areas of 0.9495 and 0.9505, but there is an asterisk and footnote indicating that an area of 0.9500 corresponds to $z = 1.645$.

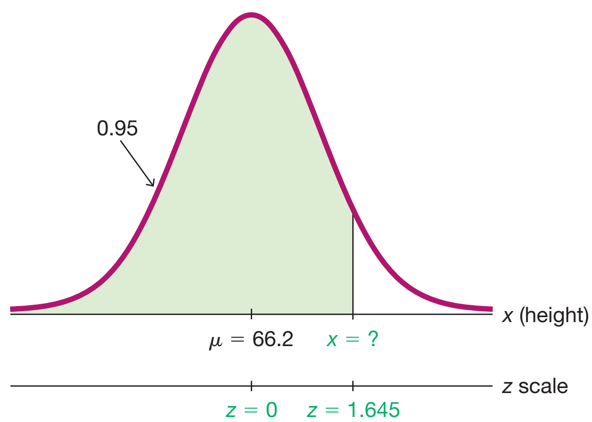
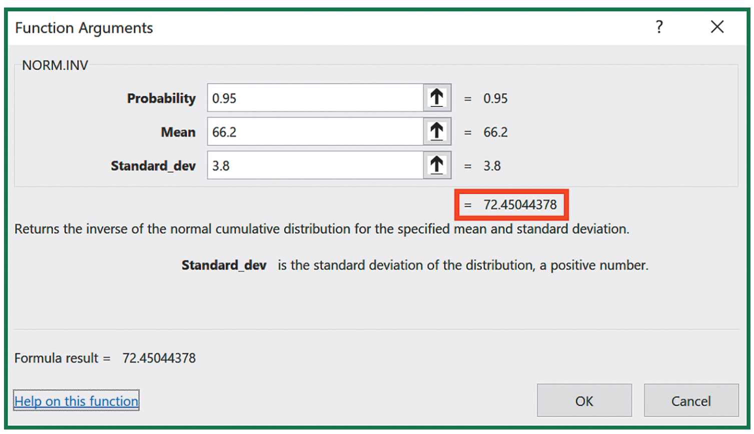


FIGURE 6-14 Finding the 95th Percentile

EXCEL



Step 3: With $z = 1.645$, $\mu = 66.2$ in., and $\sigma = 3.8$ in., we can solve for x by using Formula 6-2:

$$z = \frac{x - \mu}{\sigma} \quad \text{becomes} \quad 1.645 = \frac{x - 66.2}{3.8}$$

The result of $x = 72.451$ in. can be found directly with a little algebra, or that result can be found by using the following version of Formula 6-2:

$$x = \mu + (z \cdot \sigma) = 66.2 + (1.645 \cdot 3.8) = 72.451$$

Step 4: The solution of $x = 72.5$ in. (rounded) in Figure 6-14 is reasonable because it is greater than the mean of 66.2 in.

INTERPRETATION

A door height of 72.5 in. would allow 95% of adults to walk through without bending or hitting heads. The door height of 72.5 in. is well below the height of 80 in. required by the International Residential Code, so more than 95% of adults could walk through doors meeting this requirement without bending or hitting their heads.

YOUR TURN. Do Exercise 21 "Pulse Rates."

Significance

In Chapter 4 we saw that probabilities can be used to determine whether values are *significantly high* or *significantly low*. Chapter 4 referred to x successes among n trials, but we can adapt those criteria to apply to continuous variables as follows:

Significantly high: The value x is *significantly high* if $P(x \text{ or greater}) \leq 0.05$.*

Significantly low: The value x is *significantly low* if $P(x \text{ or less}) \leq 0.05$.*

*The value of 0.05 is not absolutely rigid, and other values such as 0.01 could be used instead.

EXAMPLE 4 Significantly Low Birth Weights

Use the preceding criteria to identify significantly low birth weights (grams) of males based on Data Set 6 "Births" in Appendix B. Assume that males have normally distributed birth weights with a mean of 3272.8 g and a standard deviation of 660.2 g.

SOLUTION

Step 1: We begin with the graph shown in Figure 6-15. We have entered the mean of 3272.8 and we have identified the x value separating the lowest 5% of male birth weights.

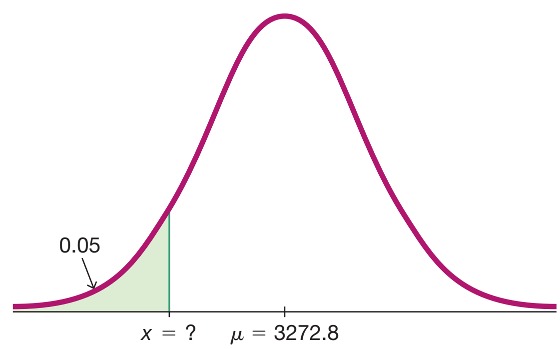


FIGURE 6-15 Male Birth Weights

Step 2: **Technology:** To use technology, refer to the instructions at the end of this section. Technology will show that the value of x in Figure 6-15 is 2186.9 g.

Table A-2: If using Table A-2, we must work with cumulative areas from the left. For the value of x, the cumulative area from the left is 0.05, so search for an area of 0.05 *in the body* of the table to get $z = -1.645$ (identified by the asterisk between 0.0505 and 0.0495). Having found the z score, we now proceed to convert it to a birth weight.

Step 3: We now solve for x by using Formula 6-2 directly or by using the following version of Formula 6-2:

$$\text{Value of } x\text{:}\ \ x = \mu + (z \cdot \sigma) = 3272.8 + (-1.645 \cdot 660.2) = 2186.8 \text{ g}$$

Step 4: Referring to Figure 6-15, we see that the value of $x = 2186.9$ g (or 2186.8 g if using Table A-2) is reasonable because it is less than the mean of 3272.8 g.

INTERPRETATION

The birth weight of 2186.9 g is on the borderline that separates significantly low male birth weights from male birth weights that are not significantly low. (The World Health Organization uses 2500 g as the cutoff for low birth weights of males and females.) Babies with low birth weights are often given special treatments, such as care in the Neonatal Intensive Care Unit or the use of a temperature-controlled bed.

YOUR TURN. Do Exercise 23 "Significance."

TECH CENTER

Finding *x* Values/Areas

Access tech supplements, videos, and data sets at **www.TriolaStats.com**

Statdisk

1. Click **Analysis** in the top menu.
2. Select **Probability Distributions** from the dropdown menu and select **Normal Distribution** from the submenu.
3. Enter the desired *z* value or cumulative area from the left of the *z* score and click **Evaluate.**

TIP: Statdisk does not work directly with nonstandard normal distributions, so use corresponding *z* scores.

Minitab

1. Click **Calc** in the top menu.
2. Select **Probability Distributions** from the dropdown menu and select **Normal** from the submenu.

Finding Cumulative Area to the Left of an *x* Value

3. Select **Cumulative probability,** enter the mean and standard deviation.
4. Select **Input constant,** enter the desired *x* value, and click **OK.**

Finding *x* Value from a Known Probability

3. Select **Inverse cumulative probability,** enter the mean and standard deviation.
4. Select **Input constant,** enter the total area to the left of the *x* value, and click **OK.**

StatCrunch

1. Click **Stat** in the top menu.
2. Select **Calculators** from the dropdown menu and **Normal** from the submenu.
3. In the calculator box enter the mean and standard deviation.
4. Enter the desired *x* value (middle box) or probability (rightmost box).
5. Click **Compute**.

TI-83/84 Plus Calculator

Unlike most other technologies, the TI-83/84 Plus bases areas on the region between two z scores, rather than cumulative regions from the left.

Finding Area Between Two *x* Values

1. Press 2ND then VARS keys to access the *DISTR* (distributions) menu.
2. Select **normalcdf** and press ENTER.
3. Enter the desired *lower x* value and *upper x* value. Enter the mean (μ) and standard deviation (σ) to complete the command **normalcdf(lower *x*,upper *x*, μ, σ).** Press ENTER.

TIP: If there is no lower *x* value, enter −99999999; if there is no upper *x* value, enter 99999999.

Finding *x* Value Corresponding to a Known Area

1. Press 2ND then VARS keys to access the *DISTR* (distributions) menu.
2. Select **invNorm** and press ENTER.
3. Enter the *area* to the left or right of the *x* value, enter the mean (μ), and the standard deviation (σ). For *Tail* select the tail where the area is located (*Left* or *Right*) relative to the *x* value. The completed command is **InvNorm(area, μ, σ,TAIL)**. Press ENTER.

TIP: The TI-83 Plus only calculates area to the left of the *x* score. The completed command is **InvNorm(area, μ, σ).**

Excel

Finding Cumulative Area to the Left of an *x* Value

1. Click **Insert Function f_x,** select the category **Statistical,** select the function **NORM.DIST,** and **click OK.**
2. For *x* enter the *x* value, enter *Mean*, enter *Standard_dev*, and enter **1** for *Cumulative*.
3. Click **OK.**

Finding *x* Value Corresponding to a Known Probability

1. Click **Insert Function f_x,** select the category **Statistical,** select the function **NORM.INV,** and click **OK.**
2. Enter the probability or the area to the left of the desired *x* value, enter *Mean,* and enter *Standard_dev*.
3. Click **OK.**

R

R command not available at time of publication.

R is rapidly evolving, and an updated list of statistical commands is available at TriolaStats.com

6-2 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Hershey Kisses Based on Data Set 38 “Candies” in Appendix B, weights of the chocolate in Hershey Kisses are normally distributed with a mean of 4.5338 g and a standard deviation of 0.1039 g.

a. What are the values of the mean and standard deviation after converting all weights of Hershey Kisses to z scores using $z = (x - \mu)/\sigma$?

b. The original weights are in grams. What are the units of the corresponding z scores?

2. Hershey Kisses Based on Data Set 38 “Candies” in Appendix B, weights of the chocolate in Hershey Kisses are normally distributed with a mean of 4.5338 g and a standard deviation of 0.1039 g.

a. For the bell-shaped graph of the normal distribution of weights of Hershey Kisses, what is the area under the curve?

b. What is the value of the median?

c. What is the value of the mode?

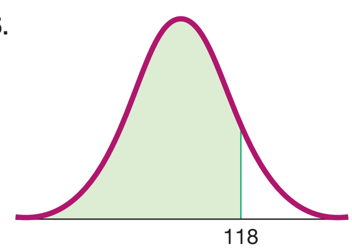
d. What is the value of the variance?

3. Normal Distributions Is the distribution described in the preceding exercise a standard normal distribution? Why or why not?

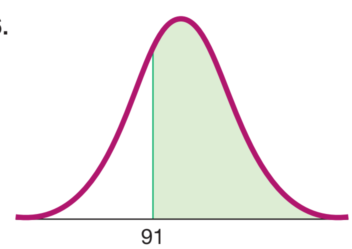
4. Mega Millions The Mega Millions lottery game requires that a number between 1 and 25 is randomly selected. What is the shape of the distribution of those selected numbers? Is it correct to say that because these selected numbers are the result of the normal selection procedure used for every drawing, the distribution of the selected numbers is a normal distribution?

IQ Scores. ***In Exercises 5–8, find the area of the shaded region. The graphs depict IQ scores of adults, and those scores are normally distributed with a mean of 100 and a standard deviation of 15 (as on the Wechsler IQ test).***

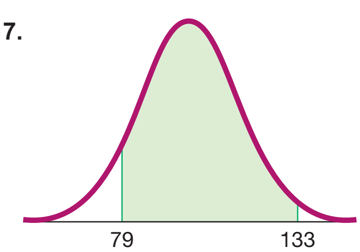
5.



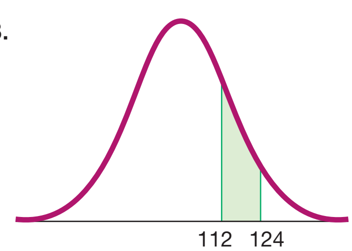
6.



7.

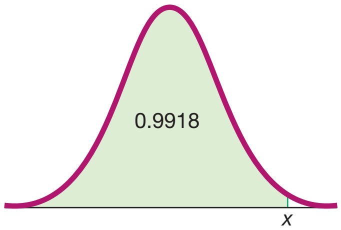


8.

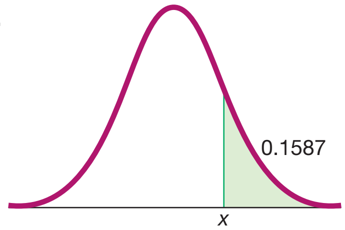


IQ Scores. ***In Exercises 9–12, find the indicated IQ score and round to the nearest whole number. The graphs depict IQ scores of adults, and those scores are normally distributed with a mean of 100 and a standard deviation of 15 (as on the Wechsler IQ test).***

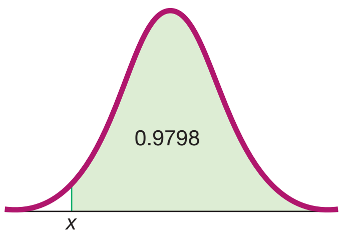
9.



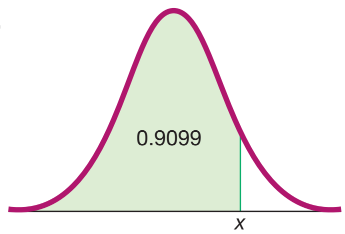
10.



11.



12.



Pulse Rates. ***In Exercises 13–24, use the data in the table below for pulse rates of adult males and females (based on Data Set 1 "Body Data" in Appendix B). Hint: Draw a graph in each case.***

Pulse Rate (beats per minute)

	Mean	St. Dev.	Distribution
Males	69.6	11.3	Normal
Females	74.0	12.5	Normal

13. Find the probability that a male has a pulse rate less than 60 beats per minute.

14. Find the probability that a female has a pulse rate less than 60 beats per minute.

15. Find the probability that a female has a pulse rate greater than 80 beats per minute.

16. Find the probability that a male has a pulse rate greater than 80 beats per minute.

17. Find the probability that a female has a pulse rate between 70 beats per minute and 90 beats per minute.

18. Find the probability that a male has a pulse rate between 65 beats per minute and 85 beats per minute.

19. Find the probability that a male has a pulse rate between 70 beats per minute and 90 beats per minute.

20. Find the probability that a female has a pulse rate between 60 beats per minute and 70 beats per minute.

21. For males, find P_{90}, which is the pulse rate separating the bottom 90% from the top 10%.

22. For females, find the first quartile Q_1, which is the pulse rate separating the bottom 25% from the top 75%.

23. Significance Instead of using 0.05 for identifying significant values, use the criteria that a value x is *significantly high* if $P(x \text{ or greater}) \leq 0.01$ and a value is *significantly low* if $P(x \text{ or less}) \leq 0.01$. Find the pulse rates for males that separate significant pulse rates from those that are not significant. Using these criteria, is a male pulse rate of 90 beats per minute significantly high?

24. Significance Instead of using 0.05 for identifying significant values, use the criteria that a value x is *significantly high* if $P(x \text{ or greater}) \leq 0.025$ and a value is *significantly low* if $P(x \text{ or less}) \leq 0.025$. Find the female pulse rates separating significant values from those that are not significant. Using these criteria, is a female pulse rate of 48 beats per minute significantly low?

In Exercises 25–28, use these parameters (based on Data Set 1 "Body Data" in Appendix B):

- ***Men's heights are normally distributed with mean 68.6 in. and standard deviation 2.8 in.***
- ***Women's heights are normally distributed with mean 63.7 in. and standard deviation 2.9 in.***

25. Navy Pilots The U.S. Navy requires that fighter pilots have heights between 62 in. and 78 in.

a. Find the percentage of women meeting the height requirement. Are many women not qualified because they are too short or too tall?

b. If the Navy changes the height requirements so that all women are eligible except the shortest 3% and the tallest 3%, what are the new height requirements for women?

26. Air Force Pilots The U.S. Air Force required that pilots have heights between 64 in. and 77 in.

a. Find the percentage of men meeting the height requirement.

b. If the Air Force height requirements are changed to exclude only the tallest 2.5% of men and the shortest 2.5% of men, what are the new height requirements?

27. Mickey Mouse Disney World requires that people employed as a Mickey Mouse character must have a height between 56 in. and 62 in.

a. Find the percentage of men meeting the height requirement. What does the result suggest about the genders of the people who are employed as Mickey Mouse characters?

b. If the height requirements are changed to exclude the tallest 50% of men and the shortest 5% of men, what are the new height requirements?

28. Snow White Disney World requires that women employed as a Snow White character must have a height between 64 in. and 67 in.

a. Find the percentage of women meeting the height requirement.

b. If the height requirements are changed to exclude the shortest 40% of women and the tallest 5% of women, what are the new height requirements?

29. Designing Helmets Engineers must consider the circumferences of adult heads when designing motorcycle helmets. Adult head circumferences are normally distributed with a mean of 570.0 mm and a standard deviation of 18.3 mm (based on Data Set 3 "ANSUR II 2012"). Due to financial constraints, the helmets will be designed to fit all adults except those with head circumferences that are in the smallest 5% or largest 5%. Find the minimum and maximum head circumferences that the helmets will fit.

30. Designing a Desk A common design requirement is that an environment must fit the range of people who fall between the 5th percentile for women and the 95th percentile for men. In designing a desk, we must consider *sitting knee height*, which is the distance from the bottom of the feet to the top of the knee. Males have sitting knee heights that are normally distributed with a mean of 21.4 in. and a standard deviation of 1.2 in.; females have sitting knee heights that are normally distributed with a mean of 19.6 in. and a standard deviation of 1.1 in. (based on data from the Department of Transportation).

a. What is the minimum desk clearance height required to satisfy the requirement of fitting 95% of men? Why is the 95th percentile for women ignored in this case?

b. The author is writing this exercise at a desk with a clearance of 23.5 in. above the floor. What percentage of men fit this desk, and what percentage of women fit this desk? Does the desk appear to be made to fit almost everyone?

31. Aircraft Seat Width Engineers want to design seats in commercial aircraft so that they are wide enough to fit 99% of all adults. (Accommodating 100% of adults would require very wide seats that would be much too expensive.) Assume adults have hip widths that are normally distributed with a mean of 14.3 in. and a standard deviation of 0.9 in. (based on data from *Applied Ergonomics*). Find P_{99}. That is, find the hip width for adults that separates the smallest 99% from the largest 1%.

32. Freshman 15 The term "Freshman 15" refers to the claim that college students typically gain 15 lb during their freshman year at college. Assume that the amounts of weight that male college students gain during their freshman year are normally distributed with a mean of 2.6 lb and a standard deviation of 10.8 lb (based on Data Set 13 "Freshman 15"). Find the probability that a randomly selected male college student gains 15 lb or more during his freshman year. What does the result suggest about the claim of the "Freshman 15"?

33. Durations of Pregnancies The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days.

a. In a letter to "Dear Abby," a wife claimed to have given birth 308 days after a brief visit from her husband, who was working in another country. Find the probability of a pregnancy lasting 308 days or longer. What does the result suggest?

b. If we stipulate that a baby is *premature* if the duration of pregnancy is in the lowest 3%, find the duration that separates premature babies from those who are not premature. Premature babies often require special care, and this result could be helpful to hospital administrators in planning for that care.

34. Body Temperatures Based on the sample results in Data Set 5 "Body Temperatures" in Appendix B, assume that human body temperatures are normally distributed with a mean of 98.20°F and a standard deviation of 0.62°F.

a. According to emedicinehealth.com, a body temperature of 100.4°F or above is considered to be a fever. What percentage of normal and healthy persons would be considered to have a fever? Does this percentage suggest that a cutoff of 100.4°F is appropriate?

b. Physicians want to select a minimum temperature for requiring further medical tests. What should that temperature be, if we want only 2.0% of healthy people to exceed it? (Such a result is a *false positive*, meaning that the test result is positive, but the subject is not really sick.)

35. Designing Cars The sitting height of drivers must be considered in the design of a new car. The sitting height is measured from the seat to the top of the driver's head, and adults have sitting heights that are normally distributed with a mean of 35.4 in. and a standard deviation of 1.8 in. (based on Data Set 3 "ANSUR II 2012"). Engineers have provided plans that can accommodate adults with sitting heights up to 38.9 in., but taller adults cannot fit. Find the percentage of adults with sitting heights that fit this car. Based on the result, is the engineering design feasible?

36. Water Taxi Safety When a water taxi sank in Baltimore's Inner Harbor, an investigation revealed that the safe passenger load for the water taxi was 3500 lb. It was also noted that the mean weight of a passenger was assumed to be 140 lb. Assume a "worst-case" scenario in which all of the passengers are adult men. Assume that weights of men are normally distributed with a mean of 188.6 lb and a standard deviation of 38.9 lb (based on Data Set 1 "Body Data" in Appendix B).

a. If one man is randomly selected, find the probability that he weighs less than 174 lb (the new value suggested by the National Transportation and Safety Board).

b. With a load limit of 3500 lb, how many male passengers are allowed if we assume a mean weight of 140 lb?

c. With a load limit of 3500 lb, how many male passengers are allowed if we assume the updated mean weight of 188.6 lb?

d. Why is it necessary to periodically review and revise the number of passengers that are allowed to board?

6-2 Beyond the Basics

37. Curving Test Scores A professor gives a test and the scores are normally distributed with a mean of 60 and a standard deviation of 12. She plans to curve the scores.

a. If she curves by adding 15 to each grade, what is the new mean and standard deviation?

b. Is it fair to curve by adding 15 to each grade? Why or why not?

c. If the grades are curved so that grades of B are given to scores above the bottom 70% and below the top 10%, find the numerical limits for a grade of B.

d. Which method of curving the grades is fairer: adding 15 to each original score or using a scheme like the one given in part (c)? Explain.

38. Outliers For the purposes of constructing modified boxplots as described in Section 3-3, outliers are defined as data values that are above Q_3 by an amount greater than $1.5 \times \text{IQR}$ or below Q_1 by an amount greater than $1.5 \times \text{IQR}$, where IQR is the interquartile range. Using this definition of outliers, find the probability that when a value is randomly selected from a normal distribution, it is an outlier.

6-3 Sampling Distributions and Estimators

Key Concept We now consider the concept of a *sampling distribution of a statistic*. Instead of working with values from the original population, we want to focus on the values of *statistics* (such as sample proportions or sample means) obtained from the population. Figure 6-16 shows the key points that we need to know, so try really, really hard to understand the story that Figure 6-16 tells.

A Statistics Story Among the population of all adults, exactly 70% do not feel comfortable in a self-driving vehicle (the author just knows this). In a TE Connectivity survey of 1000 adults, 69% said that they did not feel comfortable in a self-driving vehicle. Empowered by visions of multitudes of driverless cars, 50,000 people became so enthusiastic that they each conducted their own individual survey of 1000 randomly selected adults on the same topic. Each of these 50,000 newbie surveyors reported the percentage that they found, with results such as 68%, 72%, 70%. The author obtained each of the 50,000 sample percentages, he changed them to proportions, and then he constructed the histogram shown in Figure 6-17. Notice anything about the *shape* of the histogram? It's *normal* (unlike the 50,000 newbie surveyors). Notice anything about the mean of the sample proportions? They are centered about the value of 0.70, which happens to be the population proportion. Moral:

When samples of the same size are taken from the same population, the following two properties apply:

1. **Sample proportions tend to be normally distributed.**
2. **The mean of sample proportions is the same as the population mean.**

The implications of the preceding properties will be extensive in the chapters that follow. What a happy ending!

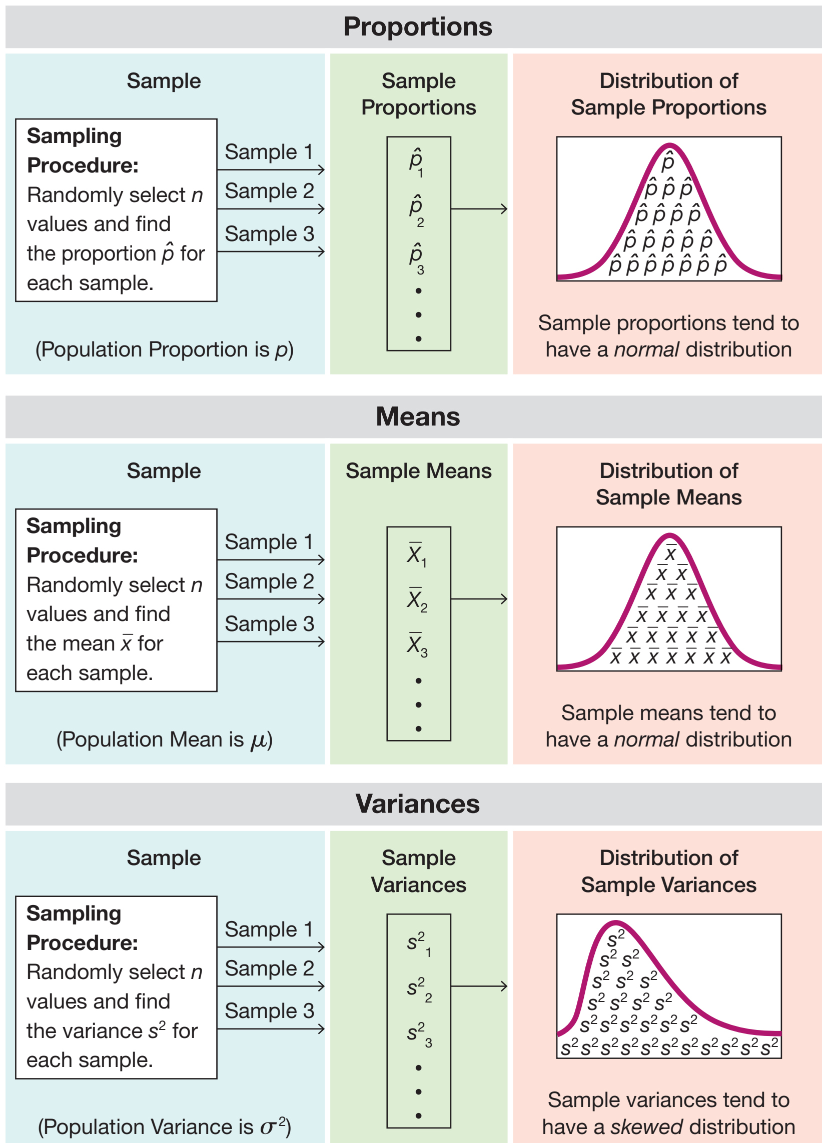


FIGURE 6-16 General Behavior of Sampling Distributions

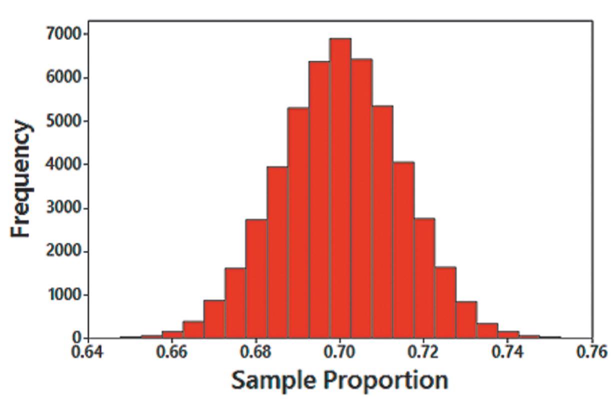


FIGURE 6-17 Histogram of 50,000 Sample Proportions

Let's formally define *sampling distribution,* the main character in the preceding statistics story.

> **DEFINITION**
>
> The **sampling distribution of a statistic** (such as a sample proportion or sample mean) is the distribution of all values of the statistic when all possible samples of the same size n are taken from the same population. (The sampling distribution of a statistic is typically represented as a probability distribution in the format of a probability histogram, formula, or table.)

Sampling Distribution of Sample Proportion

The preceding general definition of a sampling distribution of a statistic can now be restated for the specific case of a sample proportion:

> **DEFINITION**
>
> The **sampling distribution of the sample proportion** is the distribution of sample proportions (or the distribution of the variable $\hat{p}$), with all samples having the same sample size n taken from the same population. (The sampling distribution of the sample proportion is typically represented as a probability distribution in the format of a probability histogram, formula, or table.)

We need to distinguish between a population proportion p and a sample proportion, and the following notation is common and will be used throughout the remainder of this book, so it's very important.

Notation for Proportions

$x =$ number of successes

$n =$ sample size

$N =$ population size

$\hat{p} = x/n$ denotes the *sample proportion*

$p = x/N$ denotes the *population proportion*

> **HINT** $\hat{p}$ is pronounced "p-hat." When symbols are used above a letter, as in $\bar{x}$ and $\hat{p}$, they represent *statistics,* not parameters.

Behavior of Sample Proportions

1. The distribution of sample proportions tends to approximate a normal distribution.
2. Sample proportions *target* the value of the population proportion in the sense that the mean of all of the sample proportions $\hat{p}$ is equal to the population proportion p; the expected value of the sample proportion is equal to the population proportion.

EXAMPLE 1 Sampling Distribution of the Sample Proportion

Consider repeating this process: Roll a die 5 times and find the proportion of *odd* numbers (1 or 3 or 5). What do we know about the behavior of all sample proportions that are generated as this process continues indefinitely?

SOLUTION

Figure 6-18 illustrates a process of rolling a die 5 times and finding the proportion of odd numbers. (Figure 6-18 shows results from repeating this process 10,000 times, but the true sampling distribution of the sample proportion involves repeating the process indefinitely.) Figure 6-18 shows that the sample proportions are approximately normally distributed. (Because the values of 1, 2, 3, 4, 5, 6 are all equally likely, the proportion of odd numbers in the population is 0.5, and Figure 6-18 shows that the sample proportions have a mean of 0.50.)

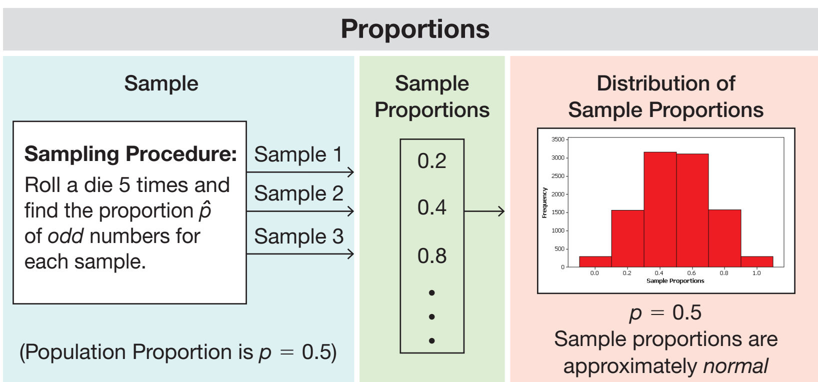


FIGURE 6-18 Sample Proportions from 10,000 Trials

YOUR TURN. Do Exercise 10 "Sampling Distribution of the Sample Proportion."

Sampling Distribution of the Sample Mean

We now consider sample means.

DEFINITION

The **sampling distribution of the sample mean** is the distribution of all possible sample means (or the distribution of the variable $\bar{x}$), with all samples having the same sample size *n* taken from the same population. (The sampling distribution of the sample mean is typically represented as a probability distribution in the format of a probability histogram, formula, or table.)

Behavior of Sample Means

1. The distribution of sample means tends to be a normal distribution. (This will be discussed further in the following section, but the distribution tends to become closer to a normal distribution as the sample size increases.)
2. The sample means *target* the value of the population mean. (That is, the mean of the sample means is the population mean. The expected value of the sample mean is equal to the population mean.)

Go Figure

90%: Percentage of the world's data that have been generated in the past two years (according to *Science Daily*).

EXAMPLE 2 Sampling Distribution of the Sample Mean

A friend of the author has three children with ages of 4, 5, and 9. Let's consider the population consisting of $\{4, 5, 9\}$. (We don't usually know all values in a population, and we don't usually work with such a small population, but it works well for the purposes of this example.) If two ages are randomly selected with replacement from the population $\{4, 5, 9\}$, identify the sampling distribution of the sample mean by creating a table representing the probability distribution of the sample mean. Do the values of the sample mean target the value of the population mean?

SOLUTION

If two values are randomly selected with replacement from the population $\{4, 5, 9\}$, the leftmost column of Table 6-2 lists the nine different possible samples. The second column lists the corresponding sample means. The nine samples are equally likely with a probability of $1/9$. We saw in Section 5-1 that a probability distribution is a description that gives the probability for each value of a random variable, as in the second and third columns of Table 6-2. The second and third columns of Table 6-2 constitute a probability distribution for the random variable representing sample means, so the second and third columns represent the sampling distribution of the sample mean. In Table 6-2, some of the sample mean values are repeated, so we combined them in Table 6-3.

TABLE 6-2 Sampling Distribution of Mean

Sample	Sample Mean $\bar{x}$	Probability
4, 4	4.0	1/9
4, 5	4.5	1/9
4, 9	6.5	1/9
5, 4	4.5	1/9
5, 5	5.0	1/9
5, 9	7.0	1/9
9, 4	6.5	1/9
9, 5	7.0	1/9
9, 9	9.0	1/9

TABLE 6-3 Sampling Distribution of Mean (Condensed)

Sample Mean $\bar{x}$	Probability
4.0	1/9
4.5	2/9
5.0	1/9
6.5	2/9
7.0	2/9
9.0	1/9

INTERPRETATION

Because Table 6-3 lists the possible values of the sample mean along with their corresponding probabilities, Table 6-3 is an example of a sampling distribution of a sample mean.

The value of the mean of the population $\{4, 5, 9\}$ is $\mu = 6.0$. Using either Table 6-2 or 6-3, we could calculate the mean of the sample values and we get 6.0. Because the mean of the sample means (6.0) is equal to the mean of the population (6.0), we conclude that the values of the sample mean do *target* the value of the population mean. It's unfortunate that this sounds so much like doublespeak, but this illustrates that *the mean of the sample means is equal to the population mean* μ.

HINT Read the last sentence of the above paragraph a few times until it makes sense.

If we were to create a probability histogram from Table 6-2, it would not have the bell shape that is characteristic of a normal distribution, but that's because we are working with such small samples. If the population of $\{4, 5, 9\}$ were much larger and if we were selecting samples much larger than $n = 2$ as in this example, we would get a probability histogram that is much closer to being bell-shaped, indicating a normal distribution, as in Example 3.

YOUR TURN. Do Exercise 11 "Sampling Distribution of the Sample Mean."

EXAMPLE 3 **Sampling Distribution of the Sample Mean**

Consider repeating this process: Roll a die 5 times to randomly select 5 values from the population $\{1, 2, 3, 4, 5, 6\}$, then find the mean $\bar{x}$ of the results. What do we know about the behavior of all sample means that are generated as this process continues indefinitely?

SOLUTION

Figure 6-19 illustrates a process of rolling a die 5 times and finding the mean of the results. Figure 6-19 shows results from repeating this process 10,000 times, but the true sampling distribution of the mean involves repeating the process indefinitely. Because the values of 1, 2, 3, 4, 5, 6 are all equally likely, the population has a mean of $\mu = 3.5$. The 10,000 sample means included in Figure 6-19 have a mean of 3.5. If the process is continued indefinitely, the mean of the sample means will be 3.5. Also, Figure 6-19 shows that the distribution of the sample means is approximately a normal distribution.

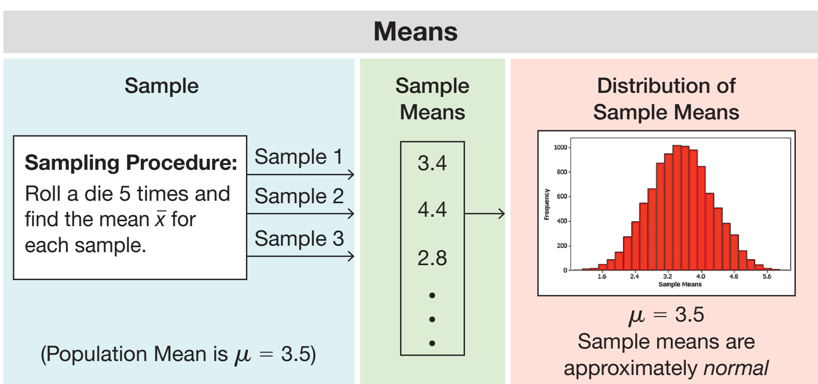


FIGURE 6-19 **Sample Means from 10,000 Trials**

Sampling Distribution of the Sample Variance

Let's now consider the sampling distribution of sample variances.

DEFINITION

The **sampling distribution of the sample variance** is the distribution of sample variances (the variable s^2), with all samples having the same sample size n taken from the same population. (The sampling distribution of the sample variance is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

CAUTION When working with population standard deviations or variances, be sure to evaluate them correctly. In Section 3-2 we saw that the computations for *population* standard deviations or variances involve division by the population size N instead of $n - 1$, as shown here.

Population standard deviation: $\sigma = \sqrt{\dfrac{\sum (x - \mu)^2}{N}}$

Population variance: $\sigma^2 = \dfrac{\sum (x - \mu)^2}{N}$

Caution: Because the calculations are typically performed with software or calculators, be careful to correctly distinguish between the variance of a sample and the variance of a population.

Behavior of Sample Variances

1. The distribution of sample variances tends to be a distribution skewed to the right.
2. The sample variances *target* the value of the population variance. (That is, the mean of the sample variances is the population variance. The expected value of the sample variance is equal to the population variance.)

EXAMPLE 4 Sampling Distribution of the Sample Variance

Consider repeating this process: Roll a die 5 times and find the variance s^2 of the results. What do we know about the behavior of all sample variances that are generated as this process continues indefinitely?

SOLUTION

Figure 6-20 illustrates a process of rolling a die 5 times and finding the variance of the results. Figure 6-20 shows results from repeating this process 10,000 times, but the true sampling distribution of the sample variance involves repeating the process indefinitely. Because the values of 1, 2, 3, 4, 5, 6 are all equally likely, the population has a variance of $\sigma^2 = 2.9$, and the 10,000 sample variances included in Figure 6-20 have a mean of 2.9. If the process is continued indefinitely, the mean of the sample variances will be 2.9. Also, Figure 6-20 shows that the distribution of the sample variances is a skewed distribution, not a normal distribution with its characteristic bell shape.

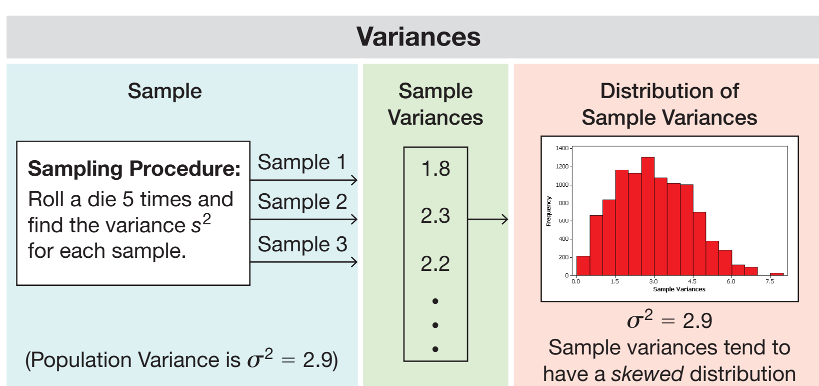


FIGURE 6-20 Sample Variances from 10,000 Trials

YOUR TURN. Do Exercise 14 "Sampling Distribution of the Variance."

Estimators: Unbiased and Biased

The preceding examples show that sample proportions, means, and variances tend to *target* the corresponding population parameters. More formally, we say that sample proportions, means, and variances are *unbiased estimators*. See the following two definitions.

DEFINITIONS

An **estimator** is a statistic used to infer (or estimate) the value of a population parameter.

An **unbiased estimator** is a statistic that targets the value of the corresponding population parameter in the sense that the sampling distribution of the statistic has a mean that is equal to the corresponding population parameter.

Unbiased Estimators These statistics are *unbiased estimators*. That is, they each target the value of the corresponding population parameter (with a sampling distribution having a mean equal to the population parameter):

- Proportion $\hat{p}$
- Mean $\bar{x}$
- Variance s^2

Biased Estimators These statistics are *biased estimators*. That is, they do *not* target the value of the corresponding population parameter:

- Median
- Range
- Standard deviation s

  Important Note: The sample standard deviations do not target the population standard deviation σ, but the bias is relatively small in large samples, so **s is often used to estimate σ** even though s is a biased estimator of σ.

EXAMPLE 5 Sampling Distribution of the Sample Range

As in Example 2, consider samples of size $n = 2$ randomly selected from the population $\{4, 5, 9\}$.

a. List the different possible samples along with the probability of each sample, then find the range for each sample.

b. Describe the sampling distribution of the sample range in the format of a table summarizing the probability distribution.

c. Based on the results, do the sample ranges target the population range, which is $9 - 4 = 5$?

d. What do these results indicate about the sample range as an estimator of the population range?

SOLUTION

a. In Table 6-4 on the next page we list the nine different possible samples of size $n = 2$ selected with replacement from the population $\{4, 5, 9\}$. The nine samples are equally likely, so each has probability $1/9$. Table 6-4 also shows the range for each of the nine samples.

continued

TABLE 6-4 Sampling Distribution of Range

Sample	Sample Range	Probability
4, 4	0	1/9
4, 5	1	1/9
4, 9	5	1/9
5, 4	1	1/9
5, 5	0	1/9
5, 9	4	1/9
9, 4	5	1/9
9, 5	4	1/9
9, 9	0	1/9

b. The last two columns of Table 6-4 list the values of the range along with the corresponding probabilities, so the last two columns constitute a table summarizing the probability distribution. Table 6-4 therefore describes the *sampling distribution* of the sample range.

c. The mean of the sample ranges in Table 6-4 is 20/9 or 2.2. The population of $\{4, 5, 9\}$ has a range of $9 - 4 = 5$. Because the mean of the sample ranges (2.2) is not equal to the population range (5), the sample ranges do *not* target the value of the population range.

d. Because the sample ranges do not target the population range, the sample range is a *biased estimator* of the population range.

INTERPRETATION

Because the sample range is a biased estimator of the population range, a sample range should generally not be used to estimate the value of the population range.

YOUR TURN. Do Exercise 13 "Sampling Distribution of the Range."

Why Sample with Replacement? All of the examples in this section involved sampling *with replacement*. Sampling *without replacement* would have the very practical advantage of avoiding wasteful duplication whenever the same item is selected more than once. Many of the statistical procedures discussed in the following chapters are based on the assumption that sampling is conducted with replacement because of the following two very important reasons.

Reasons for Sampling with Replacement

1. When selecting a relatively small sample from a large population, it makes no significant difference whether we sample with replacement or without replacement.
2. Sampling with replacement results in *independent* events that are unaffected by previous outcomes, and independent events are easier to analyze and result in simpler calculations and formulas.

6-3 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Fatal Car Crashes There are about 15,000 car crashes each day in the United States, and the proportion of car crashes that are fatal is 0.00559 (based on data from the National Highway Traffic Safety Administration). Assume that each day, 1000 car crashes are randomly selected and the proportion of fatal car crashes is recorded.

a. What do you know about the mean of the sample proportions?

b. What do you know about the shape of the distribution of the sample proportions?

2. Sampling with Replacement The Pew Research Center conducts many different surveys in the United States each year.

a. Do you think that for each individual survey, the respondents are randomly selected with replacement or without replacement?

b. Give two reasons why statistical methods tend to be based on the assumption that sampling is conducted *with* replacement, instead of without replacement.

3. Unbiased Estimators Data Set 1 "Body Data" in Appendix B includes systolic blood pressure measurements from 147 adult females. If we compute the values of sample statistics from that sample, which of the following statistics are *unbiased* estimators of the corresponding population parameters: sample mean; sample median; sample range; sample variance; sample standard deviation; sample proportion?

4. Sampling Distribution Data Set 1 "Body Data" in Appendix B includes systolic blood pressure measurements from 147 adult females. If we explore this sample by constructing a histogram and finding the mean and standard deviation, do those results describe the sampling distribution of the mean? Why or why not?

5. Good Sample? An economist is investigating the incomes of college students. Because she lives in Maine, she obtains sample data from that state. Is the resulting mean income of college students a good estimator of the mean income of college students in the United States? Why or why not?

6. College Presidents There are about 4200 college presidents in the United States, and they have annual incomes with a distribution that is skewed instead of being normal. Many different samples of 40 college presidents are randomly selected, and the mean annual income is computed for each sample.

a. What is the approximate shape of the distribution of the sample means (uniform, normal, skewed, other)?

b. What value do the sample means target? That is, what is the mean of all such sample means?

In Exercises 7–10, use the same population of {4, 5, 9} that was used in Examples 2 and 5. As in Examples 2 and 5, assume that samples of size $n = 2$ are randomly selected with replacement.

7. Sampling Distribution of the Sample Variance

a. Find the value of the population variance σ^2.

b. Table 6-2 describes the sampling distribution of the sample mean. Construct a similar table representing the sampling distribution of the sample variance s^2. Then combine values of s^2 that are the same, as in Table 6-3 (*Hint:* See Example 2 on page 276 for Tables 6-2 and 6-3, which describe the sampling distribution of the sample mean.)

c. Find the mean of the sampling distribution of the sample variance.

d. Based on the preceding results, is the sample variance an unbiased estimator of the population variance? Why or why not?

8. Sampling Distribution of the Sample Standard Deviation For the following, round results to three decimal places.

a. Find the value of the population standard deviation σ.

b. Table 6-2 describes the sampling distribution of the sample mean. Construct a similar table representing the sampling distribution of the sample standard deviation s. Then combine values of s that are the same, as in Table 6-3 (*Hint:* See Example 2 on page 276 for Tables 6-2 and 6-3, which describe the sampling distribution of the sample mean.)

c. Find the mean of the sampling distribution of the sample standard deviation.

d. Based on the preceding results, is the sample standard deviation an unbiased estimator of the population standard deviation? Why or why not?

9. Sampling Distribution of the Sample Median

a. Find the value of the population median.

b. Table 6-2 describes the sampling distribution of the sample mean. Construct a similar table representing the sampling distribution of the sample median. Then combine values of the median that are the same, as in Table 6-3. (*Hint:* See Example 2 on page 276 for Tables 6-2 and 6-3, which describe the sampling distribution of the sample mean.)

c. Find the mean of the sampling distribution of the sample median.

d. Based on the preceding results, is the sample median an unbiased estimator of the population median? Why or why not?

10. Sampling Distribution of the Sample Proportion

a. For the population, find the proportion of odd numbers.

b. Table 6-2 describes the sampling distribution of the sample mean. Construct a similar table representing the sampling distribution of the sample proportion of odd numbers. Then combine values of the sample proportion that are the same, as in Table 6-3. (*Hint:* See Example 2 on page 276 for Tables 6-2 and 6-3, which describe the sampling distribution of the sample mean.)

c. Find the mean of the sampling distribution of the sample proportion of odd numbers.

d. Based on the preceding results, is the sample proportion an unbiased estimator of the population proportion? Why or why not?

In Exercises 11–14, use the population of {2, 3, 5, 9} of the lengths of hospital stay (days) of mothers who gave birth, found from Data Set 6 "Births" in Appendix B. Assume that random samples of size $n = 2$ are selected with replacement.

11. Sampling Distribution of the Sample Mean

a. After identifying the 16 different possible samples, find the mean of each sample, and then construct a table representing the sampling distribution of the sample mean. In the table, combine values of the sample mean that are the same. (*Hint:* See Table 6-3 in Example 2 on page 276.)

b. Compare the mean of the population {2, 3, 5, 9} to the mean of the sampling distribution of the sample mean.

c. Do the sample means target the value of the population mean? In general, do sample means make good estimators of population means? Why or why not?

12. Sampling Distribution of the Median Repeat Exercise 11 using medians instead of means.

13. Sampling Distribution of the Range Repeat Exercise 11 using ranges instead of means.

14. Sampling Distribution of the Variance Repeat Exercise 11 using variances instead of means.

15. Births: Sampling Distribution of Sample Proportion When two births are randomly selected, the sample space for genders is bb, bg, gb, and gg (where b = boy and g = girl). Assume that those four outcomes are equally likely. Construct a table that describes the sampling distribution of the sample proportion of girls from two births. Does the mean of the sample proportions equal the proportion of girls in two births? Does the result suggest that a sample proportion is an unbiased estimator of a population proportion?

16. Births: Sampling Distribution of Sample Proportion Repeat Exercise 15 using three births instead of two births.

17. MCAT The Medical College Admissions Test (MCAT) is used to help screen applicants to medical schools. Like many such tests, the MCAT uses multiple-choice questions with each question having five choices, one of which is correct. Assume that you must make random guesses for two such questions. Assume that both questions have correct answers of "a."

a. After listing the 25 different possible samples, find the proportion of correct answers in each sample, then construct a table that describes the sampling distribution of the sample proportions of correct responses.

b. Find the mean of the sampling distribution of the sample proportion.

c. Is the mean of the sampling distribution [from part (b)] equal to the population proportion of correct responses? Does the mean of the sampling distribution of proportions *always* equal the population proportion?

18. Hybridization A hybridization experiment begins with four peas having yellow pods and one pea having a green pod. Two of the peas are randomly selected *with replacement* from this population.

a. After identifying the 25 different possible samples, find the proportion of peas with yellow pods in each of them, then construct a table to describe the sampling distribution of the proportions of peas with yellow pods.

b. Find the mean of the sampling distribution.

c. Is the mean of the sampling distribution [from part (b)] equal to the population proportion of peas with yellow pods? Does the mean of the sampling distribution of proportions *always* equal the population proportion?

6-3 Beyond the Basics

19. Using a Formula to Describe a Sampling Distribution Exercise 15 "Births" requires the construction of a table that describes the sampling distribution of the proportions of girls from two births. Consider the formula shown here, and evaluate that formula using sample proportions (represented by x) of 0, 0.5, and 1. Based on the results, does the formula describe the sampling distribution? Why or why not?

$$P(x) = \frac{1}{2(2 - 2x)!(2x)!} \quad \text{where } x = 0, 0.5, 1$$

20. Mean Absolute Deviation Is the mean absolute deviation of a sample a good statistic for estimating the mean absolute deviation of the population? Why or why not? (*Hint:* See Example 5.)

6-4 The Central Limit Theorem

Key Concept In the preceding section we saw that the sampling distribution of sample means tends to be a normal distribution as the sample size increases. In this section we introduce and apply the *central limit theorem*. The central limit theorem allows us to use a normal distribution for some very meaningful and important applications.

Given any population with *any* distribution (uniform, skewed, whatever), the distribution of sample means $\bar{x}$ can be approximated by a normal distribution when the samples are large enough with $n > 30$. (There are some special cases of very nonnormal distributions for which the requirement of $n > 30$ isn't quite enough, so the number 30 should be higher in those cases, but those cases are rare.)

CENTRAL LIMIT THEOREM

For all samples of the same size n with $n > 30$, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

EXAMPLE 1 Boston Commute Times

Figures 6-21 and 6-22 illustrate the central limit theorem.

- **Original Data:** Figure 6-21 is a histogram of 1000 Boston commute times (minutes) from Data Set 31 "Commute Times" in Appendix B.
- **Sample Means:** Figure 6-22 is a histogram of 1000 *sample means*, where each of the 1000 samples includes 50 Boston commute times (randomly selected from Data Set 31 "Commute Times" in Appendix B).

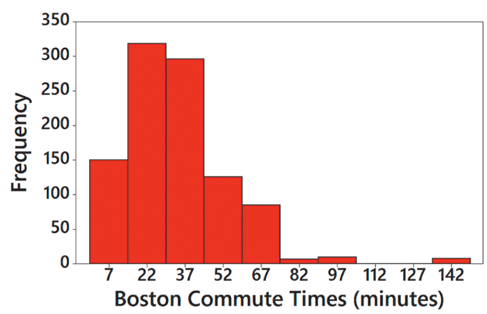


FIGURE 6-21 Nonnormal Distribution: 1000 Boston Commute Times

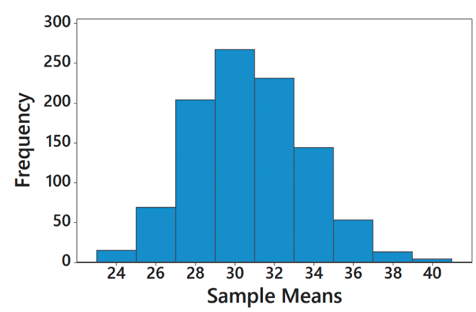


FIGURE 6-22 Approximately Normal Distribution of 1000 Sample Means

INTERPRETATION

The original Boston commute times depicted in Figure 6-21 have a skewed distribution, but when we collect samples and compute their means, those sample means tend to have a distribution that is *normal*, as in Figure 6-22.

A Universal Truth Example 1 and the central limit theorem are truly remarkable because they describe a rule of nature that works throughout the universe. If we could send a spaceship to a distant planet "in a galaxy far, far away," and if we collect samples of rocks (all of the same large sample size) and weigh them, the sample means would have a distribution that is approximately normal. Think about the significance of that!

The following key points form the foundation for estimating population parameters and hypothesis testing—topics discussed at length in the following chapters.

KEY ELEMENTS

Central Limit Theorem and the Sampling Distribution of $\bar{x}$

Given

1. Population (with any distribution) has mean μ and standard deviation σ.
2. Simple random samples all of the same size n are selected from the population.

Practical Rules for Real Applications Involving a Sample Mean $\bar{x}$

Requirements: Population has a normal distribution *or* $n > 30$:

Mean of all values of $\bar{x}$: $\mu_{\bar{x}} = \mu$

Standard deviation of all values of $\bar{x}$: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

z score conversion of $\bar{x}$: $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

Original population is *not* normally distributed *and* $n \leq 30$: The distribution of $\bar{x}$ cannot be approximated well by a normal distribution, and the methods of this section do not apply. Use other methods, such as nonparametric methods (Chapter 13) or bootstrapping methods (Section 7-4).

Considerations for Practical Problem Solving

1. **Check Requirements:** When working with the mean from a sample, verify that the normal distribution can be used by confirming that the original population has a normal distribution or the sample size is $n > 30$.
2. **Individual Value or Mean from a Sample?** Determine whether you are using a normal distribution with a *single* value x or the mean $\bar{x}$ from a sample of n values. See the following.
   - Individual value: When working with an *individual* value from a normally distributed population, use the methods of Section 6-2 with $z = \dfrac{x - \mu}{\sigma}$.
   - Mean from a sample of values: When working with a mean for some *sample* of n values, be sure to use the value of $\sigma/\sqrt{n}$ for the standard deviation of the sample means, so use $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$.

The following new notation is used for the mean and standard deviation of the distribution of $\bar{x}$.

NOTATION FOR THE SAMPLING DISTRIBUTION OF $\bar{x}$

If all possible simple random samples of size n are selected from a population with mean μ and standard deviation σ, the mean of all sample means is denoted by $\mu_{\bar{x}}$ and the standard deviation of all sample means is denoted by $\sigma_{\bar{x}}$.

Mean of all values of $\bar{x}$: $\mu_{\bar{x}} = \mu$

Standard deviation of all values of $\bar{x}$: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Note: $\sigma_{\bar{x}}$ is called the *standard error of the mean* and is sometimes denoted as SEM.

Applying the Central Limit Theorem

Many practical problems can be solved with the central limit theorem. Example 2 is a good illustration of the central limit theorem because we can see the difference between working with an *individual* value in part (a) and working with the *mean* for a sample in part (b). Study Example 2 carefully to understand the fundamental difference between the procedures used in parts (a) and (b). In particular, note that when working with an *individual* value, we use $z = \frac{x - \mu}{\sigma}$, but when working with the mean $\bar{x}$ for a collection of *sample* values, we use $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

EXAMPLE 2 Boeing 737 Airline Seats

American Airlines uses Boeing 737 jets with 126 seats in the main cabin. In an attempt to create more room, an engineer is considering a reduction of the seat width from 16.6 in. to 16.0. in. Adult males have hip widths that are normally distributed with a mean of 14.3 in. and a standard deviation of 0.9 in. (based on data from *Applied Ergonomics*).

a. Find the probability that a randomly selected adult male has a hip width greater than the seat width of 16.0 in.

b. Find the probability that 126 main cabin seats are all occupied by males with a mean hip width greater than the seat width of 16.0 in.

c. For the design of the aircraft seats, which is more relevant: The result from part (a) or the result from part (b)? Why? What do the results suggest about the reduction of the seat width to 16.0 in.?

SOLUTION

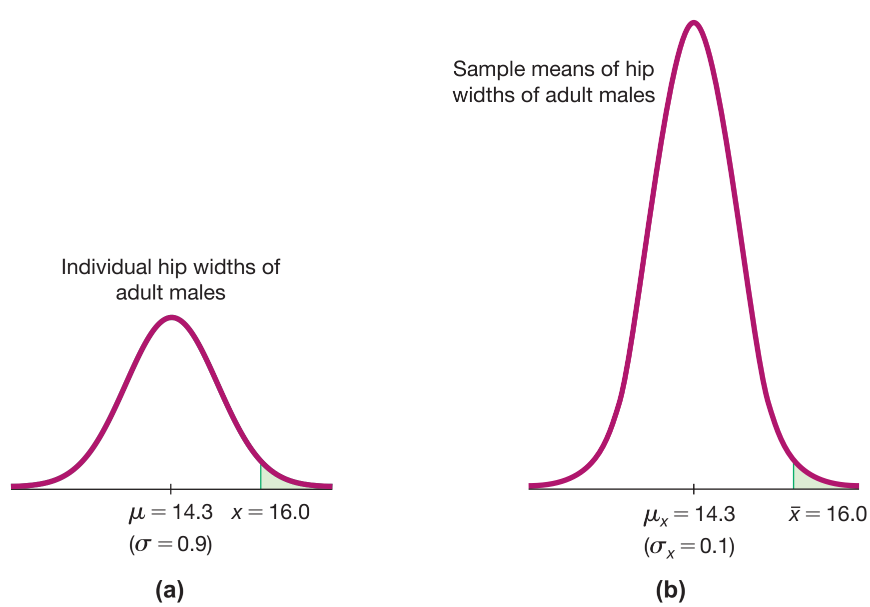


FIGURE 6-23 **Hip Widths of Adult Males**

a. Approach Used for an Individual Value: Use the methods presented in Section 6-2 because we are dealing with an *individual* value from a normally distributed population. We seek the area of the green-shaded region in Figure 6-23(a).

Technology: If using technology (as described at the end of Section 6-2), we find that the green-shaded area is 0.0295.

Table A-2: If using Table A-2, we convert the hip width of 16.0 in. to the corresponding z score of $z = 1.89$, as shown here:

$$z = \frac{x - \mu}{\sigma} = \frac{16.0 - 14.3}{0.9} = 1.89$$

We refer to Table A-2 to find that the cumulative area to the *left* of $z = 1.89$ is 0.9706, so the green-shaded area in Figure 6-23 is $1 - 0.9706 = 0.0294$. The result of 0.0295 from technology is more accurate.

b. Approach Used for the Mean of Sample Values: Because we are dealing with the *mean* of a sample of 126 males and not an individual male, use the central limit theorem.

REQUIREMENT CHECK FOR PART B We can use the normal distribution if the original population is normally distributed or $n > 30$. The sample size is greater than 30, so samples of any size will yield means that are normally distributed. ✓

Because we are now dealing with a distribution of sample means, we must use the parameters $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, which are evaluated as follows:

$$\mu_{\bar{x}} = \mu = 14.3$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.9}{\sqrt{126}} = 0.1$$

We want to find the green-shaded area shown in Figure 6-23(b). (*Note:* Figure 6-23(b) is not drawn to scale. If Figure 6-23(b) had been true to scale, the green-shaded area would not be visible and the normal curve would be much thinner and taller.)

Technology: If using technology, the green-shaded area in Figure 6-23(b) is 0+, which is a really small positive number.

Table A-2: If using Table A-2, we convert the value of $\bar{x} = 16.0$ to the corresponding z score of $z = 17.00$, as shown here:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{16.0 - 14.3}{0.1} = 17.00$$

From Table A-2 we find that the cumulative area to the left of $z = 17$ is 0.9999, so the green-shaded area of Figure 6-23(b) is $1 - 0.9999 = 0.0001$. It is highly unlikely that the 126 adult males will have a mean hip width greater than 16.0 in.

INTERPRETATION

The result from part (a) is more relevant for the design of the aircraft seats. Individual seats will be occupied by individual passengers, not groups of passengers. The result from part (a) shows that with a 0.0295 probability, approximately 3% of adult males would have hip widths greater than the seat width. Although 3% appears to be small, that would result in several passengers on every flight that would somehow require a special accommodation, and that would likely lead to significant challenges for passengers and the flight crew. The reduction of the seat width to 16.0 in. does not appear to be feasible.

YOUR TURN. Do Exercise 5 "Using the Central Limit Theorem."

The Fuzzy Central Limit Theorem

In *The Cartoon Guide to Statistics*, by Gonick and Smith, the authors describe the Fuzzy Central Limit Theorem as follows: "Data that are influenced by many small and unrelated random effects are approximately normally distributed. This explains why the normal is everywhere: stock market fluctuations, student weights, yearly temperature averages, SAT scores: All are the result of many different effects." People's heights, for example, are the results of hereditary factors, environmental factors, nutrition, health care, geographic region, and other influences, which, when combined, produce normally distributed values.

Example 2 shows that we can use the same basic procedures from Section 6-2, but we must remember to correctly adjust the standard deviation when working with a *sample mean* instead of an individual sample value. The calculations used in Example 2 are exactly the type of calculations used by engineers when they design elevators, ski lifts, escalators, airplanes, boats, amusement park rides, and other devices that carry people.

Introduction to Hypothesis Testing

Carefully examine the conclusions that are reached in the next example illustrating the type of thinking that is the basis for the important procedure of hypothesis testing (formally introduced in Chapter 8). Example 3 uses the rare event rule for inferential statistics, first presented in Section 4-1:

Identifying Significant Results with Probabilities: The Rare Event Rule for Inferential Statistics

If, under a given assumption, the probability of a particular observed event is very small and the observed event occurs *significantly less than* or *significantly greater than* what we typically expect with that assumption, we conclude that the assumption is probably not correct.

The following example illustrates the above rare event rule and it uses the author's all-time favorite data set. This example also illustrates the type of reasoning that is used for the important method of *hypothesis testing,* which is formally introduced in Chapter 8.

EXAMPLE 3 Body Temperatures

Assume that the population of human body temperatures has a mean of 98.6°F, as is commonly believed. Also assume that the population standard deviation is 0.62°F (based on Data Set 5 "Body Temperatures" in Appendix B). A sample of size $n = 106$ subjects was randomly selected and the mean body temperature of 98.2°F was obtained. If the mean body temperature is really 98.6°F, find the probability of getting a sample mean of 98.2°F or lower for a sample of size 106. Based on the result, is 98.2°F *significantly low*? What do these results suggest about the common belief that the mean body temperature is 98.6°F?

SOLUTION

We work under the assumption that the population of human body temperatures has a mean of 98.6°F. We weren't given the distribution of the population, but because the sample size $n = 106$ exceeds 30, we use the central limit theorem and conclude that the distribution of sample means is a normal distribution with these parameters:

$$\mu_{\bar{x}} = \mu = 98.6 \text{ (by assumption)}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.62}{\sqrt{106}} = 0.0602197$$

Figure 6-24 shows the shaded area (see the tiny left tail of the graph) corresponding to the probability we seek. Having already found the parameters that apply to the distribution shown in Figure 6-24, we can now find the shaded area by using the same procedures developed in Section 6-2.

Technology: If we use technology to find the shaded area in Figure 6-24, we get 0.0000000000155, which can be expressed as 0+.

Table A-2: If we use Table A-2 to find the shaded area in Figure 6-24, we must first convert the sample mean of $\bar{x} = 98.2°F$ to the corresponding z score:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{98.2 - 98.6}{0.0602197} = -6.64$$

Referring to Table A-2, we find that $z = -6.64$ is off the chart, but for values of z below -3.49, we use an area of 0.0001 for the cumulative left area up to $z = -3.49$. We therefore conclude that the shaded region in Figure 6-24 is 0.0001.

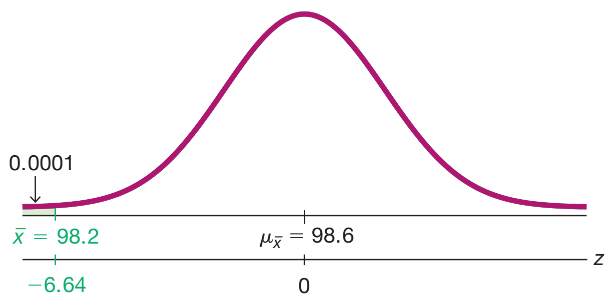


FIGURE 6-24 Means of Body Temperatures from Samples of Size $n = 106$

INTERPRETATION

The result shows that if the mean of our body temperatures is really 98.6°F, as we assumed, then there is an extremely small probability of getting a sample mean of 98.2°F or lower when 106 subjects are randomly selected. The sample mean of 98.2°F is *significantly low.* University of Maryland researchers did obtain such a sample mean, and after confirming that the sample is sound, there are two feasible explanations: (1) The population mean really is 98.6°F and their sample represents a chance event that is extremely rare; (2) the population mean is actually lower than the assumed value of 98.6°F and so their sample is typical. Because the probability is so low, it is more reasonable to conclude that the common belief of 98.6°F for the mean body temperature is a belief that is incorrect. Based on the sample data, we should reject the belief that the mean body temperature is 98.6°F.

YOUR TURN. Do Exercise 9 "Safe Loading of Elevators."

Not *Exactly,* but "At Least as Extreme"

In Example 3, we assumed that the mean body temperature is $\mu = 98.6°F$, and we determined that the probability of getting the sample mean of $\bar{x} = 98.2°F$ *or lower* is 0.0001, which suggests that the true mean body temperature is actually less than 98.6°F. In the context of Example 3, the sample mean of 98.2°F is *significantly low* not because the probability of *exactly* 98.2°F is low, but because the probability of 98.2°F *or lower* is small. (See Section 5-1 for the discussion of "Not *Exactly,* but At Least as Extreme.")

Correction for a Finite Population

In applying the central limit theorem, our use of $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ assumes that the population has infinitely many members. When we sample with replacement, the population is effectively infinite. When sampling without replacement from a finite population, we may need to adjust $\sigma_{\bar{x}}$. Here is a common rule:

> **When sampling without replacement and the sample size n is greater than 5% of the finite population size N (that is, $n > 0.05N$), adjust the standard deviation of sample means $\sigma_{\bar{x}}$ by multiplying it by this *finite population correction factor:***

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

Except for Exercise 21 "Correcting for a Finite Population," the examples and exercises in this section assume that the finite population correction factor does *not* apply, because we are sampling with replacement, or the population is infinite, or the sample size doesn't exceed 5% of the population size.

6-4 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Requirements Medical researchers once conducted experiments to determine whether Lisinopril is a drug that is effective in lowering systolic blood pressure of patients. Patients in a treatment group had their systolic blood pressure measured after being treated with Lisinopril. Under what conditions can the mean systolic blood pressure of this sample be treated as a value that is from a population having a normal distribution?

2. Small Sample Weights of M&M plain candies are normally distributed. Twelve M&M plain candies are randomly selected and weighed, and then the mean of this sample is calculated. Is it correct to conclude that the resulting sample mean cannot be considered to be a value from a normally distributed population because the sample size of 12 is too small? Explain.

3. Notation In general, what do the symbols $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ represent? What are the values of $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ for samples of size 36 randomly selected from the population of IQ scores with population mean of 100 and standard deviation of 15?

4. Incomes of Statistics Students Annual incomes of statistics students are known to have a distribution that is skewed to the right instead of being normally distributed. Assume that we collect a random sample of annual incomes of 50 statistics students. Can the distribution of incomes in that sample be approximated by a normal distribution because the sample is large? Why or why not?

Using the Central Limit Theorem. ***In Exercises 5–8, assume that the amounts of weight that male college students gain during their freshman year are normally distributed with a mean of 1.2 kg and a standard deviation of 4.9 kg (based on Data Set 13 "Freshman 15" in Appendix B).***

5. a. If 1 male college student is randomly selected, find the probability that he has no weight gain during his freshman year. (That is, find the probability that during his freshman year, his weight gain is less than or equal to 0 kg.)

b. If 25 male college students are randomly selected, find the probability that their mean weight gain during their freshman year is less than or equal to 0 kg.

c. Why can the normal distribution be used in part (b), even though the sample size does not exceed 30?

6. a. If 1 male college student is randomly selected, find the probability that he gains at least 2.0 kg during his freshman year.

b. If 16 male college students are randomly selected, find the probability that their mean weight gain during their freshman year is at least 2.0 kg.

c. Why can the normal distribution be used in part (b), even though the sample size does not exceed 30?

7. a. If 1 male college student is randomly selected, find the probability that he gains between 0 kg and 3 kg during freshman year.

b. If 9 male college students are randomly selected, find the probability that their mean weight gain during freshman year is between 0 kg and 3 kg.

c. Why can the normal distribution be used in part (b), even though the sample size does not exceed 30?

8. a. If 1 male college student is randomly selected, find the probability that he gains between 0.5 kg and 2.5 kg during freshman year.

b. If 4 male college students are randomly selected, find the probability that their mean weight gain during freshman year is between 0.5 kg and 2.5 kg.

c. Why can the normal distribution be used in part (b), even though the sample size does not exceed 30?

Ergonomics. ***Exercises 9–16 involve applications to* ergonomics, *as described in the Chapter Problem.***

9. Safe Loading of Elevators The elevator in the car rental building at San Francisco International Airport has a placard stating that the maximum capacity is “4000 lb—27 passengers.” Because $4000/27 = 148$, this converts to a mean passenger weight of 148 lb when the elevator is full. We will assume a worst-case scenario in which the elevator is filled with 27 adult males. Based on Data Set 1 “Body Data” in Appendix B, assume that adult males have weights that are normally distributed with a mean of 189 lb and a standard deviation of 39 lb.

a. Find the probability that 1 randomly selected adult male has a weight greater than 148 lb.

b. Find the probability that a sample of 27 randomly selected adult males has a mean weight greater than 148 lb.

c. What do you conclude about the safety of this elevator?

10. Designing Manholes According to the website www.torchmate.com, “manhole covers must be a minimum of 22 in. in diameter, but can be as much as 60 in. in diameter.” Assume that a manhole is constructed to have a circular opening with a diameter of 22 in. Men have shoulder widths that are normally distributed with a mean of 18.2 in. and a standard deviation of 1.0 in. (based on data from the National Health and Nutrition Examination Survey).

a. What percentage of men will fit into the manhole?

b. Assume that the Connecticut’s Evergreen company employs 36 men who work in manholes. If 36 men are randomly selected, what is the probability that their mean shoulder width is less than 18.5 in.? Does this result suggest that money can be saved by making smaller manholes with a diameter of 18.5 in.? Why or why not?

11. Water Taxi Safety Passengers died when a water taxi sank in Baltimore’s Inner Harbor. Men are typically heavier than women and children, so when loading a water taxi, assume a worst-case scenario in which all passengers are men. Assume that weights of men are normally distributed with a mean of 189 lb and a standard deviation of 39 lb (based on Data Set 1 “Body Data” in Appendix B). The water taxi that sank had a stated capacity of 25 passengers, and the boat was rated for a load limit of 3500 lb.

a. Given that the water taxi that sank was rated for a load limit of 3500 lb, what is the maximum mean weight of the passengers if the boat is filled to the stated capacity of 25 passengers?

b. If the water taxi is filled with 25 randomly selected men, what is the probability that their mean weight exceeds the value from part (a)?

c. After the water taxi sank, the weight assumptions were revised so that the new capacity became 20 passengers. If the water taxi is filled with 20 randomly selected men, what is the probability that their mean weight exceeds 175 lb, which is the maximum mean weight that does not cause the total load to exceed 3500 lb?

d. Is the new capacity of 20 passengers safe?

12. Loading a Tour Boat The Ethan Allen tour boat capsized and sank in Lake George, New York, and 20 of the 47 passengers drowned. Based on a 1960 assumption of a mean weight of 140 lb for passengers, the boat was rated to carry 50 passengers. After the boat sank, New York State changed the assumed mean weight from 140 lb to 174 lb.

a. Given that the boat was rated for 50 passengers with an assumed mean of 140 lb, the boat had a passenger load limit of 7000 lb. Assume that the boat is loaded with 50 male passengers, and assume that weights of men are normally distributed with a mean of 189 lb and a standard deviation of 39 lb (based on Data Set 1 "Body Data" in Appendix B). Find the probability that the boat is overloaded because the 50 male passengers have a mean weight greater than 140 lb.

b. The boat was later rated to carry only 14 passengers, and the load limit was changed to 2436 lb. If 14 passengers are all males, find the probability that the boat is overloaded because their mean weight is greater than 174 lb (so that their total weight is greater than the maximum capacity of 2436 lb). Do the new ratings appear to be safe when the boat is loaded with 14 male passengers?

13. Redesign of Ejection Seats When women were finally allowed to become pilots of fighter jets, engineers needed to redesign the ejection seats because they had been originally designed for men only. The ACES-II ejection seats were designed for men weighing between 140 lb and 211 lb. Weights of women are now normally distributed with a mean of 171 lb and a standard deviation of 46 lb (based on Data Set 1 "Body Data" in Appendix B).

a. If 1 woman is randomly selected, find the probability that her weight is between 140 lb and 211 lb.

b. If 25 different women are randomly selected, find the probability that their mean weight is between 140 lb and 211 lb.

c. When redesigning the fighter jet ejection seats to better accommodate women, which probability is more relevant: the result from part (a) or the result from part (b)? Why?

14. Loading Aircraft Before every flight, the pilot must verify that the total weight of the load is less than the maximum allowable load for the aircraft. The Bombardier Dash 8 aircraft can carry 37 passengers, and a flight has fuel and baggage that allows for a total passenger load of 6200 lb. The pilot sees that the plane is full and all passengers are men. The aircraft will be overloaded if the mean weight of the passengers is greater than $6200 \text{ lb}/37 = 167.6 \text{ lb}$. What is the probability that the aircraft is overloaded? Should the pilot take any action to correct for an overloaded aircraft? Assume that weights of men are normally distributed with a mean of 189 lb and a standard deviation of 39 lb (based on Data Set 1 "Body Data" in Appendix B).

15. Doorway Height The Boeing 757-200 ER airliner carries 200 passengers and has doors with a height of 72 in. Heights of men are normally distributed with a mean of 68.6 in. and a standard deviation of 2.8 in. (based on Data Set 1 "Body Data" in Appendix B).

a. If a male passenger is randomly selected, find the probability that he can fit through the doorway without bending.

b. If half of the 200 passengers are men, find the probability that the mean height of the 100 men is less than 72 in.

c. When considering the comfort and safety of passengers, which result is more relevant: the probability from part (a) or the probability from part (b)? Why?

d. When considering the comfort and safety of passengers, why are women ignored in this case?

16. Aircraft Cockpit The overhead panel in an aircraft cockpit typically includes controls for such features as landing lights, fuel booster pumps, and oxygen. It is important for pilots to be able to reach those overhead controls while sitting. Seated adult males have overhead grip reaches that are normally distributed with a mean of 51.6 in. and a standard deviation of 2.2 in.

a. If an aircraft is designed for pilots with an overhead grip reach of 53 in., what percentage of adult males would not be able to reach the overhead controls? Is that percentage too high?

b. If the cockpit is designed so that 95% of adult males would be able to reach the overhead controls, what is the overhead grip reach distance?

c. A small regional airline employs 40 male pilots. An engineer wants to design for an overhead grip reach that satisfies this criterion: There is a 0.95 probability that 40 randomly selected male pilots have a mean overhead grip reach that is greater than or equal to the designed overhead reach distance. What overhead grip reach distance satisfies that design? Why should this engineer be fired?

Hypothesis Testing. ***In Exercises 17–20, apply the central limit theorem to test the given claim. (Hint: See Example 3.)***

17. Freshman 15 The term "Freshman 15" refers to the claim that college students gain 15 lb during their freshman year at college. Data Set 13 "Freshman 15" includes measurements from 67 college students from their freshman year, and they had weight gains with a mean of 2.6 lb and a standard deviation of 8.6 lb. Assume that the mean weight gain really is 15 lb and find the probability that a random sample of 67 college students would have a mean weight gain of 2.6 lb or less. What does the result suggest about the claim of the "Freshman 15"?

18. Adult Sleep Times (hours) of sleep for randomly selected adult subjects included in the National Health and Nutrition Examination Study are listed below. Here are the statistics for this sample: $n = 12$, $\bar{x} = 6.8$ hours, $s = 2.0$ hours. The times appear to be from a normally distributed population. A common recommendation is that adults should sleep between 7 hours and 9 hours each night. Assuming that the mean sleep time is 7 hours, find the probability of getting a sample of 12 adults with a mean of 6.8 hours or less. What does the result suggest about a claim that "the mean sleep time is less than 7 hours"?

4 8 4 4 8 6 9 7 7 10 7 8

19. Weight Watchers Diet When 40 people used the Weight Watchers diet for one year, their mean weight loss was 3.0 lb, and the standard deviation was 4.9 lb (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Reduction," by Dansinger et al., *Journal of the American Medical Association*, Vol. 293, No. 1). Assuming that the diet has no effect so the true mean amount of lost weight is 0 lb, find the probability of getting a sample of 40 subjects with a mean weight loss of 3.0 lb or higher. Based on the result, is the mean weight loss of 3.0 lb *significantly high*? What do the results suggest about the effectiveness of the diet?

6-4 Beyond the Basics

20. Correcting for a Finite Population In a study of babies born with very low birth weights, 275 children were given IQ tests at age 8, and their scores approximated a normal distribution with $\mu = 95.5$ and $\sigma = 16.0$ (based on data from "Neurobehavioral Outcomes of School-age Children Born Extremely Low Birth Weight or Very Preterm," by Anderson et al., *Journal of the American Medical Association*, Vol. 289, No. 24). Fifty of those children are to be randomly selected without replacement for a follow-up study.

a. When considering the distribution of the mean IQ scores for samples of 50 children randomly selected from a population of 275 children, should $\sigma_{\bar{x}}$ be corrected by using the finite population correction factor? Why or why not? What is the value of $\sigma_{\bar{x}}$?

b. Find the probability that the mean IQ score of the follow-up sample is between 95 and 105.

6-5 Assessing Normality

Key Concept The following chapters include important statistical methods requiring that sample data are from a population having a distribution that is approximately *normal*. In this section we present these steps for determining whether sample data satisfy the requirement of a normal distribution:

1. Construct a histogram and determine whether it is roughly bell-shaped.
2. Construct a *normal quantile plot* and use the criteria given later in this section.

PART 1 Basic Concepts of Assessing Normality

When trying to determine whether a collection of data has a distribution that is approximately normal, we can visually inspect a histogram to see if it is approximately bell-shaped (as discussed in Section 2-2), and we can also use a *normal quantile plot* (discussed briefly in Section 2-2).

DEFINITION

A **normal quantile plot** (or **normal probability plot**) is a graph of points (x, y) where each x value is from the original set of sample data, and each y value is the corresponding z score that is expected from the standard normal distribution.

Procedure for Determining Whether It Is Reasonable to Assume That Sample Data Are from a Population Having a Normal Distribution

1. *Histogram:* Construct a histogram. If the histogram departs dramatically from a bell shape, conclude that the data do not have a normal distribution.
2. *Normal quantile plot:* If the histogram is basically symmetric, use technology to generate a *normal quantile plot*. Apply the following criteria to determine whether the distribution is normal. (These criteria can be used loosely for small samples, but they should be used more strictly for large samples.)

   Normal Distribution: The population distribution is normal if the pattern of the points is reasonably close to a straight line and the points do not show some systematic pattern that is not a straight-line pattern.

   Not a Normal Distribution: The population distribution is *not* normal if either or both of these two conditions applies:

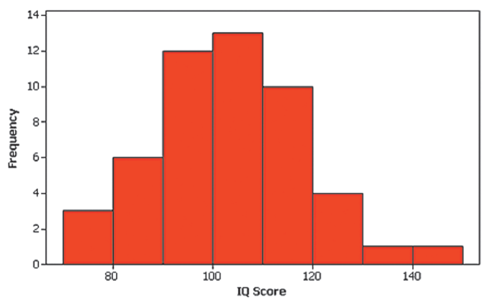
   - The points do not lie reasonably close to a straight-line pattern.
   - The points show some *systematic pattern* that is not a straight-line pattern.

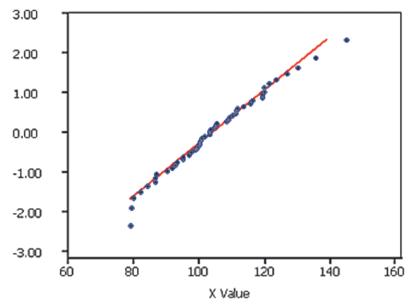
Advanced Methods: In addition to using histograms and normal quantile plots, there are other more advanced procedures for assessing normality, such as the Ryan-Joiner test (discussed briefly in Part 2 of this section). Other tests for normality include (insert drum roll here) the Shapiro-Wilk test, D'Agostino-Pearson test, chi-square goodness-of-fit test, Kolmogorov-Smirnov test, Lillefors corrected K-S test, Cramer-von Mises test, Anderson-Darling test, the Jarque-Bera test, and the Anscombe-Glynn kurtosis test.

Histograms and Normal Quantile Plots

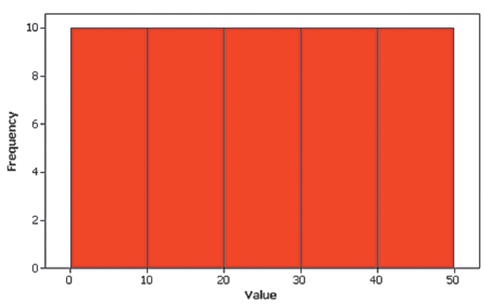
In Part 2 of this section we describe the process of constructing a normal quantile plot, but for now we focus on *interpreting* a normal quantile plot. The following displays show histograms of data and the corresponding normal quantile plots.

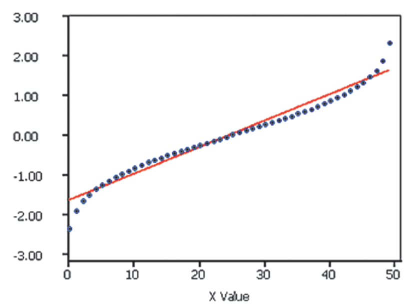
Normal: The first case shows a histogram of IQ scores that is close to being bell-shaped, so the histogram suggests that the IQ scores are from a normal distribution. The corresponding normal quantile plot shows points that are reasonably close to a straight-line pattern, and the points do not show any other systematic pattern that is not a straight line. It is safe to assume that these IQ scores are from a population that has a normal distribution.



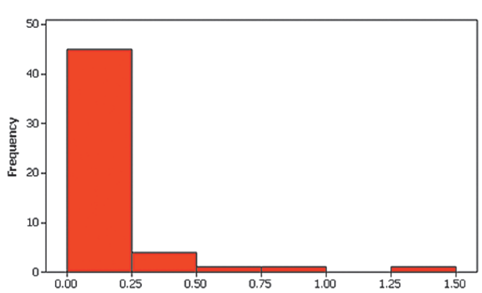


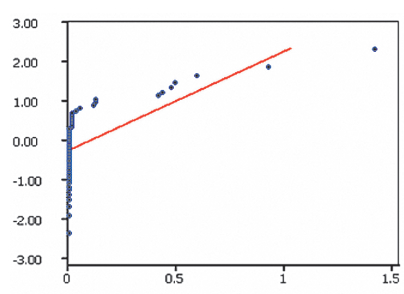
Uniform: The second case shows a histogram of data having a uniform (rectangular) distribution. The corresponding normal quantile plot suggests that the points are *not* normally distributed. Although the pattern of points is reasonably close to a straight-line pattern, *there is another systematic pattern that is not a straight-line pattern*. We conclude that these sample values are from a population having a distribution that is not normal.





Skewed: The third case shows a histogram of the amounts of rainfall (in inches) in Boston for every Monday in one year. The shape of the histogram is skewed to the right, not bell-shaped. The corresponding normal quantile plot shows points that are not at all close to a straight-line pattern. These rainfall amounts are from a population having a distribution that is not normal.





PART 2 Manual Construction of Normal Quantile Plots

The following is a relatively simple procedure for manually constructing a normal quantile plot, and it is the same procedure used by Statdisk and the TI-83/84 Plus calculator. Some statistical packages use various other approaches, but the interpretation of the graph is essentially the same.

Manual Construction of a Normal Quantile Plot

Step 1: First sort the data by arranging the values in order from lowest to highest.

Step 2: With a sample of size n, each value represents a proportion of $1/n$ of the sample. Using the known sample size n, find the values of $\frac{1}{2n}, \frac{3}{2n}, \frac{5}{2n}$, and so on, until you get n values. These values are the cumulative areas to the left of the corresponding sample values.

Step 3: Use the standard normal distribution (software or a calculator or Table A-2) to find the z scores corresponding to the cumulative left areas found in Step 2. (These are the z scores that are expected from a normally distributed sample.)

Step 4: Match the original sorted data values with their corresponding z scores found in Step 3, then plot the points (x, y), where each x is an original sample value and y is the corresponding z score.

Step 5: Examine the normal quantile plot and use the criteria given in Part 1. Conclude that the population has a normal distribution if the pattern of the points is reasonably close to a straight line and the points do not show some systematic pattern that is not a straight-line pattern.

EXAMPLE 1 Dallas Commute Times

Data set 31 "Commute Times" in Appendix B includes commute times (minutes) obtained from Dallas, Texas. Let's consider this sample of the first five commute times: 20, 16, 25, 10, 30. With only five sample values, a histogram will not be very helpful here. Instead, construct a normal quantile plot for these five values and determine whether they appear to be from a population that is normally distributed.

SOLUTION

The following steps correspond to those listed in the procedure above for constructing a normal quantile plot.

Step 1: First, sort the data by arranging them in order. We get 10, 16, 20, 25, 30.

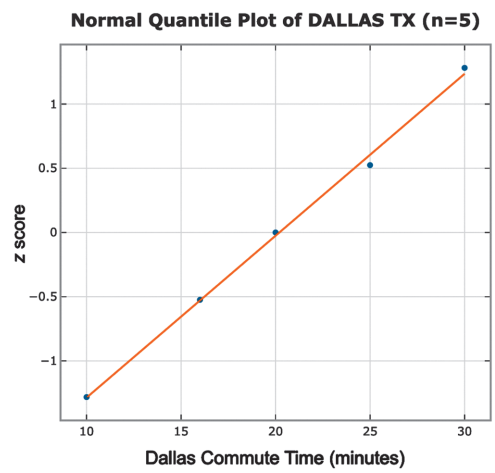
Step 2: With a sample of size $n = 5$, each value represents a proportion of $1/5$ of the sample, so we proceed to identify the cumulative areas to the left of the corresponding sample values. The cumulative left areas, which are expressed in general as $\frac{1}{2n}, \frac{3}{2n}, \frac{5}{2n}$, and so on, become these specific areas for this example with $n = 5$: $\frac{1}{10}, \frac{3}{10}, \frac{5}{10}, \frac{7}{10}, \frac{9}{10}$. These cumulative left areas expressed in decimal form are 0.1, 0.3, 0.5, 0.7, and 0.9.

Step 3: We now use technology (or Table A-2) with the cumulative left areas of 0.1000, 0.3000, 0.5000, 0.7000, and 0.9000 to find these corresponding z scores: -1.28, -0.52, 0, 0.52, and 1.28. (For example, the z score of -1.28 has an area of 0.1000 to its left.)

Step 4: We now pair the original sorted Dallas commute times with their corresponding z scores. We get these (x, y) coordinates, which are plotted in the accompanying Statdisk display:

$$(10, -1.28), (16, -0.52), (20, 0), (25, 0.52), (30, 1.28)$$

Statdisk



INTERPRETATION

We examine the normal quantile plot in the Statdisk display. The points do appear to lie reasonably close to the straight line, so we conclude that the sample of five commute times *does* appear to be from a normally distributed population.

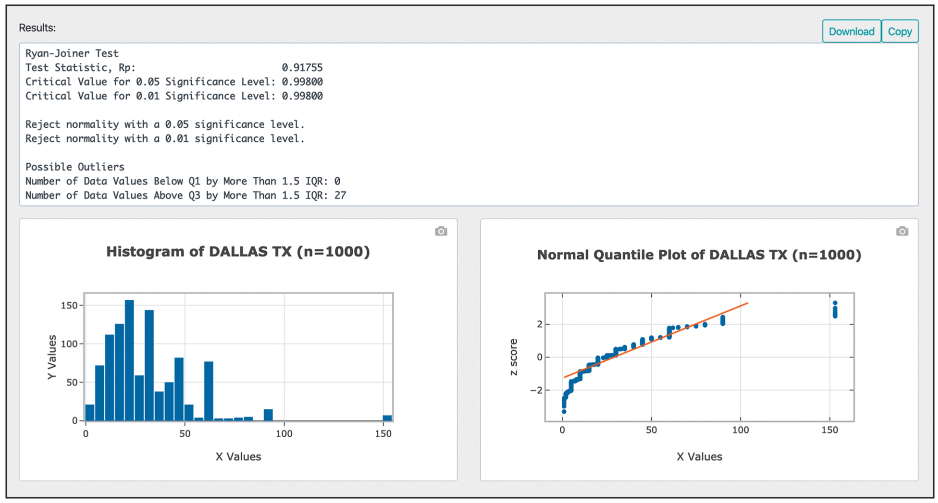
YOUR TURN. Do Exercise 5 "Ages of Presidents."

Ryan-Joiner Test The Ryan-Joiner test is one of several formal tests of normality, each having its own advantages and disadvantages. Statdisk has a feature of **Normality Assessment** that displays a histogram, normal quantile plot, the number of potential outliers, and results from the Ryan-Joiner test.

EXAMPLE 2 Dallas Commute Times

Example 1 used only five of the Dallas commute times listed in Data Set 31 “Commute Times” in Appendix B. Shown in the accompanying display is the result obtained by using the Statdisk **Normality Assessment** feature with all 1000 Dallas commute times.

Statdisk



Let’s use the display with the three criteria for assessing normality.

1. *Histogram:* We can see that the histogram is *skewed* to the right and is far from being bell-shaped.
2. *Normal quantile plot:* The points in the normal quantile plot are very far from a straight-line pattern. We conclude that the 1000 Dallas commute times do *not* appear to be from a population with a normal distribution.

YOUR TURN. Do Exercise 11 “Small World.”

Outliers We should always be aware of the presence of any outliers, particularly because they can have very dramatic effects on results. Test for the effects of outliers by applying statistical methods with these outliers included and then a second time with outliers excluded. Outliers should be investigated because they may be the most important characteristics of the data and they may reveal critical information about the data. Discard outliers only if they are identified as being errors.

Data Transformations Many data sets have a distribution that is not normal, but we can *transform* the data so that the modified values have a normal distribution. One common transformation is to transform each value of x by taking its logarithm. (You can use natural logarithms or logarithms with base 10. If any original values are 0, take logarithms of values of $x + 1$). If the distribution of the logarithms of the values is a normal distribution, the distribution of the original values is called a **lognormal distribution**. (See Exercises 21 “Transformations” and 22 “Lognormal Distribution.”) In addition to transformations with logarithms, there are other transformations, such as replacing each x value with $\sqrt{x}$, or $1/x$, or x^2. In addition to getting a required normal distribution when the original data values are not normally distributed, such transformations can be used to correct deficiencies, such as a requirement (found in later chapters) that different data sets have the same variance.

TECH CENTER

Normal Quantile Plots

Access tech supplements, videos, and data sets at **www.TriolaStats.com**

Statdisk

1. Click **Data** in the top menu.
2. Select **Normal Quantile Plot** from the dropdown menu.
3. Select the desired data column and click **Plot.**

TIP: Select **Normality Assessment** in the dropdown menu under **Data** to obtain the normal quantile plot along with other results helpful in assessing normality.

Minitab

Minitab generates a probability plot that is similar to the normal quantile plot and can be interpreted using the same criteria given in this section.

Probability Plot

1. Click **Stat** in the top menu.
2. Select **Basic Statistics** from the dropdown menu and select **Normality Test** from the submenu.
3. Select the desired column in the *Variable* box and click **OK.**

Probability Plot with Boundaries

1. Click **Graph** in the top menu.
2. Select **Probability Plot** from the dropdown menu, select **Single**,and click **OK.**
3. Select the desired column in the *Graph variables* box and click **OK.**
4. If the points all lie within the boundaries, conclude that the data are normally distributed. If points are outside the boundaries, conclude that the data are not normally distributed.

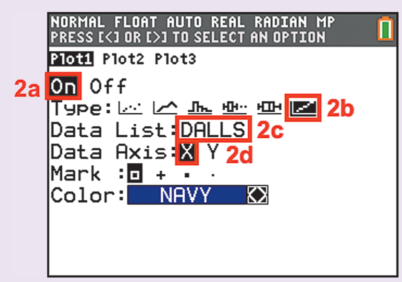
StatCrunch

StatCrunch generates a QQ Plot that is similar to the normal quantile plot and can be interpreted using the same criteria given in this section.

1. Click **Graph** in the top menu.
2. Select **QQ Plot** in the dropdown menu.
3. Select the desired data column.
4. Click **Compute!**

TI-83/84 Plus Calculator

1. Open the **STAT PLOTS** menu by pressing 2ND, Y=.
2. Press ENTER to access the Plot 1 settings screen as shown:
   a. Select **ON** and press ENTER.
   b. Select last graph type, press ENTER.
   c. Enter name of list containing data.
   d. For *Data Axis* select **X.**
3. Press ZOOM then 9 (ZoomStat) to generate the normal quantile plot.
4. Press WINDOW to customize graph and then press GRAPH to view the normal quantile plot.



Excel

XLSTAT Add-In (Required)

1. Click on the **XLSTAT** tab in the Ribbon and then click **Describing Data.**
2. Select **Normality tests** from the dropdown menu.
3. Enter the desired data range in the *Data* box. If the first row of data contains a label, check the **Sample labels** box.
4. Click the **Charts** tab and confirm that the **Normal Q-Q plots** box is checked.
5. Click **OK** and scroll down the results to view the Normal Q-Q plot.

R

R command: **qqnorm(*x*)**

TIP: The R command **qqline(*x*)** can be used to add a reference line.

A complete list of R statistical commands is available at TriolaStats.com

6-5 Basic Skills and Concepts

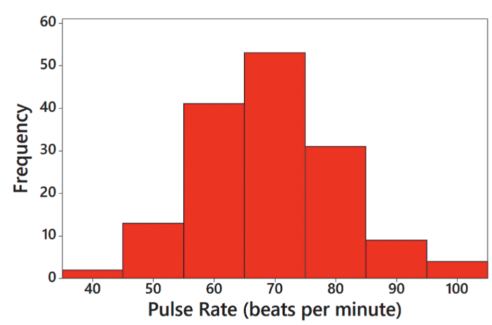
Statistical Literacy and Critical Thinking

1. Satisfying Requirements Data Set 1 "Body Data" in Appendix B includes a sample of 147 pulse rates of randomly selected women. Does that sample satisfy the following requirement: (1) The sample appears to be from a normally distributed population; or (2) the sample has a size of $n > 30$?

2. Histogram and Normal Quantile Plot Pulse rates of women are normally distributed. If we construct a histogram and normal quantile plot using the same sample of 147 pulse rates described in the preceding exercise, describe the appearance of those two graphs.

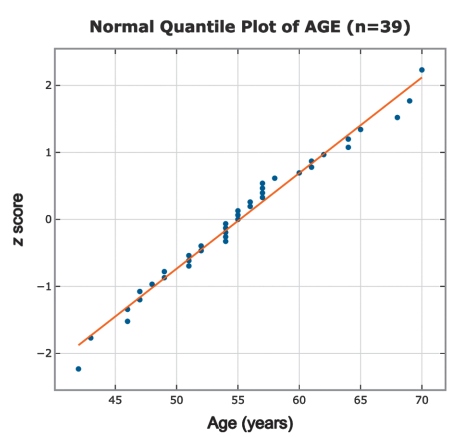
3. Normal Quantile Plot After constructing a histogram of the ages of the 147 women included in Data Set 1 "Body Data" in Appendix B, you see that the histogram is far from being bell-shaped. What do you now know about the pattern of points in the normal quantile plot?

4. Assessing Normality The accompanying histogram is constructed from the pulse rates of the 153 men included in Data Set 1 "Body Data" in Appendix B. If you plan to conduct further statistical tests and there is a requirement of a normally distributed population, what do you conclude about the population distribution based on this histogram?

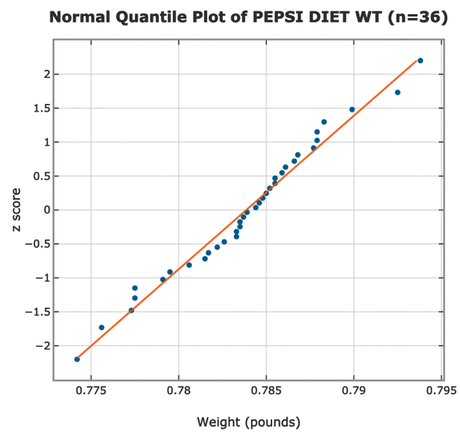


Interpreting Normal Quantile Plots. ***In Exercises 5–8, examine the normal quantile plot and determine whether the sample data appear to be from a population with a normal distribution.***

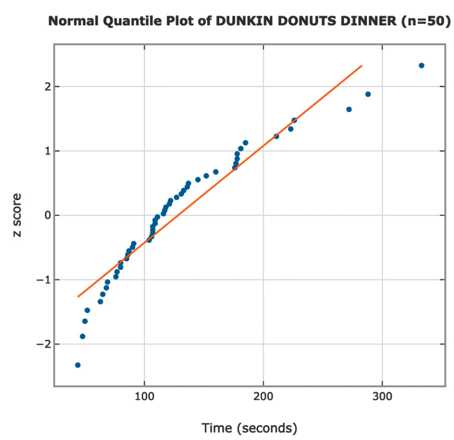
5. Ages of Presidents The normal quantile plot represents the ages of presidents of the United States at the times of their inaugurations. The data are from Data Set 22 "Presidents" in Appendix B.



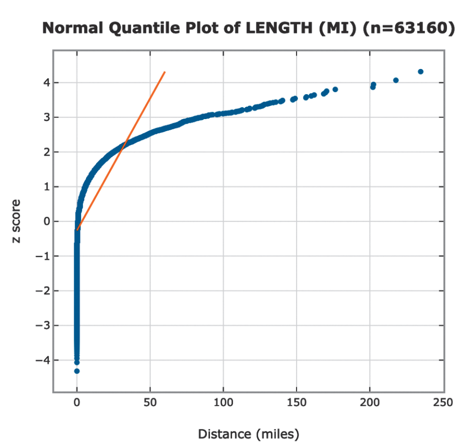
6. Diet Pepsi The normal quantile plot represents weights (pounds) of the contents of cans of Diet Pepsi (from Data Set 37 "Cola Weights and Volumes" in Appendix B).



7. Dunkin' Donuts Service Times The normal quantile plot represents service times during the dinner hours at Dunkin' Donuts (from Data Set 36 "Fast Food" in Appendix B).



8. Tornadoes The normal quantile plot represents the distances (miles) that tornadoes traveled (from Data Set 25 "Tornadoes" in Appendix B).



Determining Normality. ***In Exercises 9–12, refer to the indicated sample data and determine whether they appear to be from a population with a normal distribution. Assume that this requirement is loose in the sense that the population distribution need not be exactly normal, but it must be a distribution that is roughly bell-shaped.***

9. M&Ms The weights (grams) of the red M&M plain candies, as listed in Data Set 38 "Candies" in Appendix B

10. Taxi Trips The distances (miles) traveled by New York City taxis transporting customers, as listed in Data Set 32 "Taxis" in Appendix B

11. Small World The waiting times (minutes) for the Disney World ride Small World at 5:00 PM, as listed in Data Set 33 "Disney World Wait Times" in Appendix B

12. Dunkin' Donuts The drive-through service times (seconds) of Dunkin' Donuts lunch customers, as listed in Data Set 36 "Fast Food" in Appendix B

Using Technology to Generate Normal Quantile Plots. ***In Exercises 13–16, use the data from the indicated exercise in this section. Use software (such as Statdisk, Minitab, Excel, or StatCrunch) or a TI-83/84 Plus calculator to generate a normal quantile plot. Then determine whether the data come from a normally distributed population.***

13. Exercise 9 "M&Ms"

14. Exercise 10 "Taxi Trips"

15. Exercise 11 "Small World"

16. Exercise 12 "Dunkin' Donuts"

Constructing Normal Quantile Plots. ***In Exercises 17–20, use the given data values to identify the corresponding z scores that are used for a normal quantile plot, then identify the coordinates of each point in the normal quantile plot. Construct the normal quantile plot, then determine whether the data appear to be from a population with a normal distribution.***

17. Body Temperatures A sample of body temperatures (°F) of women from Data Set 5 "Body Temperatures" in Appendix B: 98.7, 98.4, 98.0, 97.9, 98.2

18. Earthquake Depths A sample of depths (km) of earthquakes is obtained from Data Set 24 "Earthquakes" in Appendix B: 17.3, 7.0, 7.0, 7.0, 8.1, 6.8.

19. Brain Volumes A sample of human brain volumes (cm^3) is obtained from those listed in Data Set 12 "IQ and Brain Size" in Appendix B: 1027, 1029, 1034, 1070, 1079, 1079, 963, 1439.

20. Ages of Oscar-Winning Actresses A sample of the ages (years) of actresses who won Oscars, as listed in Data Set 21 "Oscar Winner Age" in Appendix B: 25, 24, 41, 30, 27, 35, 33, 29, 80

6-5 Beyond the Basics

21. Transformations The heights (in inches) of women listed in Data Set 1 "Body Data" in Appendix B have a distribution that is approximately normal, so it appears that those heights are from a normally distributed population.

a. If 2 inches is added to each height, are the new heights also normally distributed?

b. If each height is converted from inches to centimeters, are the heights in centimeters also normally distributed?

c. Are the logarithms of the normally distributed heights also normally distributed?

22. Lognormal Distribution The following are the values of net worth (in millions of dollars) of recent members of the executive branch of the U.S. government. Test these values for normality, then take the logarithm of each value and test for normality. What do you conclude?

1297 15.8 7.08 1.05 0.73 0.08 0.23 0.19

6-6 Normal as Approximation to Binomial (available at www.TriolaStats.com)

Because of our ability to use technology to find exact values of binomial probabilities, the use of a normal approximation to a binomial distribution has become largely obsolete, so this section is included on the website www.TriolaStats.com.

Here are two key points:

- Given probabilities p and q (where $q = 1 - p$) and sample size n, if the conditions $np \geq 5$ and $nq \geq 5$ are both satisfied, then probabilities from a binomial probability distribution can be approximated reasonably well by using a normal distribution having these parameters:

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

- The binomial probability distribution is *discrete* (with whole numbers for the random variable x), but the normal approximation is *continuous*. To compensate, we use a "continuity correction" with each whole number x represented by the interval from $x - 0.5$ to $x + 0.5$.

Chapter Quick Quiz

Bone Density Test. ***In Exercises 1–4, assume that scores on a bone mineral density test are normally distributed with a mean of 0 and a standard deviation of 1.***

1. Bone Density Sketch a graph showing the shape of the distribution of bone density test scores.

2. Bone Density Find the bone density score that is the 90th percentile, which is the score separating the lowest 90% from the top 10%.

3. Bone Density For a randomly selected subject, find the probability of a bone density score greater than 1.55.

4. Bone Density For a randomly selected subject, find the probability of a bone density score between −1.00 and 2.00.

5. Notation

a. Identify the values of μ and σ for the standard normal distribution.

b. What do the symbols $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ represent?

6. Salaries It is known that salaries of college professors have a distribution that is skewed. If we repeat the process of randomly selecting 50 college professors and find the mean of each sample, what is the distribution of these sample means?

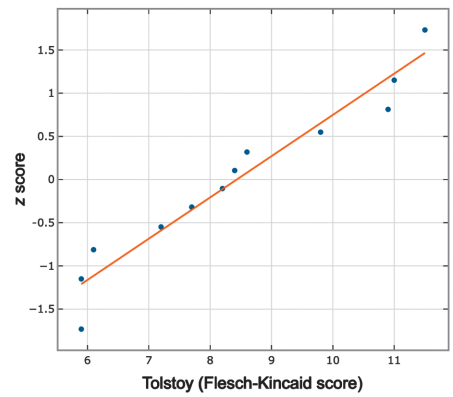
Seat Designs. ***In Exercises 7–9, assume that when seated, adult males have back-to-knee lengths that are normally distributed with a mean of 23.5 in. and a standard deviation of 1.1 in. (based on anthropometric survey data from Gordon, Churchill, et al.). These data are used often in the design of different seats, including aircraft seats, train seats, theater seats, and classroom seats.***

7. Find the probability that a male has a back-to-knee length greater than 25.0 in.

8. Find the probability that a male has a back-to-knee length between 22.0 in. and 26.0 in.

9. Find the probability that nine males have back-to-knee lengths with a mean greater than 23.0 in.

10. *War and Peace* The accompanying normal quantile plot was constructed from the Flesch-Kincaid Grade Level scores for pages randomly selected from Leo Tolstoy's classic novel *War and Peace*. (Flesch-Kincaid Grade Level scores are measures of the U.S. grade level of education that the reader should have to understand the text.) What does the graph tell us about those scores?



Review Exercises

1. Bone Density Test A bone mineral density test is used to identify a bone disease. The result of a bone density test is commonly measured as a z score, and the population of z scores is normally distributed with a mean of 0 and a standard deviation of 1.

a. For a randomly selected subject, find the probability of a bone density test score greater than −1.37.

b. For a randomly selected subject, find the probability of a bone density test score less than 2.34.

c. For a randomly selected subject, find the probability of a bone density test score between −0.67 and 1.29.

d. Find Q_1, the bone density test score separating the bottom 25% from the top 75%.

e. If the mean bone density test score is found for 9 randomly selected subjects, find the probability that the mean is greater than 0.23.

2. Unbiased Estimators

a. What is an unbiased estimator?

b. For the following statistics, identify those that are unbiased estimators: mean, median, range, variance, proportion.

c. Determine whether the following statement is true or false: "The sample standard deviation is a biased estimator, but the bias is relatively small in large samples, so s is often used to estimate σ."

3. Critical Values

a. Find the standard z score with a cumulative area to its left of 0.01.

b. Find the standard z score with a cumulative area to its right of 0.01.

c. Find the value of $z_{0.025}$.

4. Arm Circumferences Arm circumferences of adult men are normally distributed with a mean of 33.64 cm and a standard deviation of 4.14 cm (based on Data Set 1 "Body Data" in Appendix B). A sample of 25 men is randomly selected and the mean of the arm circumferences is obtained.

a. Describe the distribution of such sample means.

b. What is the mean of all such sample means?

c. What is the standard deviation of all such sample means?

5. Birth Weights Based on Data Set 6 "Births" in Appendix B, birth weights of girls are normally distributed with a mean of 3037.1 g and a standard deviation of 706.3 g.

a. For the bell-shaped graph, what is the area under the curve?

b. What is the value of the median?

c. What is the value of the mode?

d. What is the value of the variance?

6. Mensa Membership in Mensa requires a score in the top 2% on a standard intelligence test. The Wechsler IQ test is designed for a mean of 100 and a standard deviation of 15, and scores are normally distributed.

a. Find the minimum Wechsler IQ test score that satisfies the Mensa requirement.

b. If 4 randomly selected adults take the Wechsler IQ test, find the probability that their mean score is at least 131.

c. If 4 subjects take the Wechsler IQ test and they have a mean of 131 but the individual scores are lost, can we conclude that all 4 of them have scores of at least 131?

7. Tall Clubs The social organization Tall Clubs International has a requirement that women must be at least 70 in. tall. Assume that women have normally distributed heights with a mean of 63.7 in. and a standard deviation of 2.9 in. (based on Data Set 1 "Body Data" in Appendix B).

a. Find the percentage of women who satisfy the height requirement.

b. If the height requirement is to be changed so that the tallest 2.5% of women are eligible, what is the new height requirement?

In Exercises 8 and 9, assume that women have standing eye heights that are normally distributed with a mean of 59.7 in. and a standard deviation of 2.5 in. (based on anthropometric survey data from Gordon, Churchill, et al.).

8. Biometric Security In designing a security system based on eye (iris) recognition, we must consider the standing eye heights of women.

a. If an eye recognition security system is positioned at a height that is uncomfortable for women with standing eye heights less than 54 in., what percentage of women will find that height uncomfortable?

b. In positioning the eye recognition security system, we want it to be suitable for the lowest 95% of standing eye heights of women. What standing eye height of women separates the lowest 95% of standing eye heights from the highest 5%?

9. Significance Instead of using 0.05 for identifying significant values, use the criteria that a value x is *significantly high* if $P(x \text{ or greater}) \leq 0.01$ and a value is *significantly low* if $P(x \text{ or less}) \leq 0.01$. Find the standing eye heights of women that separate significant values from those that are not significant. Using these criteria, is a woman's standing eye height of 67 in. significantly high?

10. Assessing Normality Listed below are the recent salaries (in millions of dollars) of players on the LA Lakers professional basketball team. Do these salaries appear to come from a population that has a normal distribution? Why or why not?

35.6 14.4 12.0 9.0 7.5 5.8 4.4 3.5 2.4 1.8 1.7 1.7 1.5 1.5 1.0 0.1 0.1

Cumulative Review Exercises

In Exercises 1 and 2, use the following wait times (minutes) at 10:00 AM for the Tower of Terror ride at Disney World (from Data Set 33 "Disney World Wait Times" in Appendix B).

35 35 20 50 95 75 45 50 30 35 30 30

1. Tower of Terror Wait Times

a. Find the mean $\bar{x}$.

b. Find the median.

c. Find the standard deviation s.

d. Find the variance.

e. Convert the longest wait time to a z score.

f. Based on the result from part (e), is the longest wait time significantly high?

g. What level of measurement (nominal, ordinal, interval, ratio) describes this data set?

h. Are the wait times discrete data or continuous data?

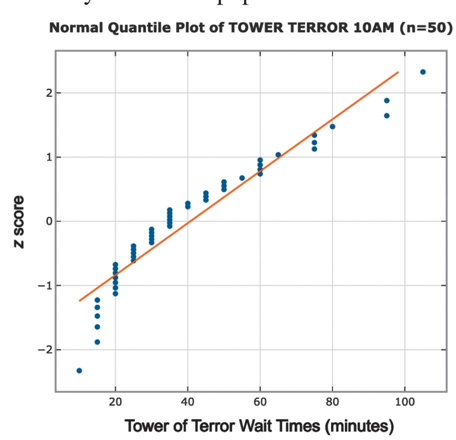
2. Tower of Terror Wait Times

a. Find Q_1, Q_2, and Q_3.

b. Construct a boxplot.

c. Based on the boxplot from part (b), do the sample data appear to be from a normally distributed population?

d. The accompanying normal quantile plot is obtained by using all 50 wait times at 10:00 AM for the Tower of Terror ride at Disney World. Based on this normal quantile plot, do the sample data appear to be from a normally distributed population?



3. Foot Lengths of Women Assume that foot lengths of adult females are normally distributed with a mean of 246.3 mm and a standard deviation of 12.4 mm (based on Data Set 3 "ANSUR II 2012" in Appendix B).

a. Find the probability that a randomly selected adult female has a foot length less than 221.5 mm.

b. Find the probability that a randomly selected adult female has a foot length between 220 mm and 250 mm.

c. Find P_{95}.

d. Find the probability that 16 adult females have foot lengths with a mean greater than 250 mm.

e. Which is more helpful in planning for production of shoes for adult females: The result from part (c) or the result from part (d)? Why?

4. Blue Eyes Assume that 35% of us have blue eyes (based on a study by Dr. P. Soria at Indiana University).

a. Let B denote the event of selecting someone who has blue eyes. What does the event $\overline{B}$ denote?

b. Find the value of $P(\overline{B})$.

c. Find the probability of randomly selecting three different people and finding that all of them have blue eyes.

d. Find the probability that among 100 randomly selected people, at least 40 have blue eyes.

e. If 35% of us really do have blue eyes, is a result of 40 people with blue eyes among 100 randomly selected people a result that is significantly high?

5. Body Temperatures Listed below are body temperatures (°F) of adult males (based on Data Set 5 "Body Temperatures" in Appendix B).

97.6 98.2 99.6 98.7 99.4 98.2 980 98.6 98.6

a. Find the mean. Does the result seem reasonable?

b. Identify a characteristic of the data that is very notable.

c. Based on the result from part (b), what appears to be the mean body temperature of adult males?

Technology Project

Testing a Claim Using a Simulation Based on the Normal Distribution When 40 people used the Weight Watchers diet for one year, their mean weight loss was 3.0 lb and the standard deviation was 4.9 lb (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Reduction," by Dansinger et al., *Journal of the American Medical Association*, Vol. 293, No. 1).

Make the following assumptions:

- The amounts of weight lost on the Weight Watchers diet are normally distributed.
- The diet has no effect, so the mean amount of weight lost is 0 lb.
- The standard deviation of amounts of lost weight is 4.9 lb.
- Samples of size $n = 40$ are randomly selected.

Use a technology capable of randomly generating values from a normal distribution with a desired sample size, mean, and standard deviation. Use the technology to randomly generate different samples with the preceding four assumptions. Find the mean of each generated sample.

a. Does it appear that the actual sample mean weight loss of 3.0 lb is significantly high? Explain.

b. Based on the values of the simulated sample means, what do you conclude about the effectiveness of the diet? Explain.

Big (or Very Large) Data Project

Refer to Data Set 45 "Births in New York" in Appendix B. Using the 465,506 birth weights (grams), test for normality. Do the histogram of the birth weights and the normal quantile plot of the birth weights suggest the same conclusion about normality?

FROM DATA TO DECISION

Critical Thinking: Designing a campus dormitory elevator

An Ohio college student died when he tried to escape from a dormitory elevator that was overloaded with 24 passengers. The elevator was rated for a maximum weight of 2500 pounds. Let's consider this elevator with an allowable weight of 2500 pounds. Let's also consider parameters for weights of adults, as shown in the accompanying table (based on Data Set 1 "Body Data" in Appendix B).

Weights of Adults

	Males	Females
μ	189 lb	171 lb
σ	39 lb	46 lb
Distribution	Normal	Normal

We could consider design features such as the type of music that could be played on the elevator. We could select songs such as "Imagine," or "Daydream Believer." Instead, we will focus on the critical design feature of weight.

a. First, elevators commonly have a 25% margin of error, so they can safely carry a load that is 25% greater than the stated load. What amount is 25% greater than 2500 pounds? Let's refer to this amount as "the maximum safe load" while the 2500 pound limit is the "placard maximum load."

b. Now we need to determine the maximum number of passengers that should be allowed. Should we base our calculations on the maximum safe load or the 2500 pound placard maximum load?

c. The weights given in the accompanying table are weights of adults not including clothing or textbooks. Add another 10 pounds for each student's clothing and textbooks. What is the maximum number of elevator passengers that should be allowed?

d. Do you think that weights of college students are different from weights of adults from the general population? If so, how? How would that affect the elevator design?

Cooperative Group Activities

1. In-class activity Each student states the last four digits of their Social Security number. For privacy concerns, those four digits can be given in any order. Construct a dotplot of the digits. What is the distribution of those digits? What value is the approximate center of the distribution? Then, each student calculates the mean of their four digits. Construct another dotplot for these means. What is the distribution of the means? What value is the approximate center of the distribution? Compare the variation in the original dotplot to the variation in the dotplot representing the sample means. What conclusions follow?

2. Out-of-class activity Use the Internet to find "Pick 4" lottery results for 50 different drawings. Find the 50 different means. Graph a histogram of the original 200 digits that were selected, and graph a histogram of the 50 sample means. What important principle do you observe?

3. In-class activity Divide into groups of three or four students and address these issues affecting the design of manhole covers.

• Which of the following is most relevant for determining whether a manhole cover diameter of 24 in. is large enough: weights of men, weights of women, heights of men, heights of women, hip widths of men, hip widths of women, shoulder widths of men, shoulder widths of women?

• Why are manhole covers usually round? (This was once a popular interview question asked of applicants at IBM, and there are at least three good answers. One good answer is sufficient here.)

4. Out-of-class activity Divide into groups of three or four students. In each group, develop an original procedure to illustrate the central limit theorem. The main objective is to show that when you randomly select samples from a population, the means of those samples tend to be *normally* distributed, regardless of the nature of the population distribution. For this illustration, begin with some population of values that does *not* have a normal distribution.

5. In-class activity Divide into groups of three or four students. Using a coin to simulate births, each individual group member should simulate 25 births and record the number of simulated girls. Combine all results from the group and record $n =$ total number of births and $x =$ number of girls. Given batches of n births, compute the mean and standard deviation for the number of girls. Is the simulated result unusual? Why or why not?

6. In-class activity Divide into groups of three or four students. Select a set of data from one of these data sets in Appendix B: 2, 3, 4, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 23, 26, 27, 28, 29, 30, 34, 35, 39, 40, 41, 42, 43. (These are the data sets that were not used in examples or exercises in Section 6-5). Use the methods of Section 6-5 to construct a histogram and normal quantile plot, then determine whether the data set appears to come from a normally distributed population.

7. Out-of-class activity Divide into groups of three or four students and have each group collect an original data set of values at the interval or ratio level of measurement. Test for normality and provide reasons why the data set does or does not appear to be from a normally distributed population.

8. In-class activity Divide into groups of three or four students. In each group, develop an original procedure to illustrate that the median is a biased estimator.

9. In-class activity Divide into groups of three or four students. In each group, develop an original procedure to illustrate that the range is a biased estimator.