Chapter 4

The Normal Distribution

Objectives

In this chapter we will study the normal distribution, including

- the use of the normal curve in modeling distributions.
- finding probabilities using the normal curve.
- assessing normality of data sets with the use of normal probability plots.

4.1 Introduction

In Chapter 2, we introduced the idea of regarding a set of data as a sample from a population. In Section 3.4 we saw that the population distribution of a quantitative variable Y can be described by its mean μ and its standard deviation σ and also by a density curve, which represents relative frequencies as areas under the curve. In this chapter we study the most important type of density curve: the **normal curve**. The normal curve is a symmetric "bell-shaped" curve whose exact form we will describe next. A distribution represented by a normal curve is called a **normal distribution**.

The family of normal distributions plays two roles in statistical applications. Its more straightforward use is as a convenient approximation to the distribution of an observed variable Y. The second role of the normal distribution is more theoretical and will be explored in Chapter 5.

An example of a natural population distribution that can be approximated by a normal distribution follows.

Example 4.1.1

Serum Cholesterol The relationship between the concentration of cholesterol in the blood and the occurrence of heart disease has been the subject of much research. As part of a government health survey, researchers measured serum cholesterol levels for a large sample of Americans including children. The distribution for children between 12 and 14 years of age can be fairly well approximated by a normal curve with mean $\mu = 162$ mg/dl and standard deviation $\sigma = 28$ mg/dl. Figure 4.1.1 shows a histogram based on a sample of 727 children between 12 and 14 years old, with the normal curve superimposed.[1] ■

To indicate how the mean μ and standard deviation σ relate to the normal curve, Figure 4.1.2 shows the normal curve for the serum cholesterol distribution of Example 4.1.1, with tick marks at 1, 2, and 3 standard deviations from the mean.

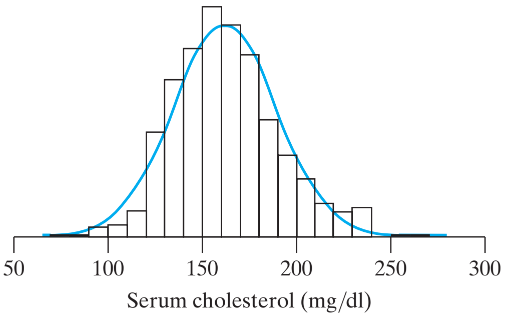


Figure 4.1.1 Distribution of serum cholesterol in 727 12- to 14-year-old children

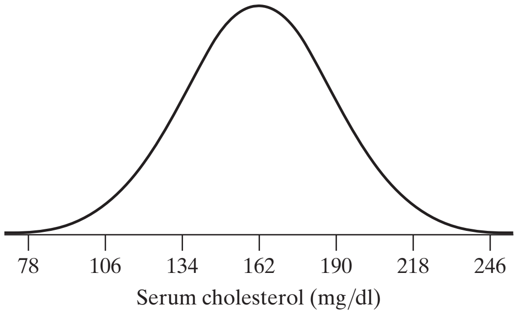


Figure 4.1.2 Normal distribution of serum cholesterol, with $\mu = 162$ mg/dl and $\sigma = 28$ mg/dl

The normal curve can be used to describe the distribution of an observed variable Y in two ways: (1) as a smooth approximation to a histogram based on a sample of Y values; and (2) as an idealized representation of the population distribution of Y. The normal curves in Figures 4.1.1 and 4.1.2 could be interpreted either way. For simplicity, in the remainder of this chapter we will consider the normal curve as representing a population distribution.

Further Examples

We now give three more examples of normal curves that approximately describe real populations. In each figure, the horizontal axis is scaled with tick marks centered at the mean and one standard deviation apart.

Example 4.1.2

Eggshell Thickness In the commercial production of eggs, breakage is a major problem. Consequently, the thickness of the eggshell is an important variable. In one study, the shell thicknesses of the eggs produced by a large flock of White Leghorn hens were observed to follow approximately a normal distribution with mean $\mu = 0.38$ mm and standard deviation $\sigma = 0.03$ mm. This distribution is pictured in Figure 4.1.3.[2] ■

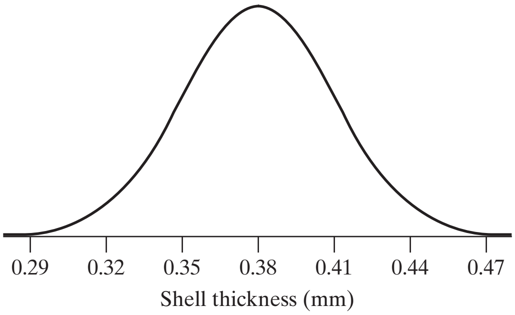


Figure 4.1.3 Normal distribution of eggshell thickness, with $\mu = 0.38$ mm and $\sigma = 0.03$ mm

Example 4.1.3

Interspike Times in Nerve Cells In certain nerve cells, spontaneous electrical discharges are observed that are so rhythmically repetitive that they are called "clock-spikes." The timing of these spikes, even though remarkably regular, does exhibit variation. In one study, the interspike-time intervals (in milliseconds) for a single housefly *(Musca domestica)* were observed to follow approximately a normal distribution with mean $\mu = 15.6$ ms and standard deviation $\sigma = 0.4$ ms; this distribution is shown in Figure 4.1.4.[3] ■

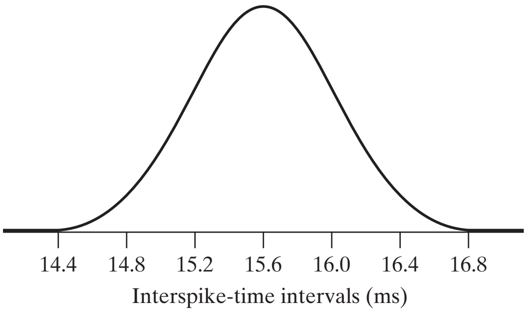


Figure 4.1.4 Normal distribution of interspike-time intervals, with $\mu = 15.6$ ms and $\sigma = 0.4$ ms

The preceding examples have illustrated very different kinds of populations. In Example 4.1.3, the entire population consists of measurements on only one fly. Still another type of population is a *measurement error* population, consisting of repeated measurements of exactly the same quantity. The deviation of an individual measurement from the "correct" value is called measurement error. Measurement error is not the result of a mistake but rather is due to lack of perfect precision in the measuring process or measuring instrument. Measurement error distributions are often approximately normal; in this case the mean of the distribution of repeated measurements of the same quantity is the true value of the quantity (assuming that the measuring instrument is correctly calibrated), and the standard deviation of the distribution indicates the precision of the instrument. One measurement error distribution was described in Example 2.2.12. The following is another example.

Example 4.1.4

Measurement Error When a certain electronic instrument is used for counting particles such as white blood cells, the measurement error distribution is approximately normal. For white blood cells, the standard deviation of repeated counts based on the same blood specimen is about 1.4% of the true count. Thus, if the true count of a certain blood specimen were 7,000 cells/mm^3, then the standard deviation would be about 100 cells/mm^3 and the distribution of repeated counts on that specimen would resemble Figure 4.1.5.[4] ■

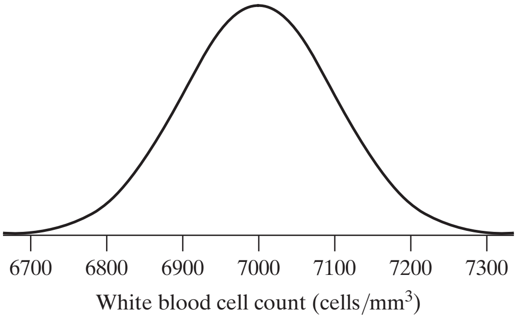


Figure 4.1.5 Normal distribution of repeated white blood cell counts of a blood specimen whose true value is $\mu = 7000$ cells/mm^3. The standard deviation is $\sigma = 100$ cells/mm^3.

4.2 The Normal Curves

As the examples in Section 4.1 show, there are many normal curves; each particular normal curve is characterized by its mean and standard deviation. If a variable Y follows a normal distribution with mean μ and standard deviation σ, then it is common to write $Y \sim N(\mu, \sigma)$. All the normal curves can be described by a single formula. Even though we will not make any direct use of the formula in this book, we present it here, both as a matter of interest and also to emphasize that a normal curve is not just any symmetric curve, but rather a *specific* kind of symmetric curve.

If a variable Y follows a normal distribution with mean μ and standard deviation σ, then the density curve of the distribution of Y is given by the following formula:

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

This function, $f(y)$, is called the *density function* of the distribution and expresses the height of the curve as a function of the position y along the y-axis. The quantities e and π that appear in the formula are constants, with e approximately equal to 2.71 and π approximately equal to 3.14.

Figure 4.2.1 shows a graph of a normal curve. The shape of the curve is like a symmetric bell, centered at $y = \mu$. The direction of curvature is downward (like an inverted bowl) in the central portion of the curve, and upward in the tail portions. The points of inflection (i.e., where the curvature changes direction) are $y = \mu - \sigma$ and $y = \mu + \sigma$; notice that the curve is almost linear near these points. In principle the curve extends to $+\infty$ and $-\infty$, never actually reaching the y-axis; however, the height of the curve is very small for y values more than three standard deviations from the mean. The area under the curve is exactly equal to 1. (Note: It may seem paradoxical that a curve can enclose a finite area, even though it never descends to touch the y-axis. This apparent paradox is clarified in Appendix 4.1.)

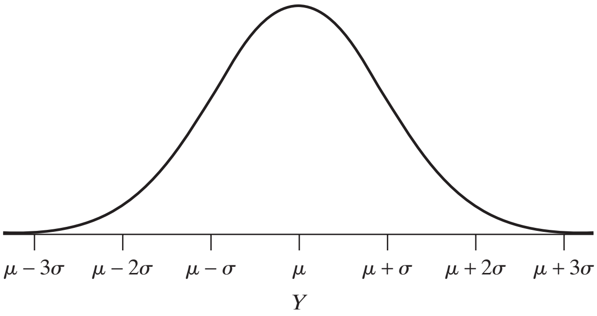


Figure 4.2.1 A normal curve with mean μ and standard deviation σ

All normal curves have the same essential shape, in the sense that they can be made to look identical by suitable choice of the vertical and horizontal scales for each. (For instance, notice that the curves in Figures 4.1.2–4.1.5 look identical.) But normal curves with different values of μ and σ will not look identical if they are all plotted to the same scale, as illustrated by Figure 4.2.2. The location of the normal curve along the y-axis is governed by μ since the curve is centered at $y = \mu$; the width of the curve is governed by σ. The height of the curve is also determined by σ. Since the area under each curve must be equal to 1, a curve with a smaller value of σ must be taller. This reflects the fact that the values of Y are more highly concentrated near the mean when the standard deviation is smaller.

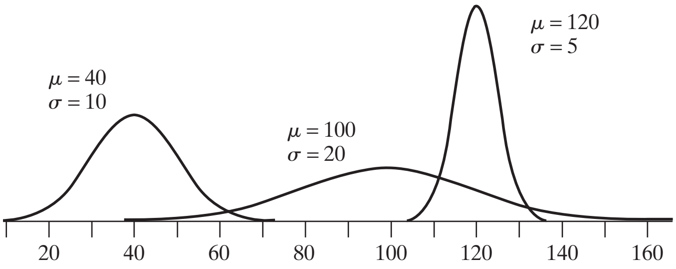


Figure 4.2.2 Three normal curves with different means and standard deviations

4.3 Areas under a Normal Curve

As explained in Section 3.4, a density curve can be quantitatively interpreted in terms of areas under the curve. While areas can be roughly estimated by eye, for some purposes it is desirable to have fairly precise information about areas.

The Standardized Scale

The areas under a normal curve have been computed mathematically and are tabulated here for practical use. The use of this tabulated information is much simplified by the fact that all normal curves can be made equivalent with respect to areas under them by suitable rescaling of the horizontal axis. The rescaled variable is denoted by Z; the relationship between the two scales is shown in Figure 4.3.1.

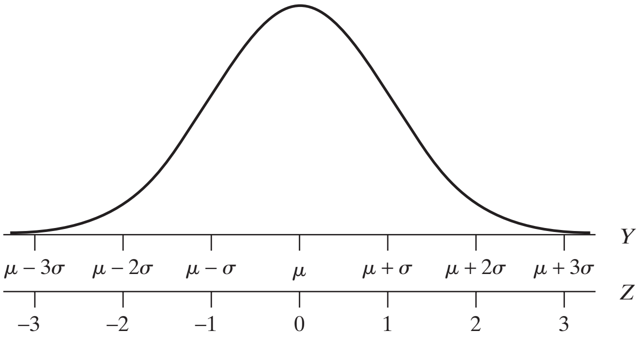


Figure 4.3.1 A normal curve, showing the relationship between the natural scale (Y) and the standardized scale (Z)

As Figure 4.3.1 indicates, the Z scale measures standard deviations from the mean: $z = 1.0$ corresponds to 1.0 standard deviation above the mean; $z = -2.5$ corresponds to 2.5 standard deviations below the mean, and so on. The Z scale is referred to as a **standardized scale**.

The correspondence between the Z scale and the Y scale can be expressed by the formula given in the following box.

Standardization Formula

$$Z = \frac{Y - \mu}{\sigma}$$

The variable Z is referred to as the **standard normal** and its distribution follows a normal curve with mean zero and standard deviation one. Table 3 at the end of this book gives areas under the standard normal curve, with distances along the horizontal axis measured in the Z scale. Each area tabled in Table 3 is the area under the standard normal curve below a specified value of z. For example, for $z = 1.53$, the tabled area is 0.9370; this area is shaded in Figure 4.3.2.

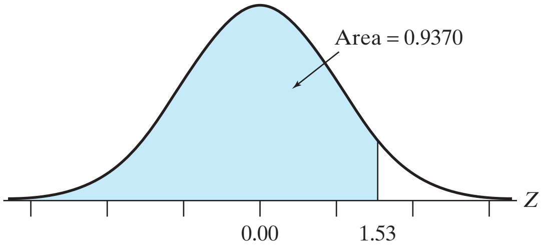


Figure 4.3.2 Illustration of the use of Table 3

If we want to find the area above a given value of z, we subtract the tabulated area from 1. For example, the area above $z = 1.53$ is $1.0000 - 0.9370 = 0.0630$ (Figure 4.3.3).

To find the area between two z values (also commonly called ***z* scores**) we can subtract the areas given in Table 3. For example, to find the area under the Z curve between $z = -1.2$ and $z = 0.8$ (Figure 4.3.4), we take the area below 0.8, which is 0.7881, and subtract the area below −1.2, which is 0.1151, to get $0.7881 - 0.1151 = 0.6730$.

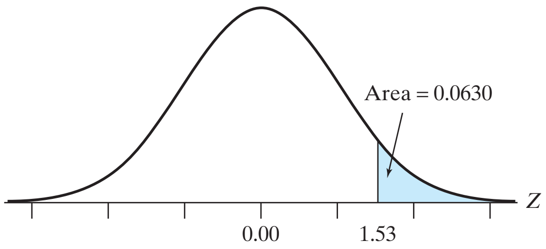


Figure 4.3.3 Area under a standard normal curve above 1.53

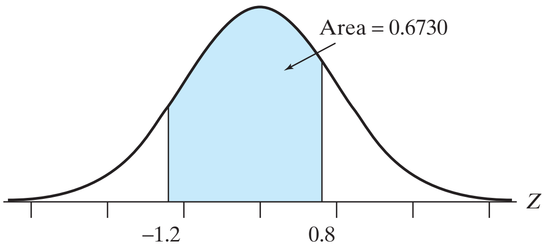


Figure 4.3.4 Area under a standard normal curve between −1.2 and 0.8

Using Table 3, we see that the area under the normal curve between $z = -1$ and $z = +1$ is $0.8413 - 0.1578 = 0.6826$. Thus, for any normal distribution, about 68% of the observations are within ±1 standard deviation of the mean. Likewise, the area under the normal curve between $z = -2$ and $z = +2$ is $0.9772 - 0.0228 = 0.9544$ and the area under the normal curve between $z = -3$ and $z = +3$ is $0.9987 - 0.0013 = 0.9974$. This means that for any normal distribution about 95% of the observations are within ±2 standard deviations of the mean and about 99.7% of the observations are within ±3 standard deviations of the mean. (See Figure 4.3.5.) For example, about 68% of the serum cholesterol values in the idealized distribution of Figure 4.1.2 are between 134 mg/dl and 190 mg/dl, about 95% are between 106 mg/dl and 218 mg/dl, and virtually all are between 78 mg/dl and 246 mg/dl. Figure 4.3.6 shows these percentages.

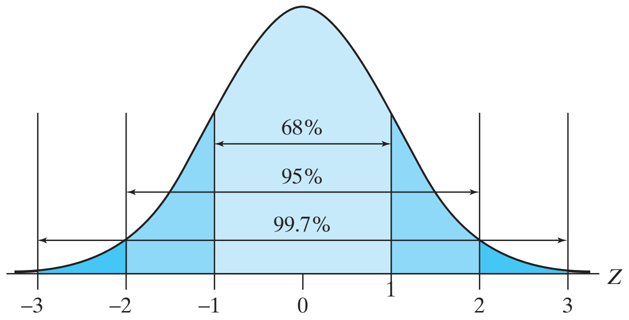


Figure 4.3.5 Areas under a standard normal curve between −1 and +1, between −2 and +2, and between −3 and +3

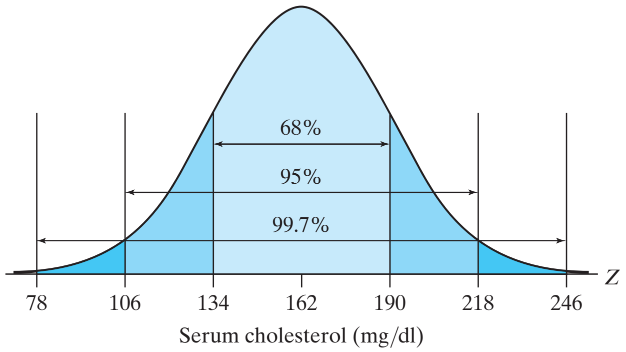


Figure 4.3.6 The 68/95/99.7 rule and the serum cholesterol distribution

If the variable Y follows a normal distribution, then

about 68% of the y's are within ±1 SD of the mean.

about 95% of the y's are within ±2 SDs of the mean.

about 99.7% of the y's are within ±3 SDs of the mean.

These statements provide a very definite interpretation of the standard deviation in cases where a distribution is approximately normal. (In fact, the statements are often approximately true for moderately nonnormal distributions; that is why, in Section 2.6, these percentages—68%, 95%, and >99%—were described as "typical" for "nicely shaped" distributions.)

Determining Areas for a Normal Curve

By taking advantage of the standardized scale, we can use Table 3 to answer detailed questions about any normal population when the population mean and standard deviation are specified. The following example illustrates the use of Table 3. (Of course, the population described in the example is an idealized one, since no actual population follows a normal distribution *exactly*.)

Example 4.3.1

Lengths of Fish In a certain population of the herring *Pomolobus aestivalis*, the lengths of the individual fish follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm.[5] We will use Table 3 to answer various questions about the population.

(a) What percentage of the fish are less than 60 mm long?

Figure 4.3.7 shows the population density curve, with the desired area indicated by shading. In order to use Table 3, we convert the limits of the area from the Y scale to the Z scale, as follows:

For $y = 60$, the z score is

$$z = \frac{y - \mu}{\sigma} = \frac{60 - 54}{4.5} = 1.33$$

Thus, the question "What percentage of the fish are less than 60 mm long?" is equivalent to the question "What is the area under the standard normal curve below the z value of 1.33?" Looking up $z = 1.33$ in Table 3, we find that the area is 0.9082; thus, 90.82% of the fish are less than 60 mm long.

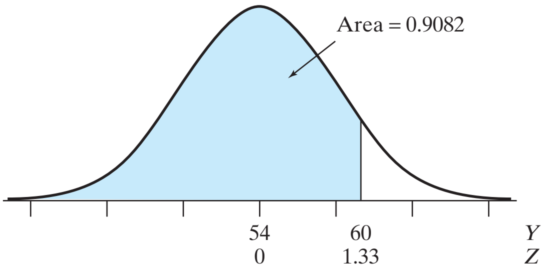


Figure 4.3.7 Area under the normal curve in Example 4.3.1(a)

(b) What percentage of the fish are more than 51 mm long?

The standardized value for $y = 51$ is

$$z = \frac{y - \mu}{\sigma} = \frac{51 - 54}{4.5} = -0.67$$

Thus, the question "What percentage of the fish are more than 51 mm long?" is equivalent to the question "What is the area under the standard normal curve above the z value of −0.67?" Figure 4.3.8 shows this relationship. Look-

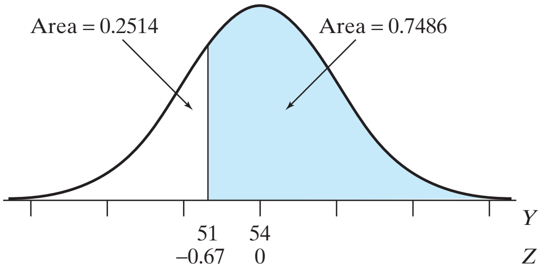


Figure 4.3.8 Area under the normal curve in Example 4.3.1(b)

ing up $z = -0.67$ in Table 3, we find that the area below $z = -0.67$ is 0.2514. This means that the area above $z = -0.67$ is $1 - 0.2514 = 0.7486$. Thus, 74.86% of the fish are more than 51 mm long.

(c) What percentage of the fish are between 51 and 60 mm long?

Figure 4.3.9 shows the desired area. This area can be expressed as a difference of two areas found from Table 3. The area below $y = 60$ is 0.9082, as found in part (a), and the area below $y = 51$ is 0.2514, as found in part (b). Consequently, the desired area is computed as

$$0.9082 - 0.2514 = 0.6568$$

Thus, 65.68% of the fish are between 51 and 60 mm long.

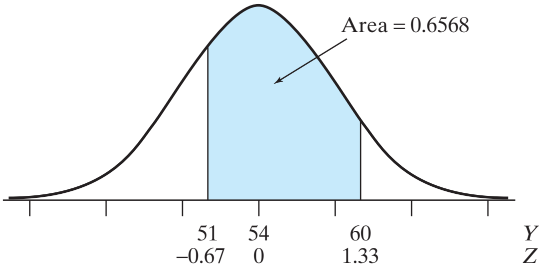


Figure 4.3.9 Area under the normal curve in Example 4.3.1(c)

(d) What percentage of the fish are between 58 and 60 mm long?

Figure 4.3.10 shows the desired area. This area can be expressed as a difference of two areas found from Table 3. The area below $y = 60$ is 0.9082, as was found in part (a). To find the area below $y = 58$, we first calculate the z value that corresponds to $y = 58$:

$$z = \frac{y - \mu}{\sigma} = \frac{58 - 54}{4.5} = 0.89$$

The area under the Z curve below $z = 0.89$ is 0.8133. Consequently, the desired area is computed as

$$0.9082 - 0.8133 = 0.0949$$

Thus, 9.49% of the fish are between 58 and 60 mm long. ■

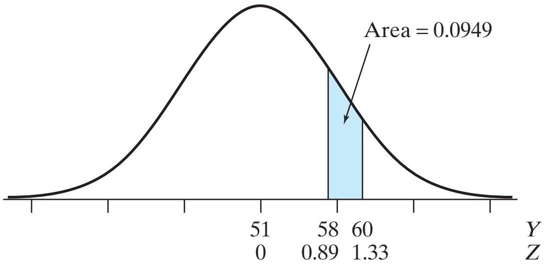


Figure 4.3.10 Area under the normal curve in Example 4.3.1(d)

Each of the percentages found in Example 4.3.1 can also be interpreted in terms of probability. Let the random variable Y represent the length of a fish randomly chosen from the population. Then the results in Example 4.3.1 imply that

$$\begin{aligned}\Pr\{Y < 60\} &= 0.9082\\ \Pr\{Y > 51\} &= 0.7486\\ \Pr\{51 < Y < 60\} &= 0.6568\end{aligned}$$

and

$$\Pr\{58 < Y < 60\} = 0.0949$$

Thus, the normal distribution can be interpreted as a continuous probability distribution.

Note that because the idealized normal distribution is perfectly continuous, probabilities such as

$$\Pr\{Y > 48\} \text{ and } \Pr\{Y \geq 48\}$$

are equal (see Section 3.4). That is,

$$\begin{aligned}\Pr\{Y \geq 48\} &= \Pr\{Y > 48\} + \Pr\{Y = 48\}\\ &= \Pr\{Y > 48\} + 0 \text{ (since } Y \text{ is taken to be continuous)}\\ &= \Pr\{Y > 48\}\end{aligned}$$

If, however, the length were measured only to the nearest mm, then the measured variable would actually be discrete, so that $\Pr\{Y > 48\}$ and $\Pr\{Y \geq 48\}$ would differ somewhat from each other. In cases where this discrepancy is important, the computation can be refined to take into account the discontinuity of the measured distribution (we will later see such an example in Section 5.4).

Inverse Reading of Table 3

In determining facts about a normal distribution, it is sometimes necessary to read Table 3 in an "inverse" way—that is, to find the value of z corresponding to a given area rather than the other way around. For example, suppose we want to find the value on the Z scale that cuts off the top 2.5% of the distribution. This number is 1.96, as shown in Figure 4.3.11.

We will find it helpful, for future reference, to introduce some notation. We will use the notation z_α to denote the number such that $\Pr\{Z < z_\alpha\} = 1 - \alpha$ and $\Pr\{Z > z_\alpha\} = \alpha$, as shown in Figure 4.3.12. Thus, $z_{0.025} = 1.96$.

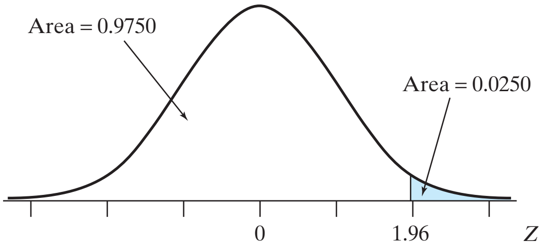


Figure 4.3.11 Area under the normal curve above 1.96

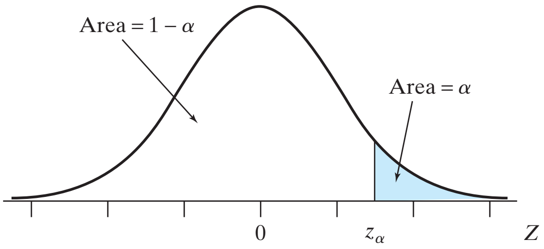


Figure 4.3.12 Area under the normal curve above α

We often need to determine a z_α value when we want to determine a *percentile* of a normal distribution. The percentiles of a distribution divide the distribution into 100 equal parts, just as the quartiles divide it into 4 equal parts [from the Latin roots *centum* ("hundred") and *quartus* ("fourth")]. For example, suppose we want to find the 70th percentile of a standard normal distribution. That means that we want to find the number $z_{0.30}$ that divides the standard normal distribution into two parts: the bottom 70% and the top 30%. As Figure 4.3.13 illustrates, we need to look in Table 3 for an area of 0.7000. The closest value is an area of 0.6985, corresponding to a z value of 0.52. Thus, $z_{0.30} = 0.52$.

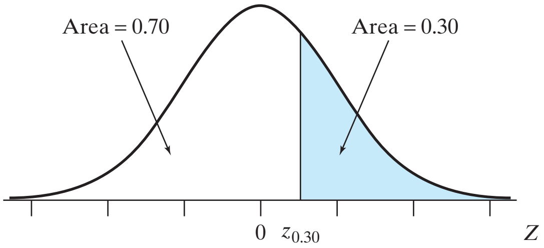


Figure 4.3.13 Determining the 70th percentile of a normal distribution

Example 4.3.2 Lengths of Fish

(a) Suppose we want to find the 70th percentile of the fish length distribution of Example 4.3.1. Let us denote the 70th percentile by y^*. By definition, y^* is the value such that 70% of the fish lengths are less than y^* and 30% are greater, as illustrated in Figure 4.3.14.

To find y^*, we use the value of $z_{0.30} = 0.52$ that we just determined. Next we convert this z value to the Y scale. We know that if we were given the value of y^*, we could convert it to a standard normal (z scale) and the result would be 0.52. Thus, from the standardization formula we obtain the equation

$$0.52 = \frac{y^* - 54}{45}$$

which can be solved to give $y^* = 54 + 0.52 \times 4.5 = 56.3$. The 70th percentile of the fish length distribution is 56.3 mm.

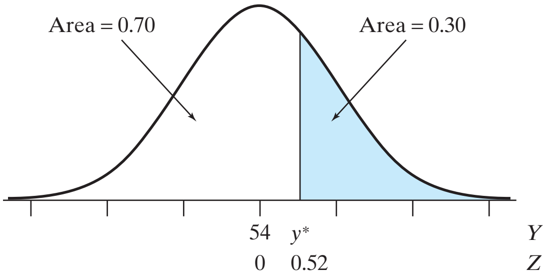


Figure 4.3.14 Determining the 70th percentile of a normal distribution, Example 4.3.2(a)

(b) Suppose we want to find the 20th percentile of the fish length distribution of Example 4.3.1. Let us denote the 20th percentile by y^*. By definition, y^* is the value such that 20% of the fish lengths are less than y^* and 80% are greater, as illustrated in Figure 4.3.15.

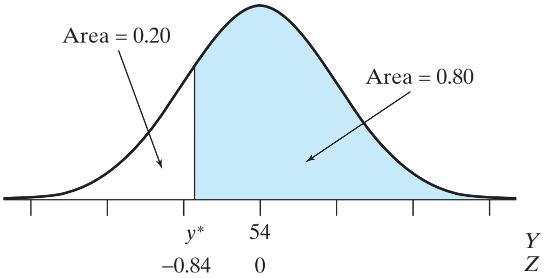


Figure 4.3.15 Determining the 20th percentile of a normal distribution, Example 4.3.2(b)

To find y^* we first determine the value of $z_{0.80}$, which is the 20th percentile in the Z scale. As Figure 4.3.15 illustrates, we need to look in Table 3 for an area of .2000. The closest value is an area of .2005, corresponding to $z = -0.84$. The next step is to convert this z value to the Y scale. From the standardization formula, we obtain the equation

$$-0.84 = \frac{y^* - 54}{45}$$

which can be solved to give $y^* = 54 - 0.84 \times 4.5 = 50.2$. The 20th percentile of the fish length distribution is 50.2 mm. ■

Problem-Solving Tip In solving problems that require the use of Table 3, a sketch of the distribution (as in Figures 4.3.7–4.3.10 and 4.3.14–4.3.15) is a very handy aid to straight thinking.

While Table 3 is handy for carrying out the sorts of computations discussed previously, computer software may also be used to find normal probabilities directly without the need for any standardization.

Exercises 4.3.1–4.3.16

4.3.1 Suppose a certain population of observations is normally distributed. What percentage of the observations in the population

(a) are within ±1.5 standard deviations of the mean?

(b) are more than 2.5 standard deviations above the mean?

(c) are more than 3.5 standard deviations away from (above or below) the mean?

4.3.2

(a) The 90th percentile of a normal distribution is how many standard deviations above the mean?

(b) The 10th percentile of a normal distribution is how many standard deviations below the mean?

4.3.3 The brain weights of a certain population of adult Swedish males follow approximately a normal distribution with mean 1,400 gm and standard deviation 100 gm.[6] What percentage of the brain weights are

(a) 1,500 gm or less?

(b) between 1,325 and 1,500 gm?

(c) 1,325 gm or more?

(d) 1,475 gm or more?

(e) between 1,475 and 1,600 gm?

(f) between 1,200 and 1,325 gm?

4.3.4 Let Y represent a brain weight randomly chosen from the population of Exercise 4.3.3. Find

(a) $\Pr\{Y \leq 1{,}325\}$

(b) $\Pr\{1{,}475 \leq Y \leq 1{,}600\}$

4.3.5 In an agricultural experiment, a large uniform field was planted with a single variety of wheat. The field was divided into many plots (each plot being 7 × 100 ft) and the yield (lb) of grain was measured for each plot. These plot yields followed approximately a normal distribution with mean 88 lb and standard deviation 7 lb.[7] What percentage of the plot yields were

(a) 80 lb or more? (b) 90 lb or more?

(c) 75 lb or less? (d) between 75 and 90 lb?

(e) between 90 and 100 lb? (f) between 75 and 80 lb?

4.3.6 Refer to Exercise 4.3.5. Let Y represent the yield of a plot chosen at random from the field. Find

(a) $\Pr\{Y > 90\}$ (b) $\Pr\{75 < Y < 90\}$

4.3.7 Consider a standard normal distribution, Z. Find

(a) $z_{0.10}$ (b) $z_{0.25}$ (c) $z_{0.05}$ (d) $z_{0.01}$

4.3.8 For the wheat-yield distribution of Exercise 4.3.5, find

(a) the 65th percentile (b) the 35th percentile

4.3.9 The serum cholesterol levels of 12- to 14-year-olds follow a normal distribution with mean 162 mg/dl and standard deviation 28 mg/dl. What percentage of 12 to 14-year-olds have serum cholesterol values

(a) 171 or more? (b) 143 or less?

(c) 194 or less? (d) 105 or more?

(e) between 166 and 194? (f) between 105 and 138?

(g) between 138 and 166?

4.3.10 Refer to Exercise 4.3.9. Suppose a 13-year-old is chosen at random and let Y be the person's serum cholesterol value. Find

(a) $\Pr\{Y \geq 166\}$ (b) $\Pr\{166 < Y < 194\}$

4.3.11 For the serum cholesterol distribution of Exercise 4.3.9, find

(a) the 80th percentile (b) the 20th percentile

4.3.12 When red blood cells are counted using a certain electronic counter, the standard deviation of repeated counts of the same blood specimen is about 0.8% of the true value, and the distribution of repeated counts is approximately normal.[8] For example, this means that if the true value is 5,000,000 cells/mm^3, then the SD is 40,000.

(a) If the true value of the red blood count for a certain specimen is 5,000,000 cells/mm^3, what is the probability that the counter would give a reading between 4,900,000 and 5,100,000?

(b) If the true value of the red blood count for a certain specimen is μ, what is the probability that the counter would give a reading between 0.98μ and 1.02μ?

(c) A hospital lab performs counts of many specimens every day. For what percentage of these specimens does the reported blood count differ from the correct value by 2% or more?

4.3.13 The amount of growth, in a 15-day period, for a population of sunflower plants was found to follow a normal distribution with mean 3.18 cm and standard deviation 0.53 cm.[9] What percentage of plants grow

(a) 4 cm or more? (b) 3 cm or less?

(c) between 2.5 and 3.5 cm?

4.3.14 Refer to Exercise 4.3.13. In what range do the middle 90% of all growth values lie?

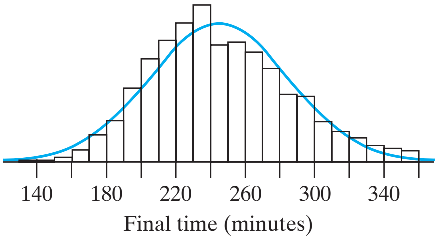
4.3.15 For the sunflower plant growth distribution of Exercise 4.3.13, what is the 25th percentile?

4.3.16 Many cities sponsor marathons each year. The following histogram shows the distribution of times that it took for 10,002 runners to complete the Rome marathon in 2008, with a normal curve superimposed. The fastest runner completed the 26.3-mile course in 2 hours and 9 minutes, or 129 minutes. The average time was 245 minutes and the standard deviation was 40 minutes. Use the normal curve to answer the following questions.[10]

(a) What percentage of times were greater than 200 minutes?

(b) What is the 60th percentile of the times?

(c) Notice that the normal curve approximation is fairly good except around the 240-minute mark. How can we explain this anomalous behavior of the distribution?



4.4 Assessing Normality

Many statistical procedures are based on having data from a normal population. In this section we consider ways to assess whether it is reasonable to use a normal curve model for a set of data and, if not, how we might proceed.

Recall from Section 4.3 that if the variable Y follows a normal distribution, then

about 68% of the y's are within ±1 SD of the mean.

about 95% of the y's are within ±2 SDs of the mean.

about 99.7% of the y's are within ±3 SDs of the mean.

We can use these facts as a check of how closely a normal curve model fits a set of data.

Example 4.4.1

Serum Cholesterol For the serum cholesterol data of Example 4.1.1, the sample mean is 162 and the sample SD is 28. The interval "mean ± SD" is

$$(162 - 28, 162 + 28) \text{ or } (134, 190)$$

This interval contains 509 of the 727 observations, or 70.0% of the data. Likewise, the interval

$$(162 - 2 \times 28, 162 + 2 \times 28) \text{ is } (106, 218)$$

which contains 685, or 94.2%, of the 727 observations. Finally, the interval

$$(162 - 3 \times 28, 162 + 3 \times 28) \text{ is } (78, 246)$$

which contains 724, or 99.6%, of the 727 observations. The three observed percentages

70.0%, 94.2%, and 99.6%

agree quite well with the theoretical percentages of

68%, 95%, and 99.7%

This agreement supports the claim that serum cholesterol levels for 12- to 14-year-olds have a normal distribution. This reinforces the visual evidence of Figure 4.1.1. ■

Example 4.4.2

Moisture Content Moisture content was measured in each of 83 freshwater fruit.[11] Figure 4.4.1 shows that this distribution is strongly skewed to the left. The sample mean of these data is 80.7 and the sample SD is 12.7. The interval

$$(80.7 - 12.7, 80.7 + 12.7)$$

contains 70, or 84.3%, of the 83 observations. The interval

$$(80.7 - 2 \times 12.7, 80.7 + 2 \times 12.7)$$

contains 78, or 94.0%, of the 83 observations. Finally, the interval

$$(80.7 - 3 \times 12.7, 80.7 + 3 \times 12.7)$$

contains 80, or 96.4%, of the 83 observations. The three percentages

84.3%, 94.0%, and 96.4%

differ from the theoretical percentages of

68%, 95%, and 99.7%

because the distribution is far from being bell-shaped. This reinforces the visual evidence of Figure 4.4.1. ■

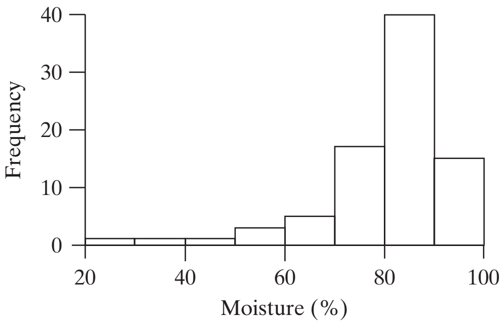


Figure 4.4.1 Moisture content in freshwater fruit

Normal Probability Plots

A **normal probability plot** is a special statistical graph that is used to assess normality. We present this statistical tool with an example using the heights (in inches) of a sample of 11 women, sorted from smallest to largest:

61, 62.5, 63, 64, 64.5, 65, 66.5, 67, 68, 68.5, 70.5

Based on these data, does it make sense to use a normal curve to model the distribution of women's heights? Figure 4.4.2 is a histogram of the data with a normal curve superimposed, using the sample mean of 65.5 and the sample standard deviation of 2.9 as the parameters of the normal curve. This histogram is fairly symmetric, but when we have a small sample, it can be hard to tell the shape of the population distribution by looking at a histogram.

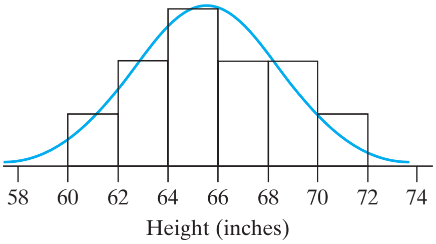


Figure 4.4.2 Histogram of the heights of 11 women

Because it is often difficult to visually examine a histogram and decide if it is bell-shaped or not, a visually simpler plot, the normal probability plot, was developed.* A normal probability plot is a scatterplot that compares our observed data values to values we would expect to see if the population were normal. If the data come from a normal population, the points in this plot should follow a straight line, which is much easier to visually recognize than a bell shape of a jagged histogram. As many statistical procedures are based on the condition that the data came from a normal population, it is important to be able to assess normality.

How Normal Probability Plots Work

In Examples 4.4.1 and 4.4.2 we compared the observed proportion of data that falls within 1, 2, and 3 SDs of the mean and then compared those values to the proportions we would expect to find if the data were from a normal population. It is natural to consider these intervals, but we could consider other intervals as well. For example, we would expect about 86.6% of normal data to fall within 1.5 SDs of the mean and 96.4% to within 2.1.† We could even consider one-sided intervals. For example, we would expect 84.1% of normal data values to be less than the mean plus 1 SD.

Rather than focus on comparing percentages, we could instead focus on comparing actual observed women's heights to heights we would expect to see if the data were from a normal population. For example, the shortest woman in our sample is 61 inches tall; that is, 1/11th (or 0.0909) of the sample is 61 inches or shorter. If heights of women really follow a normal distribution, with mean 65.5 and standard deviation 2.9, then we would expect the 9.09th percentile to be $\mu + z_{(1-.0909)}\sigma = 65.5 - 1.34 \times 2.9$ or 61.6 inches. This value is close to the observed

*Though visually simple, the construction of these plots is complex and typically performed using statistical software.

†These values can be verified using the techniques of Section 4.3.

value of 61 inches. We could repeat this sort of calculation for each of the 11 observed data values. A normal probability plot provides a visual comparison of these values.

The first step in creating a normal probability plot, therefore, is to compute the sample percentiles. Example 4.4.3 presents this computation, which is typically performed by statistical software.

Example 4.4.3

Height of Eleven Women Sorting the data from smallest to largest we observe that 1/11th (= 9.1%) of our sample is 61 inches or shorter, 2/11ths (= 18.2%) is 62.5 inches or shorter, … 10/11ths (90.9%) is 68.5 inches or shorter and 11/11ths (100%) is 70.5 inches or shorter. Unfortunately, computing percentages in this simplistic way (i.e., $100 \times i/n$ where i is the sorted observation number) creates some implausible population estimates. For example, it seems unreasonable to believe that 100% of the *population* is 70.5 inches or shorter when, after all, we are observing only a small sample; a larger sample would likely observe some taller women. To correct for this, an alternative and more reasonable percentage for each data value is computed as $100\left(i - \frac{1}{2}\right)/n$ where i is the index of the data value in the sorted list.* These adjusted percentiles are tabulated in Table 4.4.1. Note that these values actually do not depend on the data observed; they depend only on the number of data values in the sample.

Table 4.4.1 Computing indices and percentiles for the heights of eleven women

i	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Percentile $100(i/11)$	9.09	18.18	27.27	36.36	45.45	54.55	63.64	72.73	81.82	90.91	100.00
Adjusted percentile $100\left(i - \frac{1}{2}\right)/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45

Once we have the adjusted percentiles we find the corresponding z scores using Table 3 or a computer. Then, with these z scores we find the theoretical heights: $\mu + z \times \sigma$ as in Example 4.4.4.

Example 4.4.4

Heights of Eleven Women The shortest woman's adjusted percentile is 4.55%. The corresponding z score is $z_{(1-0.0455)} = z_{0.9545} = -1.69$. In this example, the sample mean and standard deviation are 65.5 and 2.9, respectively, so the expected height of the shortest woman in a sample of 11 women from a normal population is $65.5 - 1.69 \times 2.9 = 60.6$ inches. The z scores and theoretical heights for this woman and the remaining 10 women appear in Table 4.4.2.

Table 4.4.2 Computing theoretical z scores and heights for eleven women

i	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile $100\left(i - \frac{1}{2}\right)/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45
z	−1.69	−1.10	−0.75	−0.47	−0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4

*Different software packages may compute these proportions differently and may also modify the formula based on sample size. The preceding formula is used by the software package *R* when $n > 10$.

Next, by plotting the observed heights against the theoretical heights in a scatterplot as in Figure 4.4.3, we may visually compare the values. In this case our plot appears fairly linear, suggesting that the observed values generally agree with the theoretical values—that the normal model provides a reasonable approximation to the data. If the data do not agree with the normal model, then the plot will show strong nonlinear patterns such as curvature or S shapes.

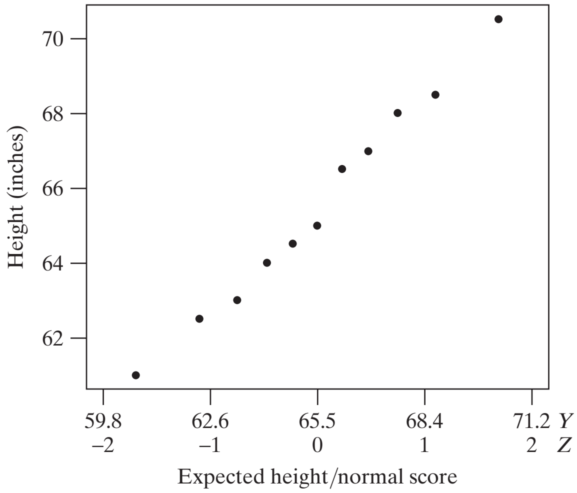


Figure 4.4.3 Normal probability plot of the heights of 11 women

Because of the one-to-one correspondence between the z scores and theoretical values, it is not common to put both sets of labels on the x-axis as in Figure 4.4.3. Traditionally only the z scores are displayed.* ■

Making Decisions about Normality

Of course, even when we sample from a perfectly normal distribution, we have to expect that there will be some variability between the sample we obtain and the theoretical normal scores. Figure 4.4.4 shows six normal probability plots based on samples taken from a $N(0, 1)$ distribution. Notice that all six plots show a general linear pattern. It is true that there is a fair amount of "wiggle" in some of the plots, but the important feature of each of these plots is that we can draw a line that captures the trend in the bulk of the points, with little deviation away from this line, even at the extremes.

If the points in the normal probability plot do not fall more or less along a straight line, then there is an indication that the data are not from a normal population. For example, if the top of the plot bends up, that means the y values at the upper end of the distribution are too large for the distribution to be bell-shaped; that is, the distribution is skewed to the right or has large outliers, as in Figure 4.4.5.

If the bottom of the plot bends down, that means the y values at the lower end of the distribution are too small for the distribution to be bell-shaped; that is, the distribution is skewed to the left or has small outliers. Figure 4.4.6 shows the distribution of moisture content in the freshwater fruit from Example 4.4.2, which is strongly skewed to the left.

*Some software programs create normal probability plots with the normal scores on the vertical axis and the observed data on the horizontal axis.

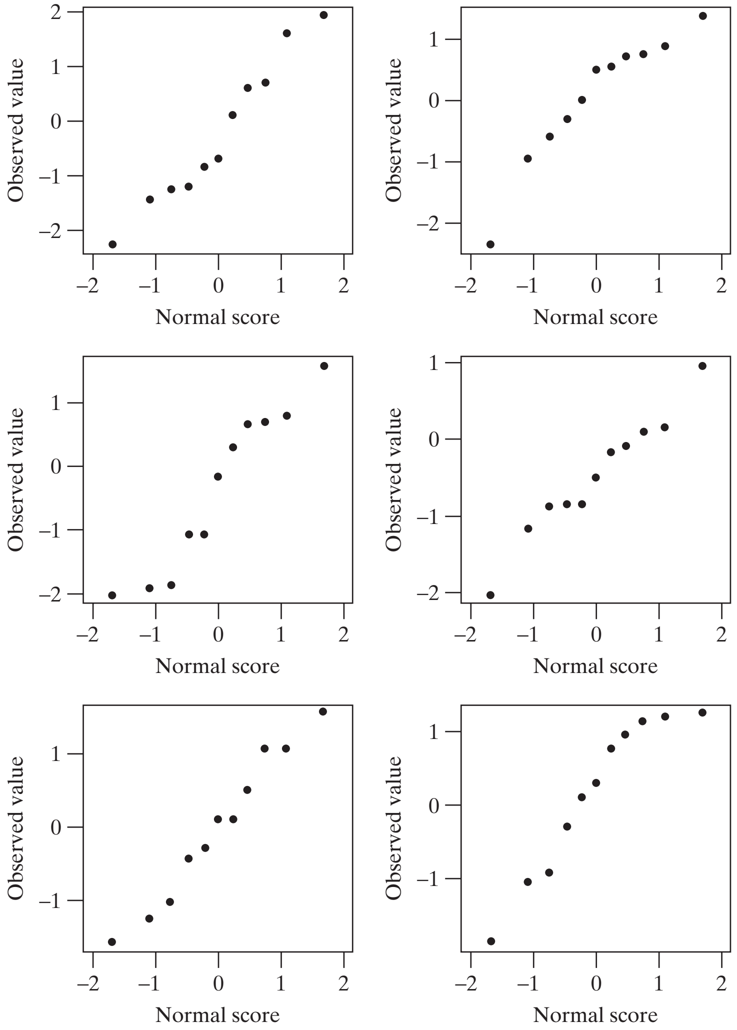


Figure 4.4.4 Normal probability plots for normal data

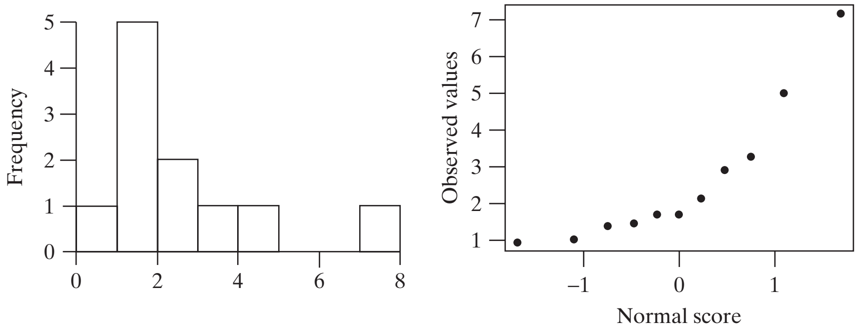


Figure 4.4.5 Histogram and normal probability plot of a distribution that is skewed to the left

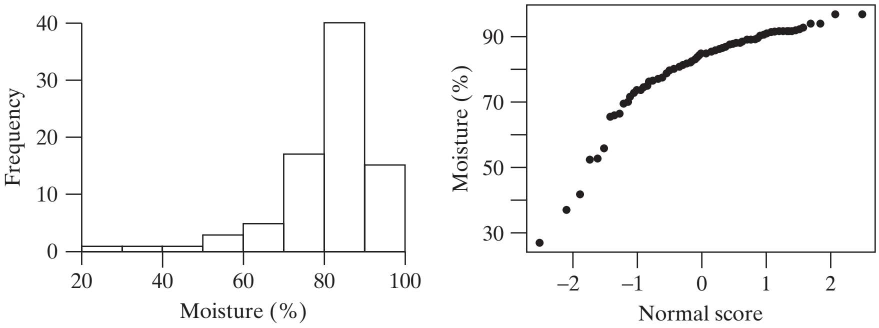


Figure 4.4.6 Histogram and normal probability plot of a distribution that is skewed to the left

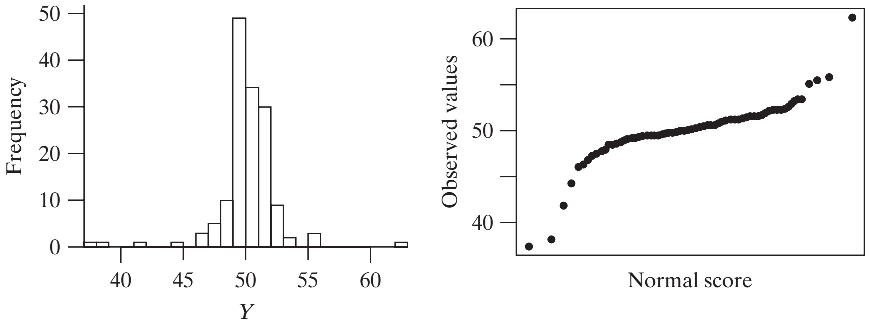


Figure 4.4.7 Histogram and normal probability plot of a distribution that has long tails

If a distribution has a very long left-hand tail and a long right-hand tail, when compared to a normal curve, then the normal probability plot will have something of an S shape. Figure 4.4.7 shows such a distribution.

Sometimes the same value shows up repeatedly in a sample, due to rounding in the measurement process. This leads to *granularity* in the normal probability plot, as in Figure 4.4.8, but this does not stop us from inferring that the underlying distribution is normal.

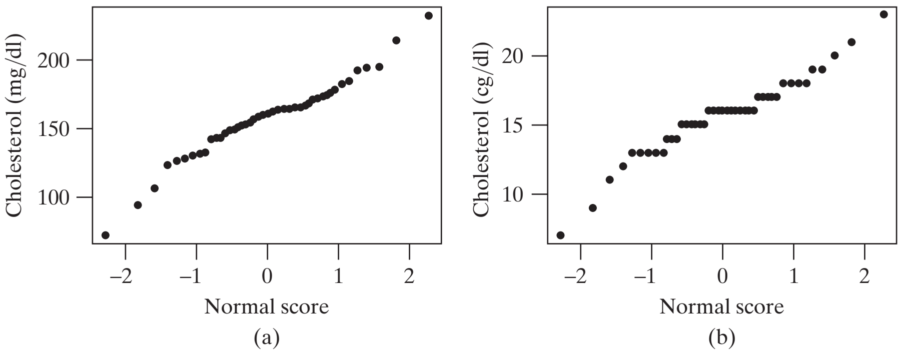


Figure 4.4.8 Normal probability plots of cholesterol values of fifty 12- to 14-year-olds measured to (a) the nearest mg/dl and (b) the nearest cg/dl

Transformations for Nonnormal Data

A normal probability plot can help us assess whether or not the data came from a normal distribution. Sometimes a histogram or normal probability plot shows that our data are nonnormal, but a transformation of the data gives us a symmetric, bell-shaped curve. In such a situation, we may wish to transform the data and continue our analysis in the new (transformed) scale.

Example 4.4.5

Lentil Growth The histogram and normal probability plot in Figure 4.4.9 show the distribution of the growth rate, in cm per day, for a sample of 47 lentil plants.[12] This distribution is skewed to the right. If we take the logarithm of each observation, we

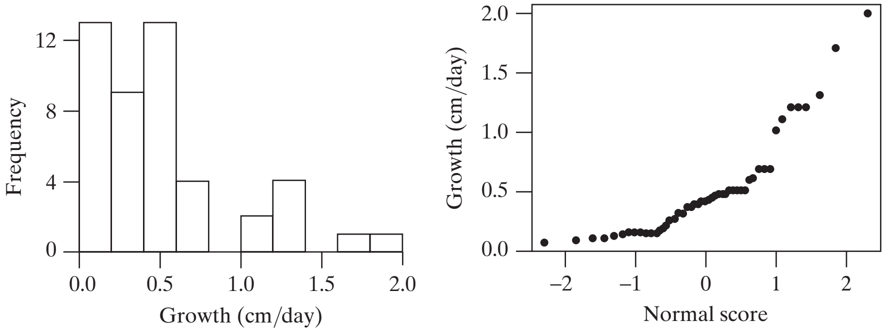


Figure 4.4.9 Histogram and normal probability plot of growth rates of 47 lentil plants

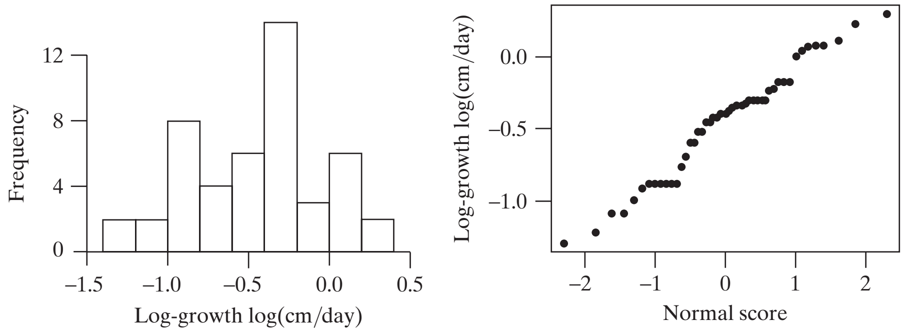


Figure 4.4.10 Histogram and normal probability plot of the logarithms of the growth rates of 47 lentil plants

get a distribution that is much more nearly symmetric. The plots in Figure 4.4.10 show that in log scale the growth rate distribution is approximately normal. (In Figure 4.4.10 the base 10 logarithm, $\log_{10}$, is used, but we could use any base, such as the natural log, $\log_e = \ln$, and the effect on the shape of the distribution would be the same.) ■

In general, if the distribution is skewed to the right then one of the following transformations should be considered: $\sqrt{Y}$, $\log Y$, $1/\sqrt{Y}$, $1/Y$. These transformations will pull in the long right-hand tail and push out the short left-hand tail, making the distribution more nearly symmetric. Each of these is more drastic than the one before. Thus, a square root transformation will change a mildly skewed distribution into a symmetric distribution, but a log transformation may be needed if the distribution is more heavily skewed, and so on. For example, we saw in Example 2.7.6 how a square root transformation pulls in a long right-hand tail and how a log transformation pulls in the right-hand tail even more. If the distribution of a variable Y is skewed to the left, then raising Y to a power greater than 1 can be helpful.

An Objective Measure of Abnormality: The Shapiro–Wilk Test (optional)

While normal probability plots are better than histograms to visually assess departures of normality, our visual perception is still subjective. The data appearing in the probability plots of Figure 4.4.4 come from a normal population, but to untrained eyes (and even to some trained ones) a few of the plots might be interpreted as being nonnormal. The **Shapiro–Wilk test** is a statistical procedure that numerically assesses evidence for certain types of nonnormality in data. As with the normal probability plot, the mechanics of the procedure is complex, but fortunately many statistical software packages will perform this or similar tests of normality.*

The output of a Shapiro–Wilk test is a P-value† and is interpreted as follows:

P-value < 0.001	Very strong evidence for nonnormality
P-value < 0.01	Strong evidence for nonnormality
P-value < 0.05	Moderate evidence for nonnormality
P-value < 0.10	Mild or weak evidence for nonnormality
P-value ≥ 0.10	No compelling evidence for nonnormality

*The Ryan–Joiner, Anderson–Darling, and Kolmogorov–Smirnoff tests are other tests of nonnormality commonly found in statistical software packages.

†As we shall see in much greater detail in Chapter 7, a P-value is not unique to testing for normality. In a test of all sorts of hypotheses, the weight of evidence for the hypothesis in question (in this case—the Shapiro–Wilk test—the hypothesis is that the data are nonnormal) can be reported using this term. Small P-values are interpreted as evidence for the hypothesis in question.

Example 4.4.6 illustrates the Shapiro–Wilk test for the lentil growth data of Example 4.4.5.

Example 4.4.6

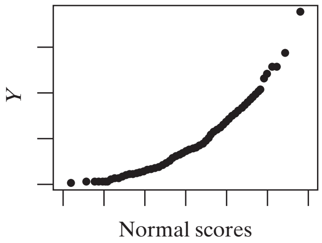
Lentil Growth For the untransformed lentil data in Figure 4.4.9, the *P*-value (reported from the statistical software package *R*) for the Shapiro–Wilk test is 0.000006. Thus, there is very strong evidence that lentil growth does not follow a normal distribution. For the transformed data in Figure 4.4.10, however, the *P*-value for the Shapiro–Wilk test is 0.2090, indicating that there is no compelling evidence for nonnormality of the log-transformed growth data. ■

Caution. The use of this test procedure and *P*-value is somewhat like the use of the "check engine light" on a car. When the *P*-value is small, there is an indication of nonnormality. This is like your engine light coming on: You pull over and assess the situation. Likewise, as we shall see in future chapters, when we have nonnormal data, we will carefully have to assess how to proceed with our analyses. On the other hand, when the *P*-value is not small ($\geq$0.10) we don't have evidence of nonnormality. This is similar to your engine light staying off: You continue to drive forward without worry, *but* this does not guarantee that your car is perfectly OK. Your car could break down at any time. Of course, if we were constantly worried about our car even when the check engine light were off, we would perpetually find ourselves paralyzed and pulled over at the side of the road. Analogously, when the *P*-value from the Shapiro–Wilk tests is not small (the light is off), this only means that there is no compelling evidence for nonnormality. It does not guarantee that the population is, in fact, normal.

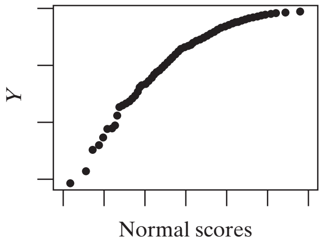
Exercises 4.4.1–4.4.8

4.4.1 In Example 4.1.2 it was stated that shell thicknesses in a population of eggs follow a normal distribution with mean μ = 0.38 mm and standard deviation σ = 0.03 mm. Use the 68%–95%–99.7% rule to determine intervals, centered at the mean, that include 68%, 95%, and 99.7% of the shell thicknesses in the distribution.

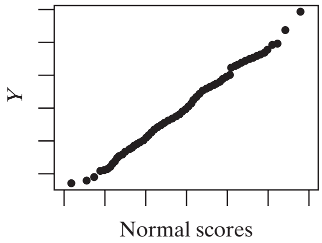
4.4.2 The following three normal probability plots, (a), (b), and (c), were generated from the distributions shown by histograms I, II, and III. Which normal probability plot goes with which histogram? How do you know?



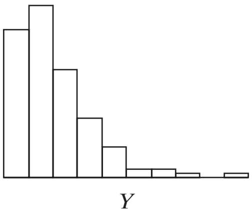
(a)



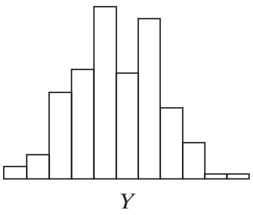
(b)



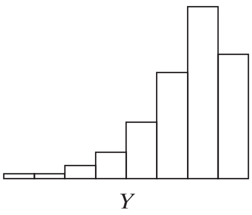
(c)



I

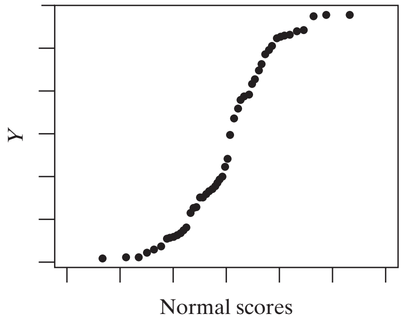


II

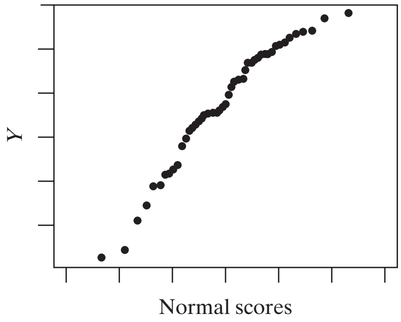


III

4.4.3 For each of the following normal probability plots, sketch the corresponding histogram of the data.



(a)



(b)

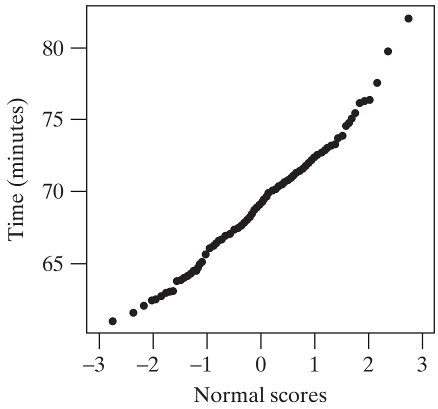
4.4.4 The mean daily rainfall between January 1, 2007, through January 1, 2009, at Pismo Beach, California, was 0.02 inches with a standard deviation of 0.11 inches. Based on this information, do you think it is reasonable to believe that daily rainfall at Pismo Beach follows a normal distribution? Explain. (*Hint*: Think about the possible values for daily rainfall.)[13]

4.4.5 The mean February 1 daily high temperature in Juneau, Alaska, between 1945 and 2005 was 1.1 °C with a standard deviation of 1.9 °C.[14]

(a) Based on this information, do you think it is reasonable to believe that the February 1 daily high temperatures in Juneau, Alaska, follow a normal distribution? Explain.

(b) Does this information provide compelling evidence that the February 1 daily high temperatures in Juneau, Alaska, follow a normal distribution? Explain.

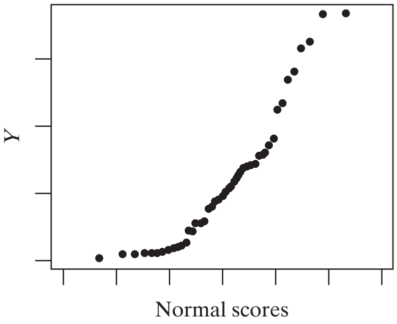
4.4.6 The following normal probability plot was created from the times that it took 166 bicycle riders to complete the stage 11 time trial, from Grenoble to Chamrousse, France, in the 2001 Tour de France cycling race.



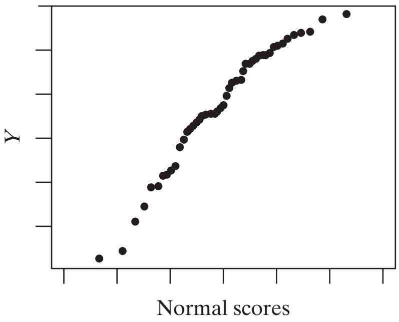
(a) Consider the fastest riders. Are their times better than, worse than, or roughly equal to the times one would expect the fastest riders to have if the data came from a truly normal distribution?

(b) Consider the slowest riders. Are their times better than, worse than, or roughly equal to the times one would expect the slowest riders to have if the data came from a truly normal distribution?

4.4.7 The *P*-values for the Shapiro–Wilk test for the data appearing in probability plots (a) and (b) are 0.235 and 0.00015. Which *P*-value corresponds to which plot? What is the basis for your decision?



(a)



(b)

4.4.8

(a) The P-value for the Shapiro–Wilk test of normality for the data in Exercise 4.4.3(b) is 0.039. Using this value to justify your answer, does it seem reasonable to believe that these data came from a normal population?

(b) The P-value for the Shapiro–Wilk test of normality for the data in Exercise 4.4.2(c) is 0.770. Using this value to justify your answer, does it seem reasonable to believe that these data came from a normal population?

(c) Does the P-value in part (b) prove that the data come from a normal population?

4.5 Perspective

The normal distribution is also called the Gaussian distribution, after the German mathematician K. F. Gauss. The term *normal*, with its connotations of "typical" or "usual," can be seriously misleading. Consider, for instance, a medical context, where the primary meaning of "normal" is "not abnormal." Thus, confusingly, the phrase "the normal population of serum cholesterol levels" may refer to cholesterol levels in ideally "healthy" people, or it may refer to a Gaussian distribution such as the one in Example 4.1.1. In fact, for many variables the distribution in the normal (nondiseased) population is decidedly not normal (i.e., not Gaussian).

The examples of this chapter have illustrated one use of the normal distribution—as an approximation to naturally occurring biological distributions. If a natural distribution is well approximated by a normal distribution, then the mean and standard deviation provide a complete description of the distribution: The mean is the center of the distribution: About 68% of the values are within 1 standard deviation of the mean, about 95% are within 2 standard deviations of the mean, and so on.

As noted in Section 2.6, the 68% and 95% benchmarks can roughly be applicable even to distributions that are rather skewed. (But if the distribution is skewed, then the 68% is not symmetrically divided on both sides of the mean, and similarly for the 95%.) However, the benchmarks do not apply to a distribution (even a symmetric one) for which one or both tails are long and thin (see Figures 2.2.13 and 2.2.16).

We will see in later chapters that many classical statistical methods are specifically designed for, and function best with, data that have been sampled from normal populations. We will further see that in many practical situations these methods also work very well for samples from nonnormal populations.

The normal distribution is of central importance in spite of the fact that many, perhaps most, naturally occurring biological distributions could be described better by a skewed curve than by a normal curve. A major use of the normal distribution is not to describe natural distributions, but rather to describe certain theoretical distributions, called sampling distributions, that are used in the statistical analysis of data. We will see in Chapter 5 that many sampling distributions are approximately normal even when the underlying data are not; it is this property that makes the normal distribution so important in the study of statistics.

Supplementary Exercises 4.S.1–4.S.21

4.S.1 The activity of a certain enzyme is measured by counting emissions from a radioactively labeled molecule. For a given tissue specimen, the counts in consecutive 10-second time periods may be regarded (approximately) as repeated independent observations from a normal distribution.[15] Suppose the mean 10-second count for a certain tissue specimen is 1,200 and the standard deviation is 35. Let Y denote the count in a randomly chosen 10-second time period. Find

(a) $\Pr\{Y \geq 1{,}250\}$

(b) $\Pr\{Y \leq 1.175\}$

(c) $\Pr\{1{,}150 \leq Y \leq 1{,}250\}$

(d) $\Pr\{1{,}150 \leq Y \leq 1{,}175\}$

4.S.2 The shell thicknesses of the eggs produced by a large flock of hens follow approximately a normal distribution with mean equal to 0.38 mm and standard deviation equal to 0.03 mm (as in Example 4.1.2). Find the 95th percentile of the thickness distribution.

4.S.3 Refer to the eggshell thickness distribution of Exercise 4.S.2. Suppose an egg is defined as thin shelled if its shell is 0.32 mm thick or less.

(a) What percentage of the eggs are thin shelled?

(b) Suppose a large number of eggs from the flock are randomly packed into boxes of 12 eggs each. What percentage of the boxes will contain at least one thin-shelled egg? (*Hint:* First find the percentage of boxes that will contain no thin-shelled egg.)

4.S.4 The heights of a certain population of corn plants follow a normal distribution with mean 145 cm and standard deviation 22 cm.[16] What percentage of the plant heights are

(a) 100 cm or more?

(b) 120 cm or less?

(c) between 120 and 150 cm?

(d) between 100 and 120 cm?

(e) between 150 and 180 cm?

(f) 180 cm or more?

(g) 150 cm or less?

4.S.5 Suppose four plants are to be chosen at random from the corn plant population of Exercise 4.S.4. Find the probability that none of the four plants will be more than 150 cm tall.

4.S.6 Refer to the corn plant population of Exercise 4.S.4. Find the 90th percentile of the height distribution.

4.S.7 For the corn plant population described in Exercise 4.S.4, find the quartiles and the interquartile range.

4.S.8 Suppose a certain population of observations is normally distributed.

(a) Find the value of z^* such that 95% of the observations in the population are between $-z^*$ and $+z^*$ on the Z scale.

(b) Find the value of z^* such that 99% of the observations in the population are between $-z^*$ and $+z^*$ on the Z scale.

4.S.9 In the nerve-cell activity of a certain individual fly, the time intervals between "spike" discharges follow approximately a normal distribution with mean 15.6 ms and standard deviation 0.4 ms (as in Example 4.1.3). Let Y denote a randomly selected interspike interval. Find

(a) $\Pr\{Y > 15\}$ (b) $\Pr\{Y > 16.5\}$

(c) $\Pr\{15 < Y < 16.5\}$ (d) $\Pr\{15 < Y < 15.5\}$

4.S.10 For the distribution of interspike-time intervals described in Exercise 4.S.9, find the quartiles and the interquartile range.

4.S.11 Among American women aged 20 to 29 years, 10% are less than 60.8 inches tall, 80% are between 60.8 and 67.6 inches tall, and 10% are more than 67.6 inches tall.[17] Assuming that the height distribution can adequately be approximated by a normal curve, find the mean and standard deviation of the distribution.

4.S.12 The intelligence quotient (IQ) score, as measured by the Stanford-Binet IQ test, is normally distributed in a certain population of children. The mean IQ score is 100 points, and the standard deviation is 16 points.[18] What percentage of children in the population have IQ scores

(a) 140 or more? (b) 80 or less?

(c) between 80 and 120? (d) between 80 and 140?

(e) between 120 and 140?

4.S.13 Refer to the IQ distribution of Exercise 4.S.12. Let Y be the IQ score of a child chosen at random from the population. Find $\Pr\{80 \leq Y \leq 140\}$.

4.S.14 Refer to the IQ distribution of Exercise 4.S.12. Suppose five children are to be chosen at random from the population. Find the probability that exactly one of them will have an IQ score of 80 or less and four will have scores higher than 80. (*Hint:* First find the probability that a randomly chosen child will have an IQ score of 80 or less.)

4.S.15 A certain assay for serum alanine aminotransferase (ALT) is rather imprecise. The results of repeated assays of a single specimen follow a normal distribution with mean equal to the true ALT concentration for that specimen and standard deviation equal to 4 U/l (see Example 2.2.12). Suppose that a certain hospital lab measures many specimens every day, performing one assay for each specimen, and that specimens with ALT readings of 40 U/l or more are flagged as "unusually high." If a patient's true ALT concentration is 35 U/l, what is the probability that his specimen will be flagged as "unusually high"?

4.S.16 Resting heart rate was measured for a group of subjects; the subjects then drank 6 ounces of coffee. Ten minutes later their heart rates were measured again. The change in heart rate followed a normal distribution, with a mean increase of 7.3 beats per minute and a standard deviation of 11.1.[19] Let Y denote the change in heart rate for a randomly selected person. Find

(a) $\Pr\{Y > 10\}$ (b) $\Pr\{Y > 20\}$

(c) $\Pr\{5 < Y < 15\}$

4.S.17 Refer to the heart rate distribution of Exercise 4.S.16. The fact that the standard deviation is greater than the average and that the distribution is normal tells us that some of the data values are negative, meaning that the person's heart rate went down, rather than up. Find the probability that a randomly chosen person's heart rate will go down. That is, find $\Pr\{Y < 0\}$.

4.S.18 Refer to the heart rate distribution of Exercise 4.S.16. Suppose we take a random sample of size 400 from this distribution. How many observations do we expect to obtain that fall between 0 and 15?

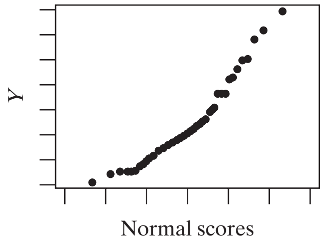
4.S.19 Refer to the heart rate distribution of Exercise 4.S.16. If we use the 1.5 × IQR rule, from Chapter 2, to identify outliers, how large would an observation need to be in order to be labeled an outlier on the upper end?

4.S.20 It is claimed that the heart rates of Exercise 4.S.16 follow a normal distribution. If this is true, which of the following Shapiro–Wilk's test P-values for a random sample of 15 subjects are consistent with this claim?

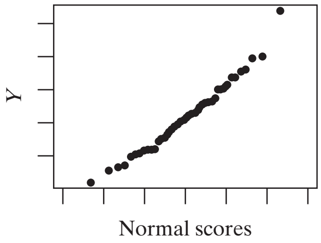
(a) P-value = 0.0149 (b) P-value = 0.1345

(c) P-value = 0.0498 (d) P-value = 0.0042

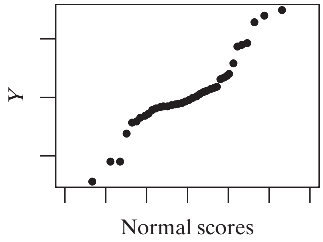
4.S.21 The following four normal probability plots, (a), (b), (c), and (d), were generated from the distributions shown by histograms I, II, and III and another histogram that is not shown. Which normal probability plot goes with which histogram? How do you know? (There will be one normal probability plot that is not used.)



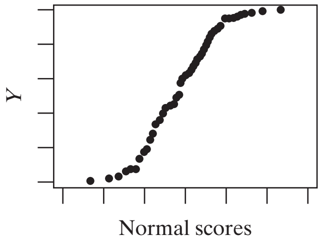
(a)



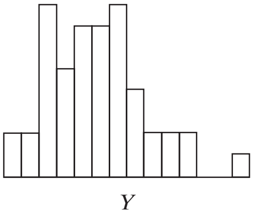
(b)



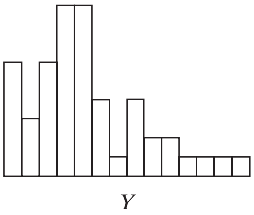
(c)



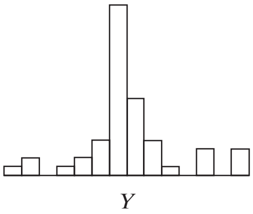
(d)



I



II



III