AN INTRODUCTION TO STATISTICS

WITH

DATA ANALYSIS

by

SHELLEY RASMUSSEN

Department of Mathematical Sciences
Olney 428T
University of Massachusetts/Lowell
Lowell, MA 01854

Originally published by Brooks/Cole Publishing Company,
Division of Wadsworth, Inc.

ISBN 0-534-13578-1

Shelley_Rasmussen@uml.edu

Minitab is a statistical package, a computer program that performs many statistical procedures. The versions of Minitab now available for use on personal computers are menu-driven and much easier to use than the main-frame version originally discussed in this text. Those sections are not included in this online edition of the text. At this time, the most recent version is Minitab 15, available at very reasonable prices for purchase or rental from:

www.e-academy.com/minitab

System Requirements

Processor:	PC with a 1 GHz 32- or 64-bit processor
Memory:	512 MB or more of available RAM
Disk Space:	125 MB free space available
Operating System:	Microsoft Windows 2000, XP, or Vista.
Display:	A display capable of 1024 X 768 or higher resolution
Software:	Adobe Acrobat Reader 5.0 or higher for Meet Minitab

CHAPTER 8

The Gaussian (Normal) Distributions

IN THIS CHAPTER

Gaussian (normal) distributions
Standard Gaussian distribution
Central Limit Theorem

We need to learn about the Gaussian distributions for a couple of reasons. One very good reason is that much of classical statistical analysis (including many techniques developed early in this century) depends on the assumption that we have independent observations from a Gaussian distribution. For instance, in Section 10-3 we will discuss the t test, a tool for making inferences about a population mean. For the probability statements associated with a t test to be valid, we must assume that we have independent observations from a Gaussian distribution. The two-sample t test, used for comparing two population means, is the subject of Section 11-3. For valid probability statements using a two-sample t test, we must assume that we have two independent random samples from Gaussian distributions. In Chapters 12 and 13 we discuss experiments that might be analyzed using analysis of variance techniques, and in Chapter 15 we discuss fitting linear regression models; all of these classical statistical techniques depend on the assumption that observations come from a Gaussian distribution. In Sections 8-1 and 8-2, we will discuss what it means for observations to come from a Gaussian distribution.

Another reason to learn about Gaussian distributions is that a number of large-sample procedures allow us to make inferences based on a particular Gaussian distribution, the *standard Gaussian distribution,* even if the observations themselves do not come from a Gaussian distribution. For instance, in Section 10-1 we discuss large-sample inferences about a population mean. With a large enough sample size (plus a random sample), these inferences can be based on the standard Gaussian distribution even if the observations do not come from a Gaussian distribution. Large-sample inference about a proportion is the subject of Section 10-2. Again, if the (random) sample size is large enough, we can base our inferences on the standard Gaussian distribution. In Section 11-1 we consider large-sample inferences about two population means and in Section 11-2, large-sample inference about two proportions. With large enough sample sizes (plus independent random samples), inferences can be based on the standard Gaussian distribution. Justification for these large-sample inferences based on the standard Gaussian distribution comes from the Central Limit Theorem, a result we discuss in Section 8-3.

If a random variable has a Gaussian probability distribution, then that variable represents a continuous-type observation or measurement such as height, weight, thickness, strength, temperature, or pressure. How can we find probabilities associated with such a random variable? We need this knowledge because, as mentioned above, many probability statements in statistical inference are based on a Gaussian distribution.

Also mentioned above, much of classical statistical inference depends on the assumption that a sample of observations comes from a Gaussian distribution. If we collect a sample of independent observations from such a distribution, what will a plot of the sample values look like? To address this question, we might use a computer program to simulate a sample of independent observations from a Gaussian distribution. (The RANDOM command in Minitab will accomplish this, as described in the appendix to this chapter.)

As an illustration, consider a particular Gaussian distribution, with mean

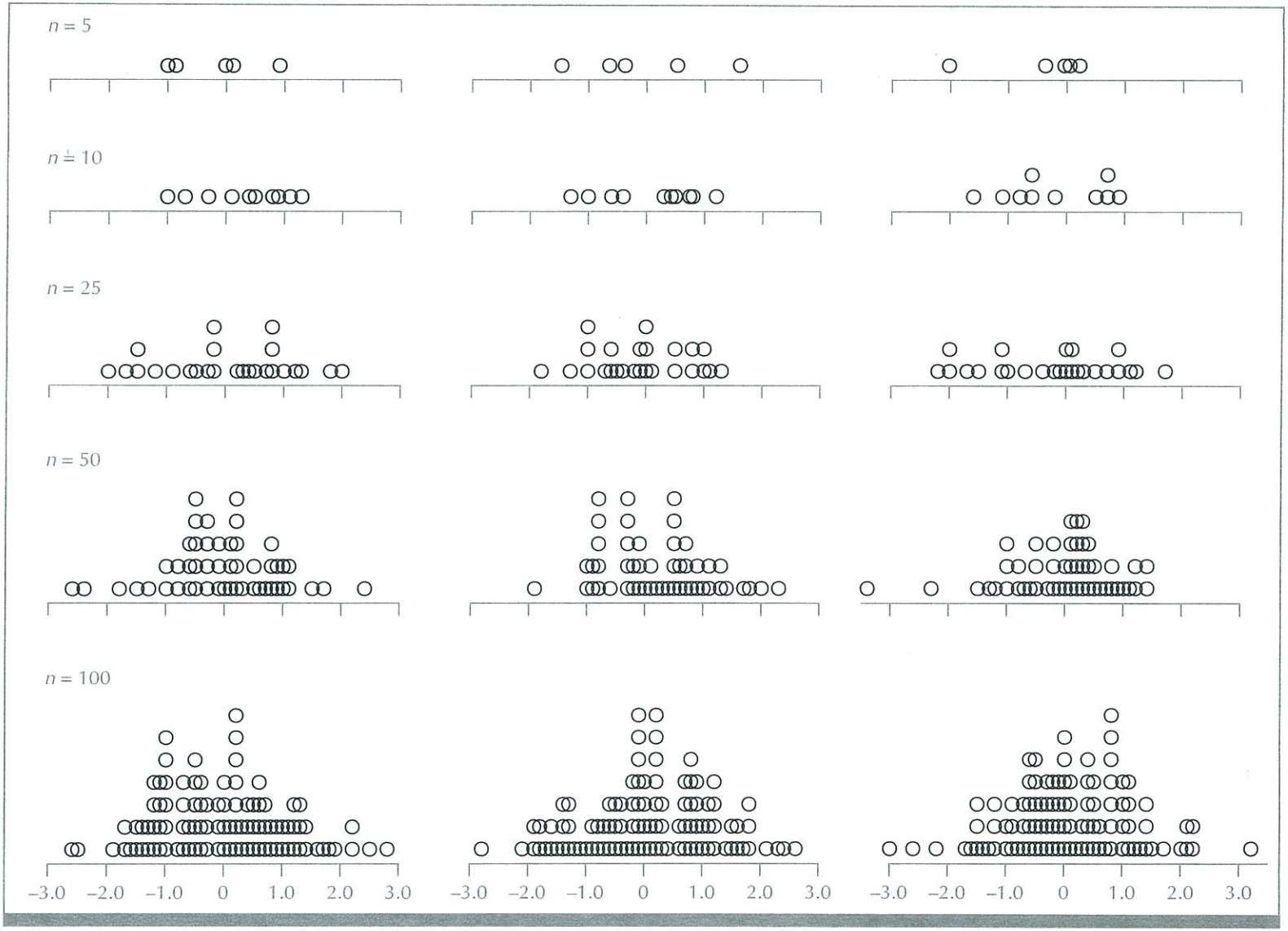

FIGURE 8-1 Dot plots of simulated samples of independent observations from the standard Gaussian distribution (obtained using the RANDOM command in Minitab). Three samples were simulated for each of these sample sizes: 5, 10, 25, 50, and 100.

0 and standard deviation 1, called the *standard Gaussian distribution.* Figure 8-1 shows plots of 15 different simulated samples of independent observations from the standard Gaussian distribution. We can think of each sample as a set of independent observations of a random variable having a standard Gaussian distribution. The first row in the figure shows dot plots of values from three samples, each of size 5. Rows 2–5 show dot plots for three samples of, respectively, size 10, 25, 50, and 100.

Looking at the plots in Figure 8-1, you might guess the standard Gaussian distribution is a unimodal distribution that is symmetric about 0. If you guessed this, you would be right. Notice the differences between plots of different samples from the same distribution. The standard Gaussian distribution has mean 0, but the *sample* means are not all 0. Just as there is variation in the sample values, there is variation among the sample means. Similarly, the standard deviation of the standard Gaussian distribution is 1, but the sample standard deviations do not all equal 1.

We expect variation among different sets of observations of a random variable. We cannot predict exactly what values of a random variable we will observe. However, we can use the distribution of the random variable to make *probability statements* about one or more observed values of the variable. In Section 8-1, we discuss how to make probability statements regarding a single observation of a random variable that has a Gaussian distribution. These ideas can then be extended to making probability statements about several independent observations from a Gaussian distribution.

In Section 8-2, we check to see if the distribution of a real data set can be approximated by a Gaussian distribution. Such a check in an experimental situation can help us decide if it is reasonable to use a statistical procedure that depends on assuming Gaussian observations. Then in Section 8-3, we discuss the Central Limit Theorem. This theorem and related results are useful in statistical inference based on large samples.

Before discussing the Gaussian distributions, a comment on their name is in order. The common term for these distributions is the *normal distributions.* The traditional designation "normal" is unfortunate because it helps to perpetuate the extremely common and *incorrect* notion that measurements "normally" follow a normal distribution. Many sets of measurements do not follow a Gaussian distribution, as we will see. To avoid incorrect interpretations, we will refer to these distributions as the Gaussian distributions, after the great mathematician Karl Friedrich Gauss, who described them.

8-1 The Gaussian Distributions

If a random variable Y has a Gaussian distribution, then Y can in theory take any real number value. How can we find the probability that a Gaussian random variable Y is in an interval of numbers? For instance, what is the probability that Y has a value from 10 to 25, $P(10 \leq Y \leq 25)$? Or, what is the probability that Y is greater than 50, $P(Y > 50)$? Before addressing such questions

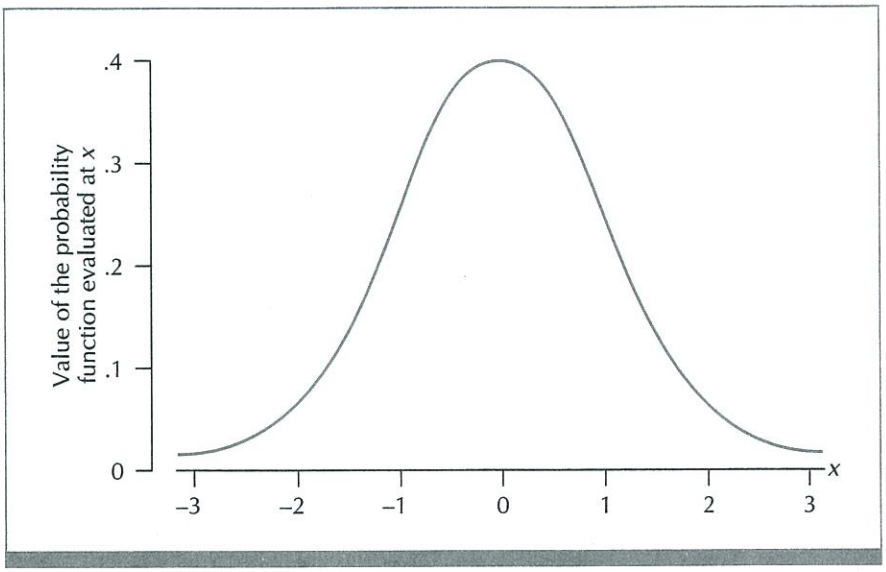

FIGURE 8-2 Graph of the standard Gaussian probability function. The area between the curve and the horizontal axis equals 1.

for a general Gaussian distribution, we will consider a special case, the standard Gaussian distribution.

The standard Gaussian distribution is defined in terms of the curve graphed in Figure 8-2, the standard Gaussian probability function. The area under this curve (between the curve and the horizontal axis) equals 1.

Let Z denote a random variable having the standard Gaussian distribution. The probability that Z takes a value from c to d, $P(c \leq Z \leq d)$, is the area under the standard Gaussian probability function between the numbers c and d. Such an area is shaded in Figure 8-3a.

We might want to find the probability that Z is greater than or equal to c, $P(Z \geq c)$. Then we find the area under the standard Gaussian probability function to the right of c. The shaded portion of Figure 8-3b shows such an area.

For the probability $P(Z \leq c)$ that Z is less than or equal to c, we find the area under the standard Gaussian probability function to the left of c. Such an area is shaded in Figure 8-3c.

Now suppose we want to find the probability that Z equals a number c, $P(Z = c)$. We find the area under the standard Gaussian probability function at the number c, as illustrated in Figure 8-3d. But this is the area of a line segment, a rectangle with width 0. Therefore, we say $P(Z = c)$ equals 0. [This is true for all continuous random variables. If X is a continuous random variable, then $P(X = c) = 0$ for any number c.] From this, we see that $P(Z \geq c) = P(Z > c)$ and $P(c \leq Z \leq d) = P(c < Z < d)$, and so on.

The **standard Gaussian distribution** is a continuous probability distribution. The probability a standard Gaussian random variable is in an inter-

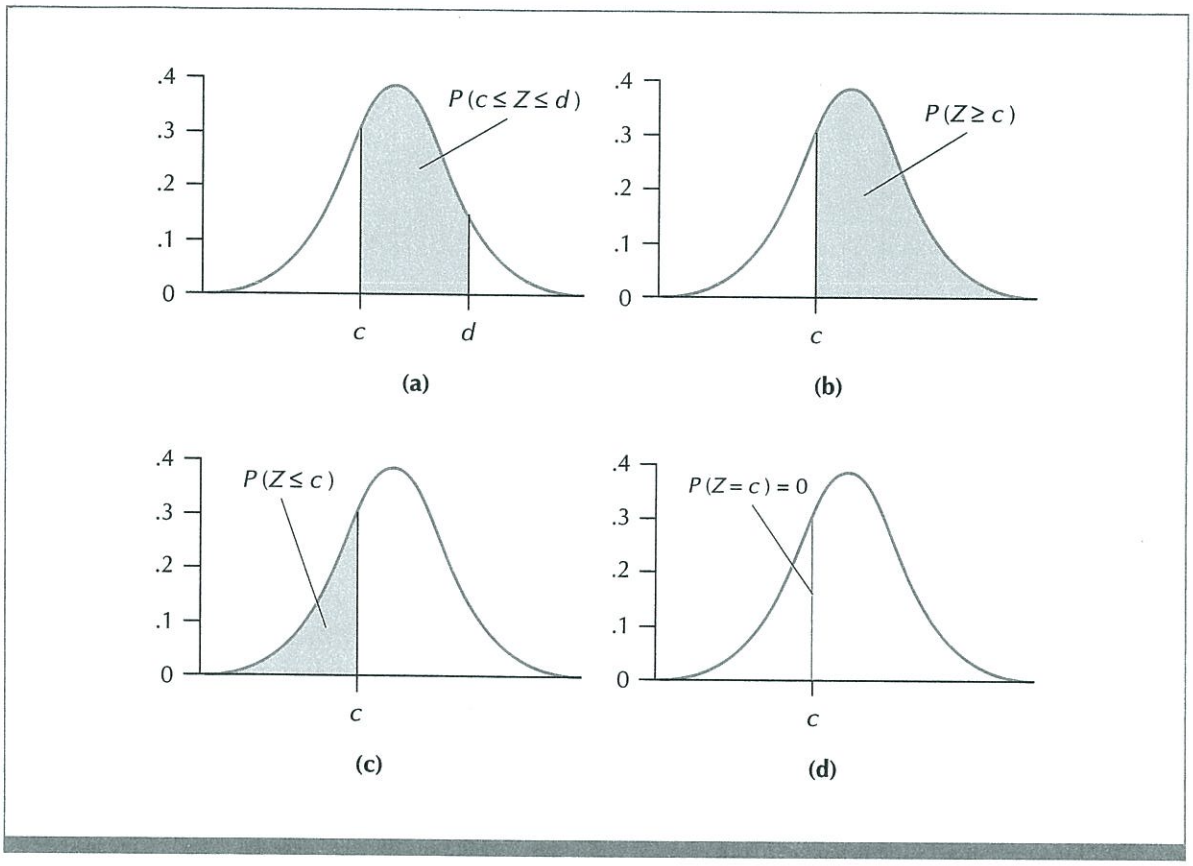

FIGURE 8-3 Let Z denote a random variable having the standard Gaussian probability distribution.
(a) The area under the graph between c and d equals $P(c \leq Z \leq d)$.
(b) The area under the curve to the right of c equals $P(Z \geq c)$.
(c) The area under the curve to the left of c equals $P(Z \leq c)$.
(d) Because Z is a continuous random variable, $P(Z = c)$ equals 0 for any number c.

val of numbers equals the area over that interval under the graph of the standard Gaussian probability function, illustrated in Figure 8-2.

Table B at the back of the book lists cumulative probabilities of the form $P(Z \leq c)$ for selected nonnegative values of c. For instance, $P(Z \leq 1) = .8413$ and $P(Z \leq 2.51) = .9940$. We can find a tail probability (so called because it represents the relatively small probability for a *tail* of the distribution) such as $P(Z \geq 2.51)$ by subtraction: $P(Z \geq 2.51) = 1 - P(Z \leq 2.51) = .0060$.

A **cumulative probability** has the form $P(X \leq c)$ where X is a random variable and c is a constant.

A **tail probability** for a random variable X is a probability that is small (less than .5) and has the form $P(X \geq c)$ or $P(X \leq c)$ for some number c.

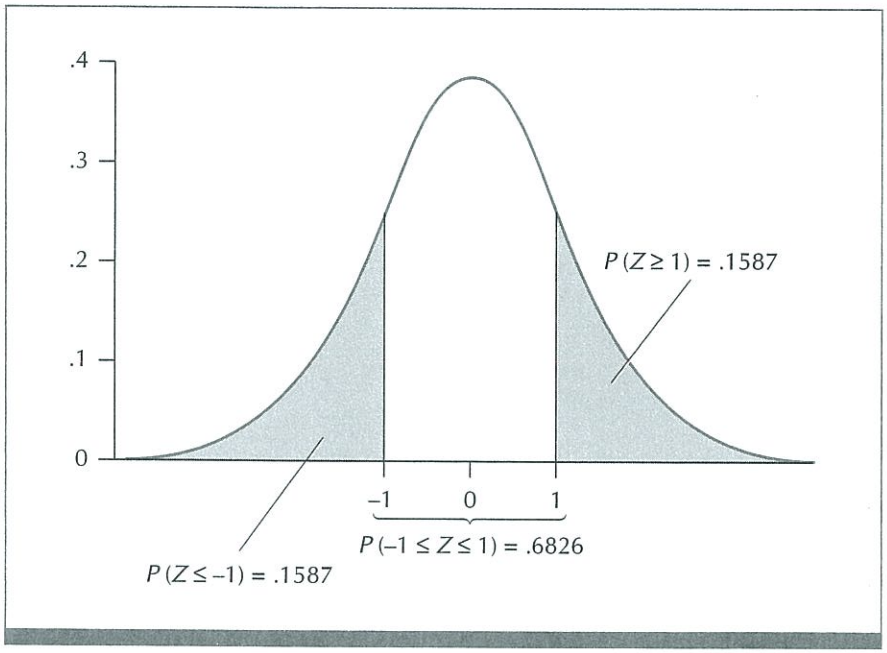

FIGURE 8-4 The standard Gaussian probability function is symmetric about 0. If Z has a standard Gaussian distribution, then $P(Z \leq -1) = P(Z \geq 1) = .1587$.

We see from Figure 8-2 that the standard Gaussian probability function is symmetric about 0. So, for any positive number c, $P(Z \geq c)$ equals $P(Z \leq -c)$. For example, Figure 8-4 illustrates the fact that $P(Z \leq -1) = P(Z \geq 1) = 1 - P(Z \leq 1) = .1587$. With this information, we can calculate $P(-1 \leq Z \leq 1)$:

$$\begin{aligned} P(-1 \leq Z \leq 1) &= 1 - P(Z \leq -1) - P(Z \geq 1) \\ &= 1 - .1587 - .1587 = .6826 \end{aligned}$$

A picture helps us visualize the probability we want to calculate. For instance, suppose we want to find $P(-2 \leq Z \leq -1)$. The corresponding area is shaded in Figure 8-5. Because the standard Gaussian curve is symmetric about 0, $P(-2 \leq Z \leq -1) = P(1 \leq Z \leq 2)$. Also, $P(1 \leq Z \leq 2) = P(Z \leq 2) - P(Z < 1)$. We can read these last two probabilities from Table B. Therefore,

$$\begin{aligned} P(-2 \leq Z \leq -1) = P(1 \leq Z \leq 2) &= P(Z \leq 2) - P(Z < 1) \\ &= .9772 - .8413 = .1359 \end{aligned}$$

Suppose we want to find $P(-1 \leq Z \leq 2)$. From Figure 8-6, we see that the probability we want is $P(Z \leq 2) - P(Z < -1)$. Using Table B, we find that $P(Z \leq 2) = .9772$ and $P(Z < -1) = P(Z > 1) = 1 - P(Z \leq 1) = .1587$. Therefore,

$$P(-1 \leq Z \leq 2) = .9772 - .1587 = .8185$$

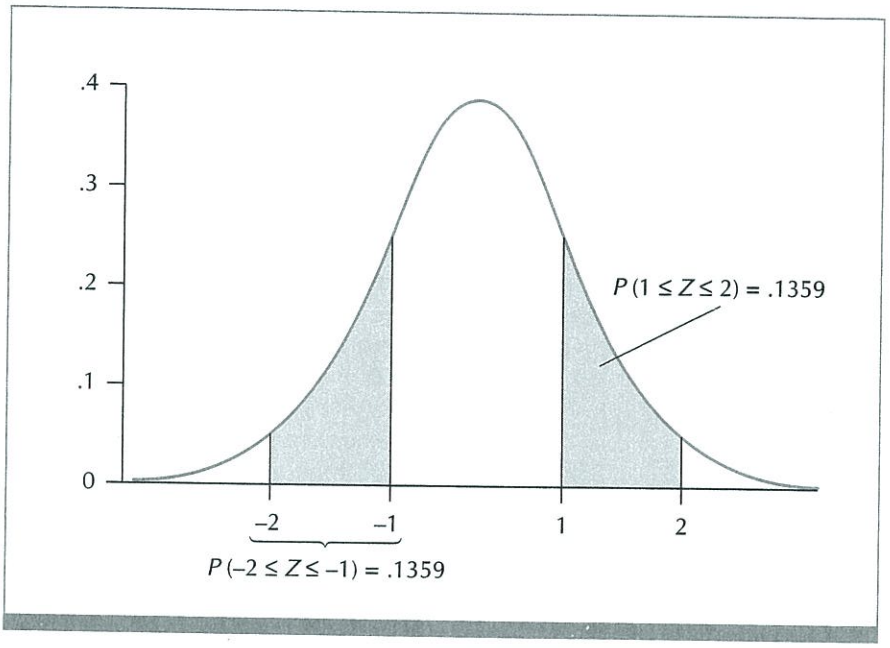

FIGURE 8-5 If Z has the standard Gaussian distribution, then $P(-2 \le Z \le -1) = P(1 \le Z \le 2) = P(Z \le 2) - P(Z < 1) = .9772 - .8413 = .1359$.

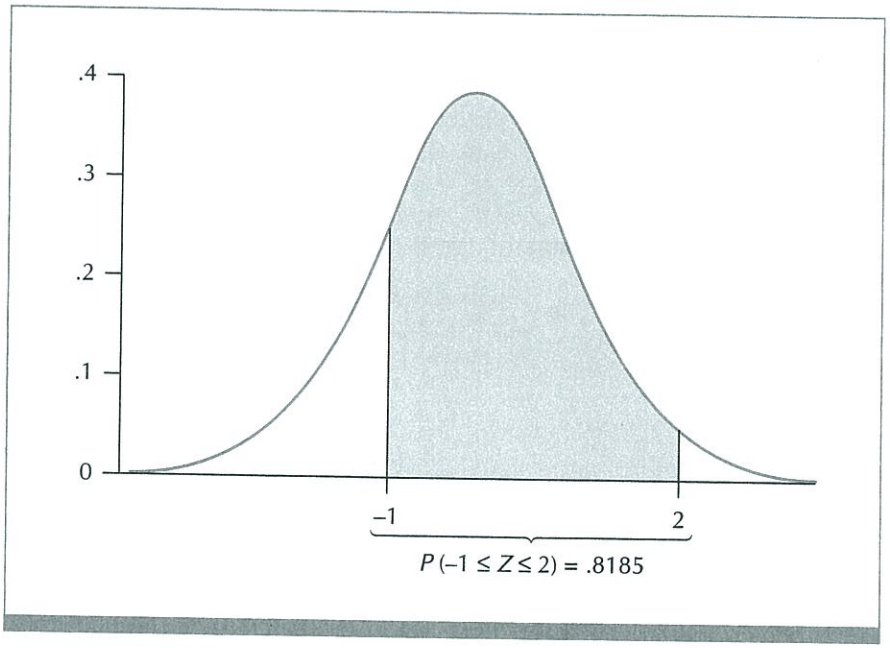

FIGURE 8-6 If Z has the standard Gaussian distribution, then $P(-1 \le Z \le 2) = P(Z \le 2) - P(Z < -1) = .9772 - .1587 = .8185$.

Again, let Z denote a random variable having the standard Gaussian distribution. Because this distribution is symmetric about 0, the mean or expected value of Z is 0. The variance of Z equals 1. We summarize this information as follows:

If Z is a standard Gaussian random variable,
then $E(Z) = 0$ and $\text{Var}(Z) = 1$.

Thus, as noted in the introduction to this chapter, the standard Gaussian distribution is a Gaussian distribution with mean 0 and variance 1. Now let σ denote a positive number and let μ denote any number. If $X = \sigma Z$, then X has a Gaussian distribution with mean 0 and variance σ^2, standard deviation σ. If $Y = X + \mu = \sigma Z + \mu$, then Y has a Gaussian distribution with mean μ and variance σ^2. That is, $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$.

Suppose the random variable Y has a Gaussian distribution with mean μ and variance σ^2. We want to find the probability that Y takes a value between two numbers c and d, $P(c \le Y \le d)$. To find this probability, we standardize the random variable Y.

Suppose a random variable X has mean μ and variance σ^2 (or standard deviation σ). We **standardize X** by subtracting the mean μ and then dividing by the standard deviation σ. The **standardized random variable**

$$\frac{X - \mu}{\sigma}$$

has mean 0 and standard deviation 1.

When we standardize a random variable, we are converting the variable into standard units. A value of the standardized variable shows the distance, in standard deviations, that the corresponding value of the original variable lies above or below its mean.

If the random variable Y has a Gaussian distribution with mean μ and variance σ^2, then the standardized variable $Z = (Y - \mu)/\sigma$ has the standard Gaussian distribution, with mean 0 and variance 1.

A **Gaussian distribution** is a continuous probability distribution. If a random variable Y has a Gaussian distribution with mean μ and variance σ^2 (or standard deviation σ), then

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

has the standard Gaussian distribution.

If we write Y as $Y = \sigma Z + \mu$, then we see that

$$P(c \le Y \le d) = P(c \le \sigma Z + \mu \le d) = P\left(\frac{c - \mu}{\sigma} \le Z \le \frac{d - \mu}{\sigma}\right)$$

Therefore, we can use a table of probabilities for the standard Gaussian distribution, such as Table B, to find $P(c \le Y \le d)$. That is, we can use a table of

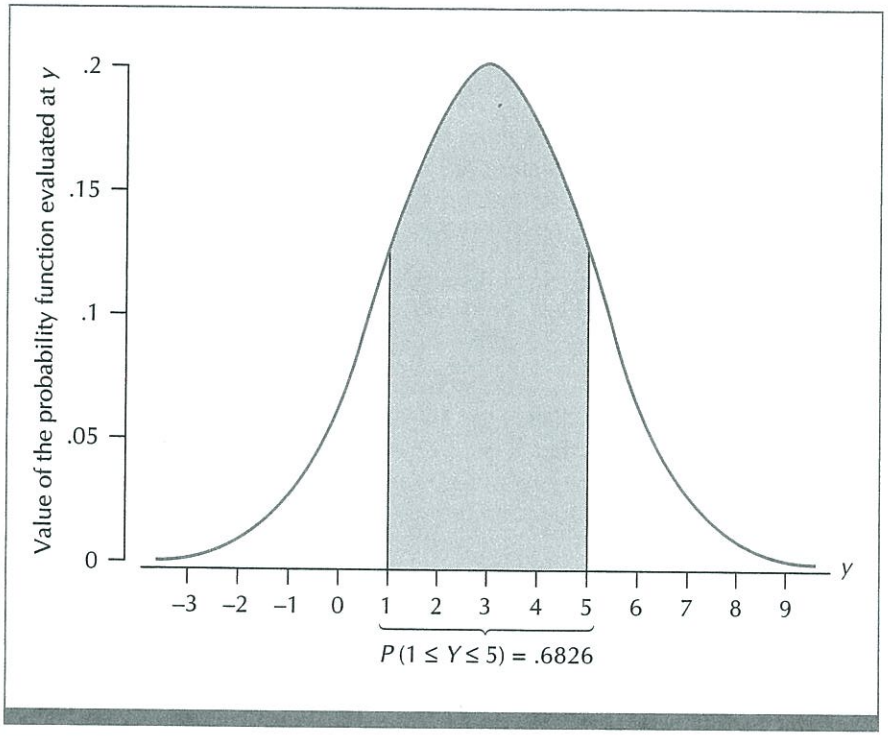

FIGURE 8-7 Graph of the probability function of a random variable Y having the Gaussian distribution with mean 3 and variance 4. The shaded area corresponds to $P(1 \leq Y \leq 5)$.

probabilities for the standard Gaussian distribution to make probability statements about any Gaussian random variable.

Suppose, for example, that Y has a Gaussian distribution with mean 3 and variance 4. Let's find the probability that Y takes a value from 1 to 5, $P(1 \leq Y \leq 5)$. The corresponding area is shaded in Figure 8-7. Let Z denote a standard Gaussian random variable. Then

$$P(1 \leq Y \leq 5) = P\left(\frac{1-3}{\sqrt{4}} \leq \frac{Y-3}{\sqrt{4}} \leq \frac{5-3}{\sqrt{4}}\right)$$
$$= P(-1 \leq Z \leq 1) = .6826$$

We can use the ideas in this section to decide whether the distribution of a set of values can be approximated by a Gaussian distribution. In Section 8-2, we evaluate whether the distribution of a set of baseball statistics can be approximated by a Gaussian distribution. In an experimental situation, we use such an evaluation to decide if a statistical procedure based on the assumption of Gaussian observations is justified.

8-2

Approximating a Distribution of Values by a Gaussian Distribution

What does it mean to say that a set of values approximately follows a Gaussian distribution? It means that the proportion of values between any two numbers c and d approximately equals the area between c and d under that Gaussian probability function.

> A **distribution of data values is approximately Gaussian** if the proportion of values in any interval approximately equals the area over that interval under the appropriate Gaussian curve.

Let's consider an example. Table 8-1 shows the number of hits allowed during the 1987 season for each of the 26 major-league baseball teams. We see that $\frac{15}{26} = .577$ is the proportion of teams with 1,400 to 1,500 hits allowed. Is there a Gaussian distribution with an area about .577 under the probability function between 1,400 and 1,500? Let's see if we can find one.

A Gaussian distribution is symmetric about its mean. So, for a Gaussian approximation to work, the data values should be symmetrically distributed. Figure 8-8 shows a dot plot and box plot of the 26 values of hits allowed. The plots look reasonably symmetric, so we will continue.

We must specify the mean and standard deviation of the Gaussian distribution we want to use. It is reasonable to use the mean and standard deviation of the data values. The mean number of hits allowed per team in 1987 was 1,457.5 and the standard deviation was 66.7. Can we approximate the distribu-

TABLE 8-1 Hits allowed during the 1987 season by the 26 major-league baseball teams (*The Sporting News,* October 12, 1987, pages 37, 42; *USA Today,* October 6, 1987, page 4c; *USA Today,* October 7, 1987, page 5c; kindly provided by Lee Panas)

Team	Hits allowed	Team	Hits allowed
Blue Jays	1,323	Phillies	1,453
Astros	1,363	Twins	1,465
Pirates	1,377	Yankees	1,475
Rangers	1,388	Angels	1,481
Padres	1,402	Cardinals	1,484
Giants	1,407	Reds	1,486
Mets	1,407	Mariners	1,503
Dodgers	1,415	Cubs	1,524
Royals	1,424	Braves	1,529
Expos	1,428	Brewers	1,548
Tigers	1,430	Orioles	1,555
White Sox	1,436	Indians	1,566
Athletics	1,442	Red Sox	1,584

Mean = 1,457.5 Standard deviation = 66.7

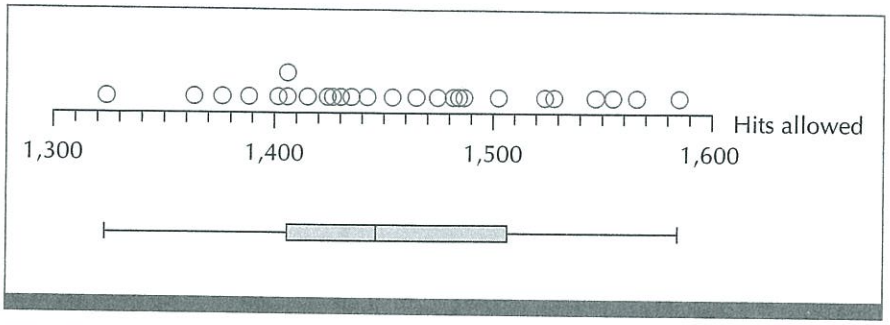

FIGURE 8-8 Dot plot and box plot of the number of hits allowed in 1987 by the 26 major-league baseball teams

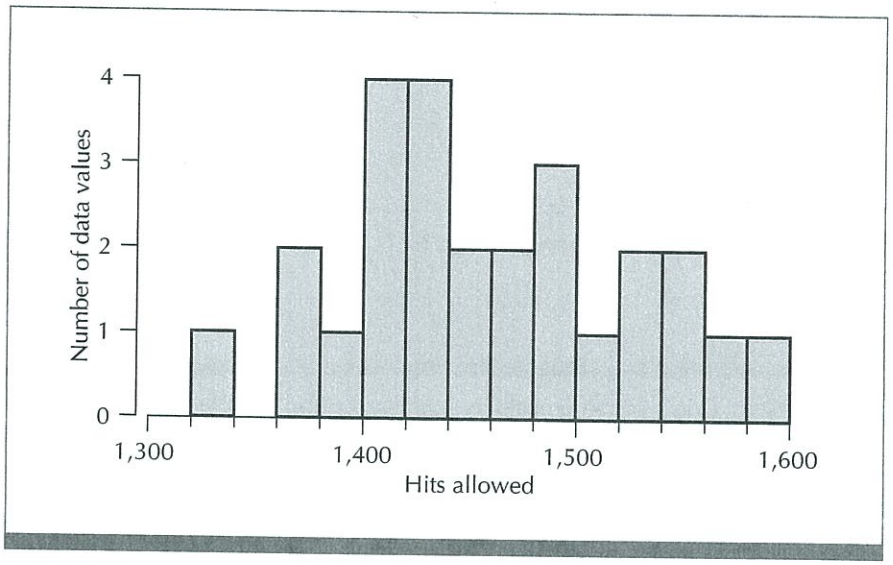

FIGURE 8-9 Histogram of number of hits allowed by major-league baseball teams in 1987

tion of 1987 hits allowed by the Gaussian distribution with mean 1,457.5 and standard deviation 66.7?

A histogram of hits allowed is shown in Figure 8-9. The proportion of values between two interval endpoints is the proportion of the total area of the histogram between those two endpoints. For instance, the proportion of teams with 1,420 to 1,480 hits allowed is $\frac{8}{26} = .308$. The area between 1,420 and 1,480 is $\frac{8}{26}$ or roughly 31% of the total area of the histogram.

The probability function for the Gaussian distribution with mean 1,457.5 and standard deviation 66.7 is graphed in Figure 8-10. Does the area under this curve between two points approximate the proportion of the area of the histo-

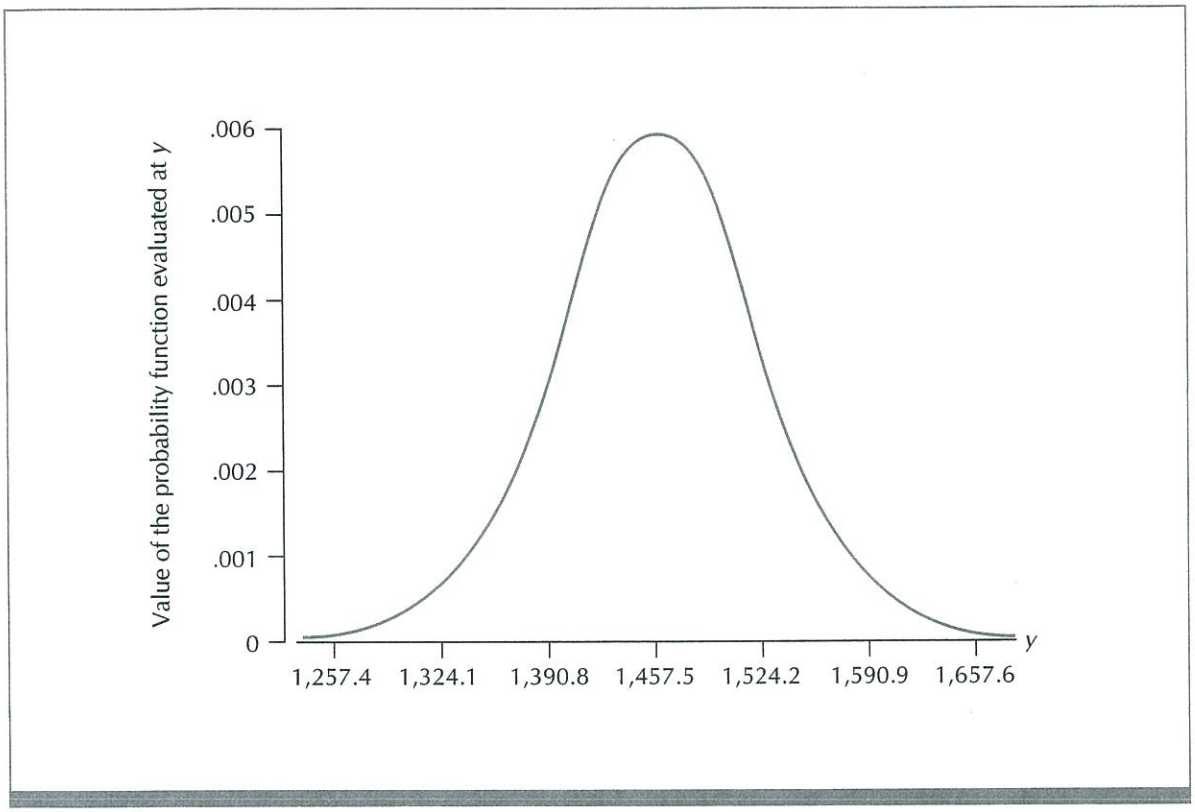

FIGURE 8-10 Graph of the probability function for the Gaussian distribution with mean 1,457.5 and standard deviation 66.7

gram between the two points? Let's find, for instance, the area under the curve between 1,420 and 1,480.

Let Y denote a random variable having the Gaussian distribution with mean 1,457.5 and standard deviation 66.7. Let Z denote a standard Gaussian random variable. The area under the curve in Figure 8-10 between 1,420 and 1,480 equals $P(1{,}420 \leq Y \leq 1{,}480)$. Then

$$
\begin{aligned}
&P(1{,}420 \leq Y \leq 1{,}480) \\
&\qquad = P\left(\frac{1{,}420 - 1{,}457.5}{66.7} \leq \frac{Y - 1{,}457.5}{66.7} \leq \frac{1{,}480 - 1{,}457.5}{66.7}\right) \\
&\qquad = P(-.56 \leq Z \leq .34) \\
&\qquad = P(Z \leq .34) - P(Z \leq -.56) \\
&\qquad = .6331 - .2877 = .3454
\end{aligned}
$$

The area under the curve in Figure 8-10 between 1,420 and 1,480 is .3454. This is not too far from the observed proportion .308 of teams with 1,420 to 1,480

TABLE 8-2 Comparison of observed frequencies with those expected based on the Gaussian distribution with mean 1,457.5 and standard deviation 66.7, for number of hits allowed in 1987

Interval of values for a Gaussian distribution with mean 1,457.5 and standard deviation 66.7	Interval of values for a standard Gaussian distribution	Area under the curve over this interval	Expected frequency in this interval	Observed frequency in this interval
1,457.5 to 1,490.9	0 to .5	.1915	.1915 × 26 = 4.98	5
1,424.1 to 1,457.5	−.5 to 0	.1915	.1915 × 26 = 4.98	5
1,390.8 to 1,524.2	−1 to 1	.6826	.6826 × 26 = 17.75	17
Greater than 1,524.2	Greater than 1	.1587	.1587 × 26 = 4.13	5
Less than 1,390.8	Less than −1	.1587	.1587 × 26 = 4.13	4
1,324.1 to 1,590.9	−2 to 2	.9544	.9544 × 26 = 24.81	25
Greater than 1,590.9	Greater than 2	.0228	.0228 × 26 = .59	0
Less than 1,324.1	Less than −2	.0228	.0228 × 26 = .59	1
1,257.3 to 1,657.7	−3 to 3	.9974	.9974 × 26 = 25.93	26
Greater than 1,657.7	Greater than 3	.0013	.0013 × 26 = .03	0
Less than 1,257.3	Less than −3	.0013	.0013 × 26 = .03	0
1,400 to 1,500	−.86 to .64	.5440	.5440 × 26 = 14.14	15

hits allowed. Using this Gaussian distribution, we would expect .3454 × 26 = 8.98 of the 26 teams to have 1,420 to 1,480 hits allowed. This agrees pretty well with the observed frequency of 8.

Table 8-2 summarizes similar calculations. An entry in column 1 shows an interval of values for hits allowed. Column 2 shows the same interval standardized by subtracting the sample mean of 1,457.5 and then dividing by the sample standard deviation of 66.7. Column 3 shows the area under the standard Gaussian curve over the interval shown in column 2; column 3 is the proportion expected in the interval in column 1 using this Gaussian distribution. This proportion times 26 gives the expected number of data values in the interval, listed in column 4. For comparison, column 5 gives the observed number of values of hits allowed in the interval.

Consider, for example, the interval that extends from the sample mean to half a standard deviation above the sample mean: from 1,457.5 to 1,457.5 + .5 × 66.7, or from 1,457.5 to 1,490.9. This interval is the first entry in Table 8-2. If we standardize by subtracting 1,457.5 and then dividing by 66.7, this interval becomes 0 to .5. Using Table B at the back of the book, we see that the probability that a standard Gaussian random variable Z is between 0 and .5 is

$$P(0 \leq Z \leq .5) = P(Z \leq .5) - P(Z \leq 0) = .6915 - .5000 = .1915$$

If the Gaussian approximation is reasonable, we expect about 19% of the data values in the interval from 1,457.5 to 1,490.9. Nineteen percent of 26 is about 5, and 5 is in fact the observed number of values of hits allowed in this interval. For each of the intervals in Table 8-2, we see that there is close agreement between the observed frequencies and those expected based on the Gaussian approximation.

TABLE 8-3 Comparison of observed proportions with those expected based on the Gaussian distribution with mean 1,457.5 and standard deviation 66.7, for number of hits allowed by major-league baseball teams in 1987

Interval of values	Proportion expected by Gaussian approximation	Proportion observed
1,457.5 to 1,490.9	.1915	$\frac{5}{26} = .19$
1,424.1 to 1,457.5	.1915	$\frac{5}{26} = .19$
1,390.8 to 1,524.2	.6826	$\frac{17}{26} = .65$
Greater than 1,524.2	.1587	$\frac{5}{26} = .19$
Less than 1,390.8	.1587	$\frac{4}{26} = .15$
1,324.1 to 1,590.9	.9544	$\frac{25}{26} = .96$
Greater than 1,590.9	.0228	$\frac{0}{26} = 0$
Less than 1,324.1	.0228	$\frac{1}{26} = .04$
1,257.3 to 1,657.7	.9974	$\frac{26}{26} = 1$
Greater than 1,657.7	.0013	$\frac{0}{26} = 0$
Less than 1,257.3	.0013	$\frac{0}{26} = 0$
1,400 to 1,500	.5440	$\frac{15}{26} = .58$

Column 1 of Table 8-3 lists the intervals shown in the first column of Table 8-2. For each interval, column 2 shows the proportion of data values expected in the interval based on the Gaussian approximation (column 3 of Table 8-2). Column 3 shows the proportion of data values observed in the interval (column 5 of Table 8-2, divided by 26). We have checked only a few intervals, but the similarity between these observed and expected proportions suggests that the Gaussian distribution with mean 1,457.5 and standard deviation 66.7 gives a reasonable approximation to the distribution of hits allowed by major-league baseball teams in 1987.

Standardized Data Values

We indicated previously that a standardized random variable shows standard deviations above or below the mean for a random variable. Similarly, a *standardized data value* shows how many standard deviations above or below the mean a data value is. We standardize a data value by subtracting the mean of the values and then dividing by the standard deviation:

> We obtain a **standardized data value** by subtracting the mean of the set of data values and dividing by the standard deviation.

For instance, in 1987 the Rangers allowed 1,388 hits. When we subtract the sample mean and divide by the sample standard deviation, we get a standardized value of

$$\frac{1{,}388 - 1{,}457.5}{66.7} = -1.04$$

The Rangers were about one standard deviation below the mean number of hits allowed among the major-league teams. The Orioles allowed 1,555 hits in 1987. Subtracting the sample mean and dividing by the sample standard deviation, we get this standardized value:

$$\frac{1{,}555 - 1{,}457.5}{66.7} = 1.46$$

The Orioles were about one and a half standard deviations above the mean number of hits allowed among the major-league teams in 1987.

Referring to Table 8-2, we find that 17 of the 26 values of hits allowed were within one standard deviation of the mean, 5 were more than one standard deviation above the mean, and 4 were more than one standard deviation below the mean. We can interpret all of the intervals in column 1 of Table 8-2 in terms of standard deviations above or below the sample mean. This is the meaning of the standardized intervals in column 2 of the table. (For more discussion of standard units and Gaussian approximations, see Freedman, Pisani, and Purves, 1978, Chapter 5.)

More on Gaussian Approximations

The assumption that we can approximate the distribution of a set of data values by a Gaussian distribution is the basis for many techniques in classical statistical inference. However, many distributions *cannot* be approximated by a Gaussian distribution. Consider, for instance, the data displayed in Figure 8-11. (We saw this plot in Chapter 2.) The figure shows a dot plot of number enrolled in higher education in 1984 as percentage of 20–24-year-old age group, for 119

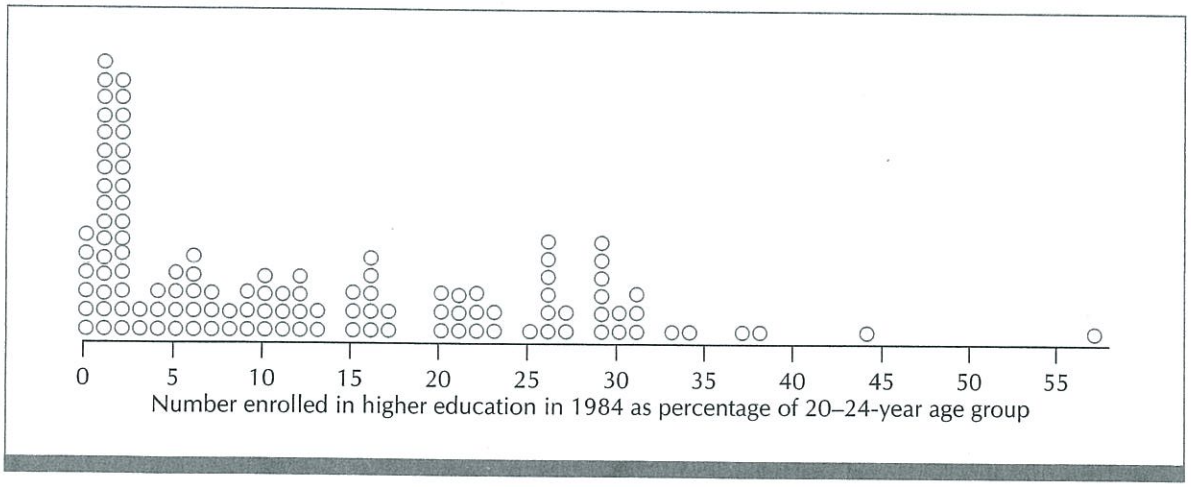

FIGURE 8-11 Dot plot of number enrolled in higher education in 1984 as percentage of 20–24-year-old age group for 119 countries (*World Development Report 1987*, pages 262, 263)

countries. The plot is skewed to the right, not at all symmetrical. The distribution of the 119 data values cannot be well approximated by a Gaussian distribution (see Exercise 8-16).

If the proportion of data values in any interval approximately equals the area over that interval under a Gaussian curve, then we say we can approximate the distribution of data values using that Gaussian distribution. Similarly, if the probability that a random variable is in any interval approximately equals the area over that interval under a Gaussian curve, then we say the random variable has approximately that Gaussian distribution.

> The **distribution of a random variable is approximately Gaussian** if the probability that the random variable is in any interval approximately equals the area over that interval under a Gaussian curve.

In Section 8-3 we discuss a famous result known as the Central Limit Theorem. This result states that if we have a large number of independent observations from the same distribution, then the average of these observations has approximately a Gaussian distribution. This is very useful in large-sample statistical analysis, when we can base our inferences on a Gaussian distribution.

8-3 The Central Limit Theorem

We say random variables X_1, X_2 through X_n are *independent* if they represent independent numerical observations made during an experiment. Observations are *independent* if the result of any one observation does not in any way affect the results for the other observations.

Let X_1, X_2 through X_n denote independent random variables with the same probability distribution. Then we say X_1 through X_n represent a *random sample in the probability sense.*

> A **random sample in the probability sense** is a collection of independent random variables with the same probability distribution.

For our purposes, a random sample in the probability sense represents numerical observations on the same variable made during independent and identical repetitions of an experiment. For example, suppose we have a group of mice with the same genetic background, each with a tumor of the same type and size. We house, feed, and care for the animals in the same way and treat them all with the same antitumor regimen. At the end of the treatment period, we plan to measure the tumor size on each mouse. Under these experimental conditions, we can think of the collection of tumor measurements as a random sample in the probability sense.

Suppose the random variables X_1 through X_n represent a random sample in the probability sense. Let μ denote the mean and σ^2 the variance of each X_i. Let Y denote the sum and $\bar{X} = Y/n$ the average of X_1 through X_n. The *Central Limit Theorem* states that if the sample size n is large enough, then Y and $\bar{X}$

are each approximately Gaussian distributed. In particular, the Central Limit Theorem tells us that the standardized version of $\bar{X}$ (or Y):

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{Y - n\mu}{\sqrt{n}\sigma}$$

has approximately the standard Gaussian distribution if the sample size n is large enough.

> **A version of the Central Limit Theorem:** Suppose X_1 through X_n are independent random variables with the same probability distribution. Let μ denote the mean and σ the standard deviation of each X_i. Let $\bar{X}$ denote the average of X_1 through X_n. If the sample size n is large enough, the distribution of
>
> $$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
>
> is approximately the standard Gaussian distribution. The larger the sample size, the better the approximation.

How large does the sample size n have to be for the Gaussian approximation to apply to the distribution of $\bar{X}$ or Y? The answer depends on how different the probability distribution of the X_i's is from a Gaussian distribution. Let's illustrate this idea.

Suppose we toss a fair coin many times. The tosses are independent of one another in that the result of one toss does not affect in any way the result of another. We record a 1 for each head and a 0 for each tail. Let X_1 denote the result of the first toss, X_2 the result of the second toss, and so on. Then the X_i's are independent random variables, each with the same probability distribution defined by the two probabilities:

$$P(X_i = 0) = \frac{1}{2} \quad \text{and} \quad P(X_i = 1) = \frac{1}{2}$$

Let $Y_1 = X_1$, $Y_2 = X_1 + X_2$, and so on. Since Y_n is the sum of X_1 through X_n, counting the number of heads in n independent tosses of a fair coin, Y_n has the Binomial(n, $\frac{1}{2}$) distribution. The left-hand side of Figure 8-12 shows plots of the Binomial(n, $\frac{1}{2}$) distribution for several values of n. When the probability of success is $\frac{1}{2}$, a binomial distribution is symmetrical for any sample size n. For sample size 10, the shape of the binomial probability distribution closely resembles that of a Gaussian distribution. For sample size 25, the similarity in shape between the binomial and Gaussian distributions is striking (see Exercise 8-17).

Suppose now the coin we toss is not fair. Instead, the probability of a head is $\frac{1}{10}$. Then Y_n counts the number of successes in n independent trials with the probability of success equal to $\frac{1}{10}$ on each trial, and so has a Binomial(n, $\frac{1}{10}$) distribution. The right-hand side of Figure 8-12 shows plots of the Binomial(n, $\frac{1}{10}$) distribution for several values of n. For small sample sizes, the distribution is very asymmetrical. The binomial distribution is more symmetrical for sample size 25 than for smaller sample sizes. The binomial distri-

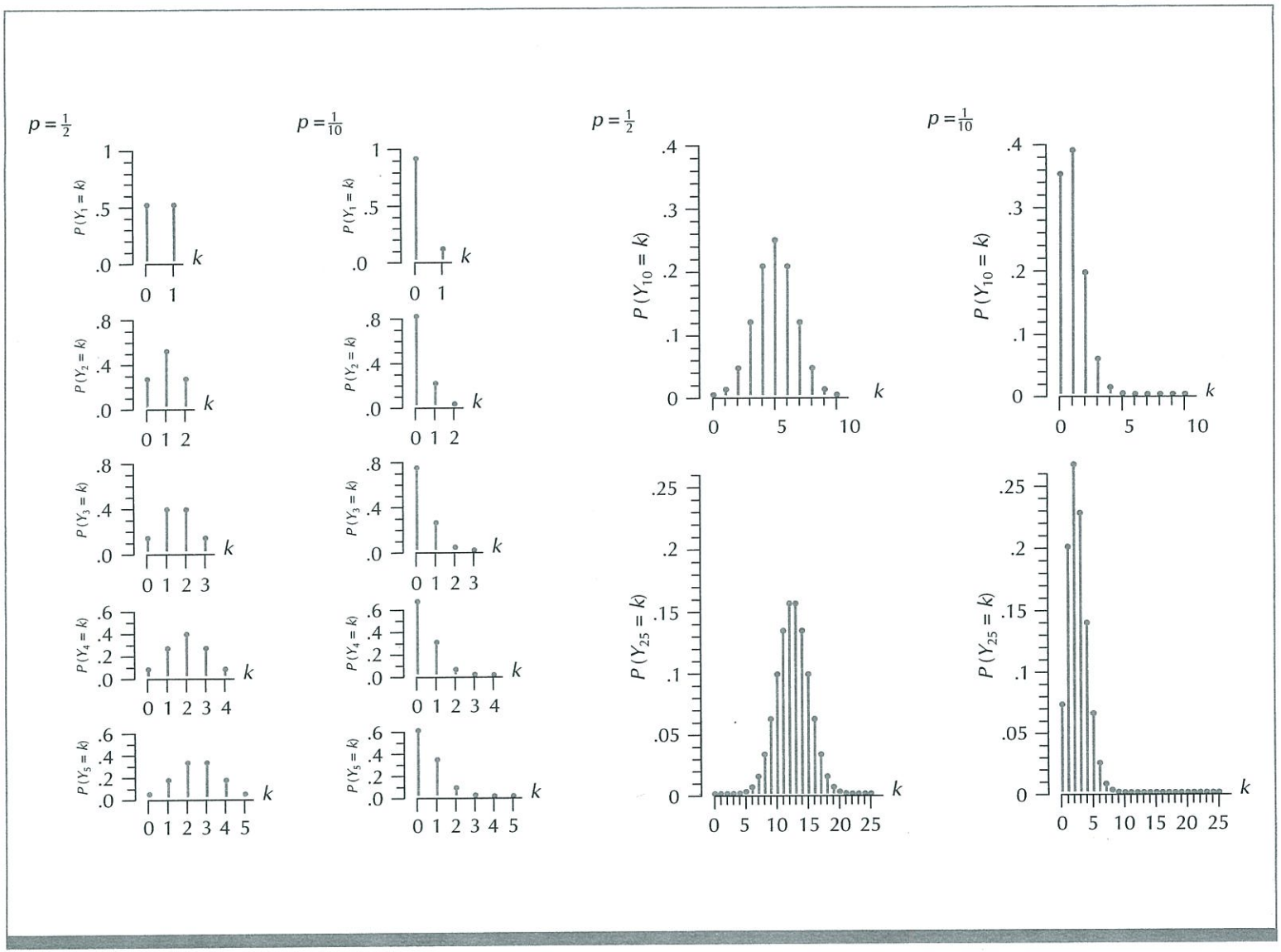

FIGURE 8-12 Plot of the Binomial(n, p) distribution for $p = \frac{1}{2}$ and $p = \frac{1}{10}$ and several values of n

bution will more and more resemble a Gaussian distribution as the sample size increases. For a comparison of the Binomial(25, $\frac{1}{10}$) distribution with a Gaussian distribution, see Exercise 8-18.

As Figure 8-12 illustrates, the sample size necessary for $\bar{X}$ (or Y) to have approximately a Gaussian distribution depends on the shape of the distribution of the random variables X_1 through X_n. If the observations have a symmetrical distribution, a sample of size 30 may be enough. For skewed or multimodal distributions, the sample size must be larger; how much larger depends on how far from symmetrical and unimodal is the distribution of the individual observations.

Another Large-Sample Result

We will find a result related to the Central Limit Theorem useful in large-sample statistical inference. As before, suppose X_1 through X_n are independent random variables with the same probability distribution. Let μ denote the mean and σ the standard deviation of each X_i. Let $\bar{X}$ and s denote the sample mean and sample standard deviation, respectively, of X_1 through X_n. The Central Limit Theorem says that if the sample size n is large enough, then

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

has approximately the standard Gaussian distribution. Suppose we use the sample standard deviation s to estimate σ. A related result states that if n is large enough, then

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has approximately the standard Gaussian distribution. Sometimes we call $s/\sqrt{n}$ the *standard error of* $\bar{X}$ and denote it by SE, so SE $= s/\sqrt{n}$.

> Suppose X_1 through X_n represent a random sample in the probability sense. Let $\bar{X}$ denote the average and s the sample standard deviation of X_1 through X_n. The **standard error of the mean**, SE $= s/\sqrt{n}$, is the estimated standard deviation of the random variable $\bar{X}$.

The large-sample result stated above says that if n is large enough, $(\bar{X} - \mu)/\text{SE}$ has approximately the standard Gaussian distribution.

> **A large-sample result related to the Central Limit Theorem:** Suppose X_1 through X_n are independent random variables having the same probability distribution, with mean μ. Let $\bar{X}$ denote the sample average and s the sample standard deviation of X_1 through X_n. Let SE $= s/\sqrt{n}$ denote the standard error of the mean. If the sample size n is large enough, then
>
> $$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{\text{SE}}$$
>
> has approximately the standard Gaussian distribution. The larger the sample size, the better the approximation.

What is the usefulness of this large-sample approximation? We are often interested in asking questions about the mean μ of a population. Suppose we want to use a sample of independent observations from the population, say X_1 through X_n, to learn about the population mean. The sample mean $\bar{X}$ is a good guess or estimate of the unknown population mean μ. In large-sample statistical analysis, we use $\bar{X}$ as the basis of our inferences about μ. That is, we use the distribution of $\bar{X}$ to make probability statements about μ. The various versions of the Central Limit Theorem say that if the sample size is large enough, the probability distribution of $\bar{X}$ is approximately Gaussian, *no matter what the probability distribution of the individual random variables* X_1 through X_n. This amazing result tells us that if we have a random sample in the probability sense and if the sample size is large enough, we can base inferences about μ on a Gaussian distribution; we do not have to worry about the distribution of the individual observations.

We will see an illustration of the use of this large-sample result in Example 9-3. Section 10-1 gives a more general discussion of large-sample inference about a population mean, based on a Gaussian distribution. In Section 10-2 we look at large-sample inferences about a proportion. These large-sample procedures are based on the Central Limit Theorem result described above. This Central Limit Theorem result can be extended to differences between two sample means. The large-sample procedures discussed in Sections 11-1 and 11-2 are based on this extension to two large samples. In Section 11-1 we consider large-sample inferences about two means and in Section 11-2, large-sample inferences about two proportions.

Summary of Chapter 8

We introduce the standard Gaussian probability distribution and show how to find probabilities for a general Gaussian distribution using probabilities for the standard Gaussian distribution. We discuss the idea of approximating a distribution with a Gaussian distribution.

A version of the Central Limit Theorem states that if X_1 through X_n are independent and identically distributed random variables (with finite mean and variance), then for large enough sample size n, the sample mean $\bar{X}$ has approximately a Gaussian distribution. This result is extremely useful for making inferences about a population mean when the sample size is large.

```
Histogram of C1   N = 20

Midpoint   Count
       0      14  **************
       1       6  ******

Histogram of C41   N = 20

Midpoint   Count
       0       6  ******
       1      10  **********
       2       1  *
       3       2  **
       4       1  *

Histogram of C43   N = 20

Midpoint   Count
       2       2  **
       3       5  *****
       4       5  *****
       5       3  ***
       6       2  **
       7       1  *
       8       1  *
       9       1  *

Histogram of C42   N = 20

Midpoint   Count
       0       1  *
       1       7  *******
       2       7  *******
       3       0
       4       4  ****
       5       0
       6       0
       7       1  *

Histogram of C44   N = 20

Midpoint   Count
       3       1  *
       4       0
       5       2  **
       6       1  *
       7       2  **
       8       2  **
       9       3  ***
      10       5  *****
      11       2  **
      12       1  *
      13       1  *
```

FIGURE M8-3 Histogram of 20 observations on the sum of 1, 5, 10, 20, and 40, respectively, random values simulated from the Bernoulli(.2) distribution

We can display histograms of C1 and C41–C44 to get a feel for the relationship between sample size and the Central Limit Theorem Gaussian approximation to the distribution of a sum (or average):

```
MTB> histogram c1, c41-c44
```

The histograms are shown in Figure M8-3. You may try similar manipulations to investigate the Central Limit Theorem result.

Exercises for Chapter 8

EXERCISE 8-1

Let Z be a standard Gaussian random variable. Draw a sketch of the appropriate area under the standard Gaussian curve and use Table B to find the following probabilities:

a. $P(-2.5 \le Z \le -1.5)$

b. $P(Z \le -2.4)$

c. $P(-1.5 \le Z \le .5)$

d. $P(.5 \le Z \le 1.5)$

e. $P(Z \ge 2.55)$

EXERCISE 8-2

Suppose the random variable Y has a Gaussian distribution with mean 20 and variance 9. Draw a sketch of the appropriate area under the Gaussian curve and find the following probabilities:

a. $P(Y \leq 15)$

b. $P(10 \leq Y \leq 30)$

c. $P(10 \leq Y < 15)$

d. $P(Y \geq 18)$

e. $P(25 \leq Y \leq 30)$

f. $P(14 \leq Y \leq 26)$

g. $P(Y > 23)$

h. $P(Y < 17)$

i. $P(Y \geq 28)$

EXERCISE 8-3

Mercury concentration (in parts per million) is listed for each of 115 swordfish inspected for mercury (Lee and Krutchkoff, 1980).

.05	.07	.07	.13	.13	.19	.24	.25
.28	.32	.39	.45	.46	.53	.54	.56
.60	.60	.61	.62	.65	.71	.72	.75
.76	.79	.81	.81	.82	.82	.82	.83
.83	.83	.84	.85	.89	.90	.91	.92
.92	.93	.95	.95	.97	.97	.98	1.00
1.00	1.01	1.02	1.04	1.05	1.05	1.08	1.10
1.12	1.12	1.14	1.14	1.15	1.16	1.20	1.20
1.20	1.20	1.20	1.21	1.22	1.25	1.25	1.26
1.27	1.27	1.29	1.29	1.29	1.29	1.30	1.31
1.32	1.32	1.37	1.37	1.39	1.39	1.40	1.40
1.41	1.42	1.43	1.44	1.45	1.54	1.54	1.58
1.58	1.60	1.60	1.62	1.62	1.66	1.66	1.68
1.69	1.72	1.74	1.85	1.89	1.96	2.06	2.10
2.23	2.25	2.72					

a. Plot these observations.

b. Can the distribution of these observations be approximated by a Gaussian distribution?

c. At the time these measurements were made, the Food and Drug Administration stated that seafood with more than 1 part per million mercury contamination should not be eaten. Discuss this safety limit in relation to this data set.

EXERCISE 8-4

Serum total cholesterol concentration (mg/dl) is listed below for each of 23 24-month-old baboons, raised on a high cholesterol, saturated fat diet (McMahan, 1981).

141	135	127	200	184	122	219	114	136	253
243	188	239	135	165	140	186	134	110	103
144	252	169							

Plot these observations. Can the distribution of these observations be approximated by a Gaussian distribution?

EXERCISE 8-5

In this study of the association between hyperglycemia and relative hyperinsulinemia, investigators administered standard glucose tolerance tests to 13 control patients and 20 obese patients on the Pediatric Clinical Research Ward, University of Colorado Medical Center. As part of the study, workers determined plasma inorganic phosphate levels from blood samples zero hours after a standard-dose oral glucose challenge. The recorded values of plasma inorganic phosphate (mg/dl) are shown below (Zerbe, 1979).

Control patients:	4.3	3.7	4.0	3.6	4.1	3.8	3.8	4.4
	5.0	3.7	3.7	4.4	4.7			
Obese patients:	4.3	5.0	4.6	4.3	3.1	4.8	3.7	5.4
	3.0	4.9	4.8	4.4	4.9	5.1	4.8	4.2
	6.6	3.6	4.5	4.6				

a. Plot these observations.

b. Do the plasma inorganic phosphate measurements for the control patients appear to be approximately Gaussian distributed?

c. Do the plasma inorganic phosphate measurements for the obese patients appear to be approximately Gaussian distributed?

d. Compare center and variation for these two distributions.

EXERCISE 8-6

In this experiment, investigators studied a method of determining aflatoxin levels in contaminated peanuts. (This problem is important in sampling inspection. Inspectors want to protect consumers while not rejecting too many good peanuts.) They ground the peanuts into meal and then blended a sample of the meal in a chemical solution. They divided this blend equally among 16 centrifuge bottles. The determination of aflatoxin concentration for each bottle (units not given) is shown below (Quesenberry, Whitaker, and Dickens, 1976; from Waltking, Bleffert, and Kiernan, 1968).

121.23	71.69	117.91	91.09	104.86	151.00	125.40
83.94	137.53	83.49	116.78	90.72	100.54	74.59
137.19	146.25					

Plot these observations. Do these observations appear to come from a Gaussian distribution?

EXERCISE 8-7

Mental retardation is associated with some metabolic diseases. In this experiment, investigators studied metabolism of tyrosine among 36 mentally handi-

capped patients (Geertsema and Reinecke, 1984). The measured response was the total amount of tyrosine catabolites excreted in the urine (in μmoles per 100 ml urine).

a. Ten separate measurements on one patient resulted in the following observations:

.325 .317 .375 .325 .508 .117 .150 .317 .275 .383

Plot these observations. Do they appear to come from a Gaussian distribution?

b. The average of ten observations on each of the 36 patients is shown below. These averages are listed in increasing order.

.309	.328	.355	.368	.379	.381	.383	.391
.393	.411	.440	.444	.447	.464	.505	.521
.554	.593	.613	.620	.628	.650	.674	.697
.699	.715	.725	.754	.818	.835	.868	.995
1.099	1.115	1.185	1.693				

Plot these averages. Can the distribution of these averages be approximated by a Gaussian distribution?

c. The investigators selected 1.0 μmole per 100 ml urine as a cutoff for classifying a patient's tyrosine metabolism as clinically negative (less than 1.0, no metabolic problem) or clinically positive (greater than 1.0, a metabolic disorder). Considering the average determinations in part (b), what can you say about these patients in light of this cutoff?

EXERCISE 8-8 Life expectancy at birth in 1985 is plotted for 125 countries in Figure 2-1. Can the distribution of these 125 life expectancies be approximated by a Gaussian distribution?

EXERCISE 8-9 Total fertility rate in 1985 is plotted for 125 countries in Figures 2-4 and 2-15. Can the distribution of these 125 fertility rates be approximated by a Gaussian distribution?

EXERCISE 8-10 A frequency plot of number of cities of over 500,000 persons in 1980, summarizing information for 123 countries, is shown in Figure 2-6. Can the distribution of these 123 data values be approximated by a Gaussian distribution?

EXERCISE 8-11 Number enrolled in primary school in 1984 as percentage of 6–11-year age group is summarized for 119 countries with a histogram in Figure 2-7, with a dot plot in Figure 2-12. Can the distribution of these 119 percentages be approximated by a Gaussian distribution?

EXERCISE 8-12 Number of deaths of children 1–4 years of age per 1,000 children in this age group in 1985 is plotted for 124 countries in Figure 2-13. Can the distribution of these 124 data values be approximated by a Gaussian distribution?

EXERCISE 8-13 In men's golf, all 19 top money winners for 1986 made over $300,000 that year. The average drive for each of these 19 winners is listed below (*USA Today*, January 15, 1987, page 6C).

Player	Average drive (yards)	1986 earnings (dollars)
Greg Norman	277.5	653,295
Bob Tway	268.2	652,780
Payne Stewart	266.4	535,389
Andy Bean	273.9	491,937
Dan Pohl	273.1	463,629
Hal Sutton	262.2	429,433
Tom Kite	253.9	394,164
Ben Crenshaw	261.4	388,168
Ray Floyd	258.6	380,508
B. Langer	259.0	379,799
John Mchaffey	262.9	378,172
Calvin Peete	248.6	374,953
Fuzzy Zoeller	266.4	358,115
Joey Sindelar	277.7	341,230
Jim Thorpe	266.2	326,086
Ken Green	267.0	317,834
Larry Mize	254.9	314,050
Doug Tewell	258.1	310,285
Corey Pavin	260.3	304,557

a. Plot average drive for these 19 golfers.

b. Find the mean and standard deviation of the 19 average drives.

c. Let W be a Gaussian random variable with mean and standard deviation that you calculated in part (b). Find the following probabilities:

(i) $P(255 \leq W \leq 260)$
(ii) $P(260 \leq W \leq 265)$
(iii) $P(265 \leq W \leq 270)$
(iv) $P(W < 255)$
(v) $P(W \geq 270)$

d. Find the proportion of the 19 average drives listed above that are

(i) from 255 to 260
(ii) from 260 to 265
(iii) from 265 to 270
(iv) less than 255
(v) greater than or equal to 270

Compare these proportions with the probabilities you found in part (c).

e. Can the distribution of these 19 average drives be approximated by a Gaussian distribution?

f. Can the distribution of the 19 earnings be approximated by a Gaussian distribution?

EXERCISE 8-14 The percentage of mothers wholly or partially breastfeeding their babies for at least 6 months in 1980–1984 is listed below for 29 countries (Grant, James P., 1987, pages 130–131).

Country	Percentage of mothers breastfeeding their babies	Country	Percentage of mothers breastfeeding their babies
Sierra Leone	94	Peru	72
Malawi	95	Indonesia	97
Niger	30	Congo	97
Rwanda	98	Kenya	84
Yemen	76	Honduras	28
Yemen, Dem.	73	Brazil	19
Burundi	95	Burma	90
Bangladesh	97	El Salvador	77
Sudan	86	Philippines	58
Bolivia	91	Colombia	58
Haiti	85	Thailand	47
Uganda	70	Panama	48
Pakistan	96	Chile	28
Ghana	70	Costa Rica	20
Egypt	91		

a. Plot these data.

b. Find the mean and the standard deviation for these 29 data values.

c. Let Y be a Gaussian random variable with mean and standard deviation that you calculated in part (b). Find the following probabilities:

(i) $P(Y < 30)$
(ii) $P(Y > 90)$
(iii) $P(50 \leq Y \leq 80)$

d. Find the proportion of data values that are:

(i) less than 30
(ii) greater than 90
(iii) from 50 to 80

Compare these proportions with the probabilities you found in part (c).

e. Can the distribution of these 29 percentages be approximated by a Gaussian distribution?

EXERCISE 8-15 Base pay of the governor in 1986 is listed in Exercise 2-19 for each of the 50 states (*USA Today,* December 11, 1986, page 9C).

a. Plot these salaries.

b. Find the mean and the standard deviation for these 50 salaries.

c. Let W be a Gaussian random variable with mean and standard deviation that you calculated in part (b). Find the following probabilities:

(i) $P(58{,}000 \leq W \leq 68{,}000)$
(ii) $P(70{,}000 \leq W \leq 76{,}000)$
(iii) $P(W > 80{,}000)$
(iv) $P(W < 50{,}000)$

d. Find the proportion of the 50 salaries that are:

(i) from \$58,000 to \$68,000
(ii) from \$70,000 to \$76,000
(iii) greater than \$80,000
(iv) less than \$50,000

Compare these proportions with the probabilities you found in part (c).

e. Can the distribution of these 50 salaries be approximated by a Gaussian distribution?

EXERCISE 8-16 Number enrolled in higher education in 1984 as percentage of 20–24-year-old age group is plotted in Figure 8-11 for 119 countries. Can the distribution of these 119 data values be approximated by a Gaussian distribution?

EXERCISE 8-17 Suppose the random variable Y has a Binomial(25, $\frac{1}{2}$) distribution. This distribution is illustrated in Figure 8-12.

a. Find the mean and variance of Y. (See Section 7-2.)

b. Find the following probabilities:

(i) $P(10 \leq Y \leq 15)$
(ii) $P(Y \geq 15)$
(iii) $P(Y \leq 10)$
(iv) $P(8 \leq Y \leq 17)$
(v) $P(Y \leq 7)$
(vi) $P(Y \geq 18)$

c. Let W be a Gaussian random variable with the same mean and variance as Y, which you found in part (a). Find the following probabilities:

(i) $P(10 \leq W \leq 15)$
(ii) $P(W \geq 15)$
(iii) $P(W \leq 10)$
(iv) $P(8 \leq W \leq 17)$
(v) $P(W \leq 7)$
(vi) $P(W \geq 18)$

d. Compare the probabilities you found in part (b) with those you found in part (c). Does the binomial distribution of Y seem to be well approximated by the Gaussian distribution of W?

EXERCISE 8-18 Suppose the random variable X has a Binomial(25, $\frac{1}{10}$) distribution. This distribution is illustrated in Figure 8-12.

a. Find the mean and variance of X. (See Section 7-2.)

b. Find the following probabilities:

(i) $P(1 \leq X \leq 4)$
(ii) $P(X \geq 4)$
(iii) $P(X \leq 1)$
(iv) $P(0 \leq X \leq 5)$
(v) $P(X \geq 6)$

c. Let U be a Gaussian random variable with the same mean and variance as X, which you found in part (a). Find the following probabilities:

(i) $P(1 \leq U \leq 4)$
(ii) $P(U \geq 4)$
(iii) $P(U \leq 1)$
(iv) $P(0 \leq U \leq 5)$
(v) $P(U \geq 6)$

d. Compare the probabilities you found in part (b) with those you found in part (c). Does the binomial distribution of X seem to be well approximated by the Gaussian distribution of U?

EXERCISE 8-19 Fifty college students participated in a coin-tossing experiment (Alex Olsen, personal communication, 1986). For the first stage of the experiment, each student tossed a coin once and recorded whether a head or tail landed face up. For the second stage, each student tossed a coin twice and recorded the number of heads and tails in the two tosses. Similarly, each student recorded the number of heads in five other stages of the experiment, involving 3, 4, 5, 6, and 10 tosses of a coin. The results are shown below:

One toss:	Number of heads	0	1									
	Number of students	22	28									
Two tosses:	Number of heads	0	1	2								
	Number of students	11	23	16								
Three tosses:	Number of heads	0	1	2	3							
	Number of students	7	15	22	6							
Four tosses:	Number of heads	0	1	2	3	4						
	Number of students	4	18	16	10	2						
Five tosses:	Number of heads	0	1	2	3	4	5					
	Number of students	1	7	21	16	4	1					
Six tosses:	Number of heads	0	1	2	3	4	5	6				
	Number of students	1	2	16	17	9	5	0				
Ten tosses:	Number of heads	0	1	2	3	4	5	6	7	8	9	10
	Number of students	0	0	1	8	9	15	10	2	3	2	0

a. Make a frequency plot of the results of each of the seven stages of the experiment.

b. What assumptions must you make to discuss this experiment in terms of the Central Limit Theorem? Discuss the reasonableness of these assumptions for this experiment.

c. Suppose the assumptions you discussed in part (b) hold for this experiment. Discuss how the results of the experiment relate to the Central Limit Theorem.