CHAPTER FIVE

CONTINUOUS PROBABILITY DISTRIBUTIONS

SECTION 5.1 Introduction

In this chapter continuous probability distributions are discussed. In particular, the normal distribution, which is the most widely used distribution in statistical work, is explored in depth.

The normal, or Gaussian, or "bell-shaped," distribution is the cornerstone of most of the methods of estimation and hypothesis testing that are developed in the rest of this text. Many random variables, such as the distribution of birthweights or blood pressures in the general population, tend to approximately follow a normal distribution. In addition, many random variables that are not themselves normal, are closely approximated by a normal distribution when summed many times. In such cases, using the normal distribution is desirable, since tables for the normal distribution are more widely available than those for many other distributions.

EXAMPLE 5.1 **Infectious Disease** The number of neutrophils in a sample of 2 white blood cells is not normally distributed, but the number in a sample of 100 white blood cells is very close to being normally distributed. ■■■

SECTION 5.2 General Concepts

We want to develop an analogue for a continuous random variable to the concept of a probability mass function, as was developed for a discrete random variable in Section 4.3. Thus, we would like to know which values are more probable than others and how probable they are.

EXAMPLE 5.2 **Hypertension** Consider the distribution of diastolic blood-pressure measurements in 35–44-year-old men. In actual practice this distribution is discrete because only a finite number of blood-pressure values are possible, since the measurement is only accurate to within 2 mm Hg, or in some cases 5 mm Hg. However, assume that there is no measurement error and hence the random variable can take on a continuum of possible values. One consequence of this assumption is that the probabilities of specific blood-pressure measurement values such as 117.3 are 0 and, thus, the concept of a probability mass function cannot be used. The proof of this statement is beyond the scope of this text. Instead, we speak in terms of the probability that blood pressure falls within a range of values. Thus, the probabilities of blood pressures (denoted by X) falling in the ranges of $90 \leq X < 95$, $95 \leq X < 100$, and $X \geq 100$ might be 15%, 5%, and 2%, respectively. People whose blood pressures fall in these ranges might be denoted as borderline, mild hypertensive, and severe hypertensive, respectively. ■■■

Although the probability of exactly obtaining any value is 0, we still have the intuitive notion that certain ranges of values occur more frequently than others. This notion can be quantified using the concept of a probability density function.

DEFINITION 5.1

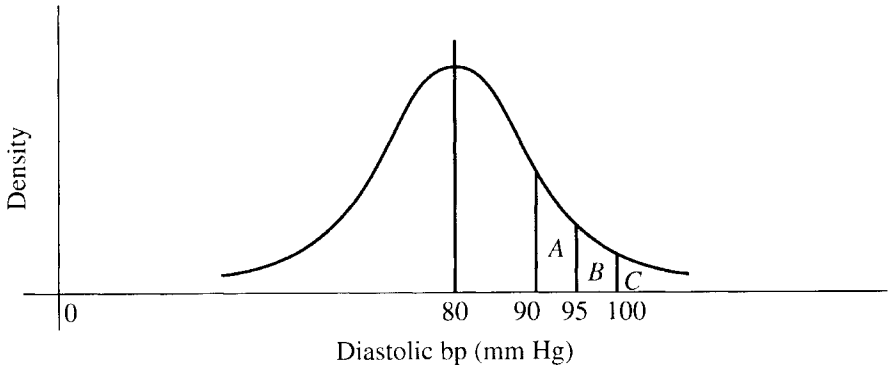
The **probability density function** of the random variable X is a curve such that the area under the curve between any two points a and b is equal to the probability that the random variable X falls between a and b. Thus, the total area under the curve over the possible range of values for the random variable is 1. ■

The probability density function takes on high values in regions of high probability and low values in regions of low probability.

EXAMPLE 5.3

Hypertension The probability density function for diastolic blood pressure in 35–44-year-old men is given in Figure 5.1. Areas A, B, and C correspond to the probabilities of being borderline, mild hypertensive, and severe hypertensive, respectively. Furthermore, the most likely range of values for diastolic blood pressure occurs around 80 mm Hg, with the values becoming increasingly unlikely as we move further away from 80. ■■■

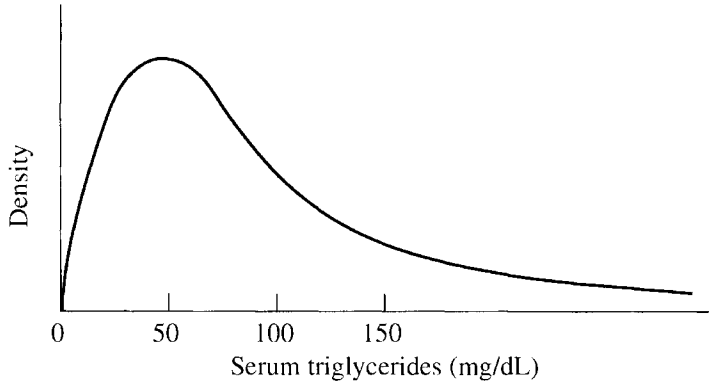
FIGURE 5.1
Probability density function for diastolic blood pressure in 35–44-year-old men



EXAMPLE 5.4

Cardiovascular Disease Serum triglycerides is an asymmetric, positively skewed, continuous random variable whose probability density function appears in Figure 5.2. ■■■

FIGURE 5.2
Probability density function for serum triglycerides



The cumulative distribution function is defined similarly to that for a discrete random variable (Section 4.6).

DEFINITION 5.2

The **cumulative distribution function** for the random variable X evaluated at the point a is defined as the probability that X will take on values $\leq a$. It is represented by the area under the probability density function to the left of a. ■

EXAMPLE 5.5

Obstetrics The probability density function for the random variable representing the distribution of birthweights (oz) in the general population is given in Figure 5.3. The cumulative distribution function evaluated at 88 oz $= Pr(X \leq 88)$ is represented by the area under this curve to the left of 88 oz. The region $X \leq 88$ oz has a special meaning in obstetrics, since 88 oz is the cutoff point usually used by obstetricians for identifying low-birthweight infants. Such infants are generally at higher risk for various unfavorable outcomes, such as mortality in the first year of life. ■■■

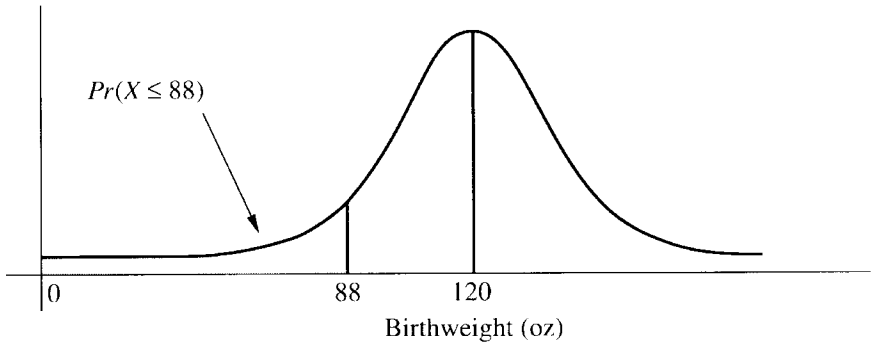


FIGURE 5.3
Cumulative distribution function evaluated at 88 oz for the distribution of birthweights in the general population

Generally, a distinction will not be made between the probabilities $Pr(X \leq x)$ and $Pr(X < x)$ when X is a continuous random variable. The reason is that they represent the same quantity, because the probability of individual values is 0, that is, $Pr(X = x) = 0$.

The expected value and variance for continuous random variables have the same meaning as for discrete random variables (Sections 4.4 and 4.5). However, the mathematical definition of these terms is beyond the scope of this book.

DEFINITION 5.3

The **expected value** of a continuous random variable X, denoted by $E(X)$, or μ, is the average value taken on by the random variable. ■

DEFINITION 5.4

The **variance** of a continuous random variable X, denoted by $Var(X)$ or σ^2, is the average squared distance of each value of the random variable from its expected value. The standard deviation, or σ, is the square root of the variance, that is, $\sigma = \sqrt{Var(X)}$. ■

EXAMPLE 5.6

Hypertension The expected value and standard deviation of the distribution of diastolic blood pressures in 35–44-year-old men are 80 mm Hg and 12 mm Hg, respectively. ■■■

SECTION 5.3 The Normal Distribution

The normal distribution is the most widely used continuous distribution. It is also frequently referred to as the Gaussian distribution, after the well-known mathematician, Gauss.

EXAMPLE 5.7 **Hypertension** The distribution of body weights or of diastolic blood pressures for a group of 35–44-year-old males will follow a normal distribution. ■■■

Many other distributions that are not themselves normal can be made normal by transforming the data onto a different scale.

EXAMPLE 5.8 **Cardiovascular Disease** The distribution of serum-triglyceride concentrations from this same group of 35–44-year-old males is likely to be positively skewed. However, the log transformation of these measurements will usually follow a normal distribution. ■■■

Generally speaking, any random variable that can be expressed as a sum of many other random variables can be well approximated by a normal distribution.

EXAMPLE 5.9 **Infectious Disease** The distribution of the number of lymphocytes in a differential of 100 white blood cells (refer to Example 4.15 for the definition of a differential) will tend to be normally distributed, since this random variable is a sum of 100 random variables, each representing whether or not an individual cell is a lymphocyte. ■■■

Thus, because of its omnipresence, the normal distribution is vital to statistical work, and most of the estimation procedures and hypothesis tests that we will study are based on the assumption that the random variable being considered has an underlying normal distribution.

Another important area of application of the normal distribution is as an approximating distribution to other distributions. The normal distribution is generally more convenient to work with than any other distribution, particularly in hypothesis testing. Thus, if an accurate normal approximation to some other distribution can be found, then we will often use it.

DEFINITION 5.5

The **normal distribution** is defined by its **probability density function**, which is given as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty < x < \infty$$

for some parameters μ, σ, where $\sigma > 0$. ■

The exp function merely implies that the quantity to the right in brackets is the power to which "e" (≈ 2.71828) is raised. A plot of this probability density function is given in Figure 5.4.

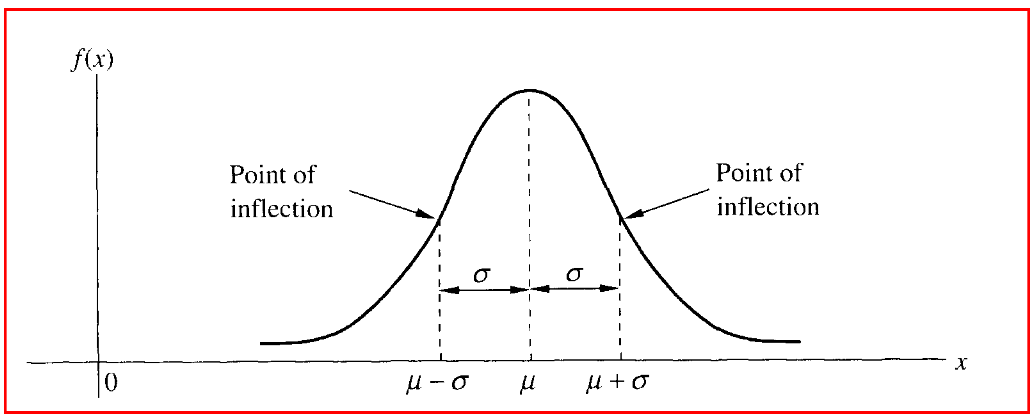


FIGURE 5.4 Probability density function for a normal distribution with mean μ and variance σ^2

The density function follows a bell-shaped curve, with the most frequently occurring value at μ. The curve is symmetric about μ, with points of inflection on each side of μ at $\mu - \sigma$ and $\mu + \sigma$, respectively. A point of inflection is a point where the slope of the curve changes direction. In Figure 5.4, the slope of the curve increases to the left of $\mu - \sigma$ and then starts to decrease to the right of $\mu - \sigma$ and continues to decrease until reaching $\mu + \sigma$, after which it starts increasing again. Thus, the distances from μ to the points of inflection provide a good visual sense of the magnitude of the parameter σ.

You may wonder why the parameters μ and σ^2 have been used to define the normal distribution when the expected value and variance of an arbitrary distribution were previously defined as μ and σ^2. Indeed, from the definition of the normal distribution it can be shown, using calculus methods, that μ and σ^2 are, respectively, the expected value and variance of this distribution.

EXAMPLE 5.10

For diastolic blood pressure the parameters might be $\mu = 80$ mm Hg, $\sigma = 12$ mm Hg; for birthweight they might be $\mu = 120$ oz, $\sigma = 15$ oz. ■■■

Interestingly, the entire shape of the normal distribution is determined by the two parameters μ and σ^2. If two normal distributions with the same variance σ^2 and different means μ_1, μ_2, where $\mu_2 > \mu_1$, are compared, then their density functions will appear as in Figure 5.5.

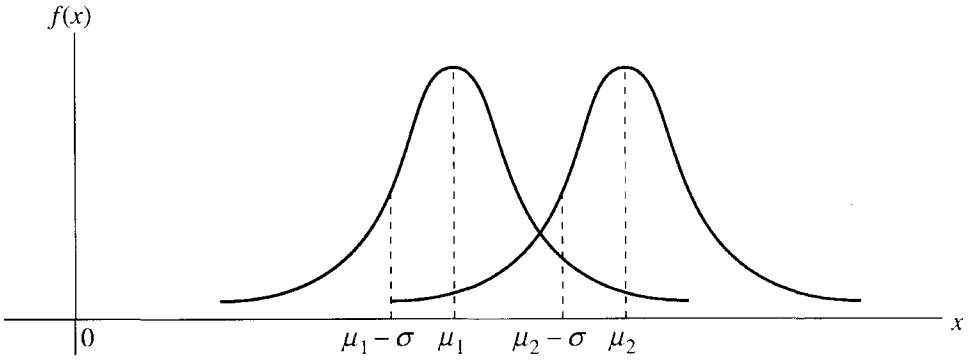


FIGURE 5.5
Comparison of two normal distributions with the same variance and different means

Similarly, two normal distributions with the same mean but different variances ($\sigma_2^2 > \sigma_1^2$) can be compared, as shown in Figure 5.6. Note that the area under any normal density function must be 1. Thus, the two normal distributions shown in Figure 5.6 must cross, since otherwise one curve would remain completely above the other and the areas under both curves could not simultaneously be 1.

DEFINITION 5.6

A **normal distribution with mean μ and variance σ^2** will generally be referred to as an $N(\mu, \sigma^2)$ distribution. ■

Note that the second parameter is always the variance σ^2 and not the standard deviation σ.

Another property of the normal distribution is that the height $= 1/(\sqrt{2\pi}\sigma)$. Thus, the height is inversely proportional to σ. This helps us to visualize σ, since the height

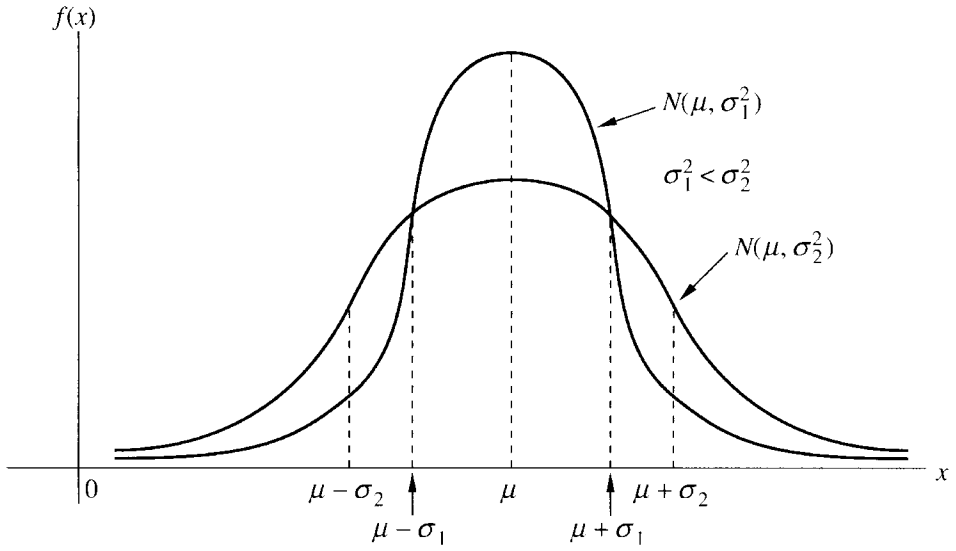


FIGURE 5.6 Comparison of two normal distributions with the same mean and different variances

of the $N(\mu, \sigma_1^2)$ distribution in Figure 5.6 is greater than the height of the $N(\mu, \sigma_2^2)$ distribution.

DEFINITION 5.7

A normal distribution with mean 0 and variance 1 will be referred to as a **standard**, or **unit**, normal distribution. This distribution is denoted by $N(0, 1)$. ■

We will see that any information concerning an $N(\mu, \sigma^2)$ distribution can be obtained from appropriate manipulations of an $N(0, 1)$ distribution.

SECTION 5.4 Properties of the Standard Normal Distribution

To become familiar with the $N(0, 1)$ distribution, some of its properties will be reviewed. First, the probability density function in this case reduces to

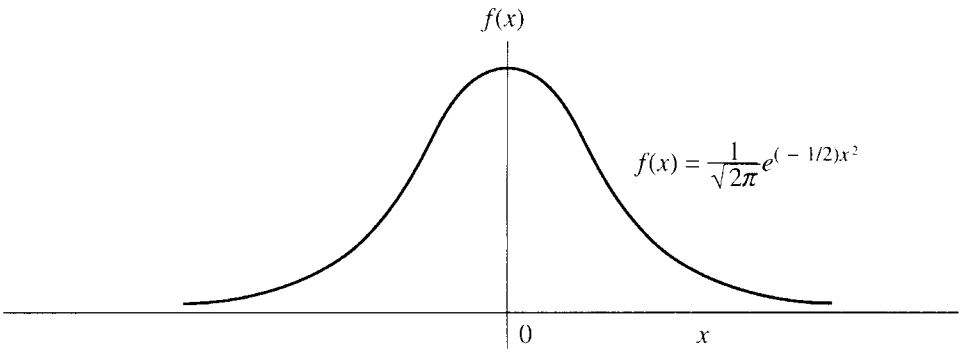
5.1

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2}, \qquad -\infty < x < +\infty$$

This distribution is symmetrical about 0, since $f(x) = f(-x)$, and is depicted in Figure 5.7.

It can be shown that about 68% of the area under the normal density lies between $+1$ and -1, about 95% of the area lies between $+2$ and -2, and about 99% lies between $+2.5$ and -2.5.

FIGURE 5.7
Probability density function for a standard normal distribution



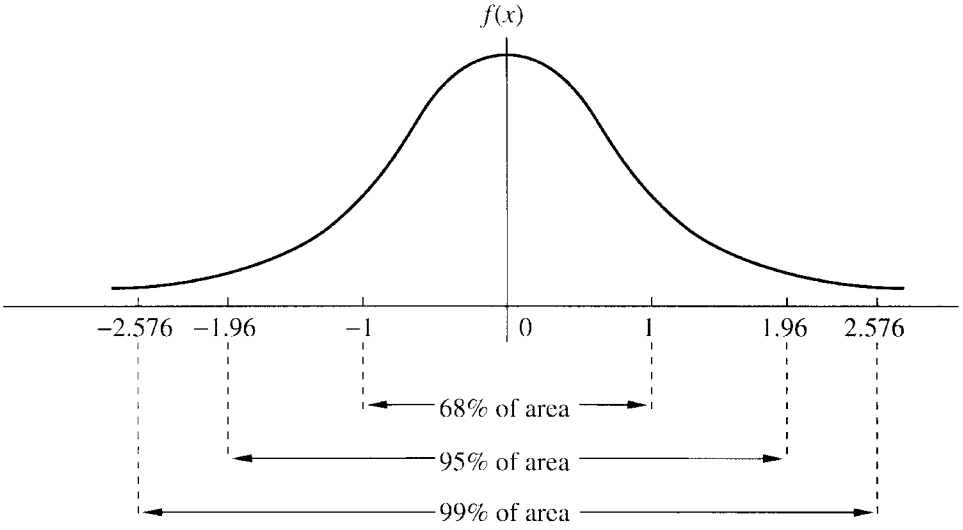
These relationships can be expressed more precisely by saying that

$$Pr(-1 < X < +1) = .6827 \qquad Pr(-1.96 < X < +1.96) = .95$$

$$Pr(-2.576 < X < +2.576) = .99$$

Thus, the standard normal distribution slopes off very rapidly, and absolute values greater than 3 are unlikely. These relationships are depicted in Figure 5.8.

FIGURE 5.8
Empirical properties of the standard normal distribution



Tables of the area under the normal density function, or so-called normal tables, take advantage of the symmetry properties of the normal distribution and generally are concerned with areas for positive values of x.

DEFINITION 5.8

The **cumulative distribution function for a standard normal distribution** is denoted by

$$\Phi(x) = Pr(X \leq x)$$

where X follows an $N(0, 1)$ distribution. This function is depicted in Figure 5.9. ■

DEFINITION 5.9

The symbol $\sim$ is used as shorthand for the phrase "**is distributed as**." Thus, $X \sim N(0, 1)$ means that the random variable X is distributed as an $N(0, 1)$ distribution.

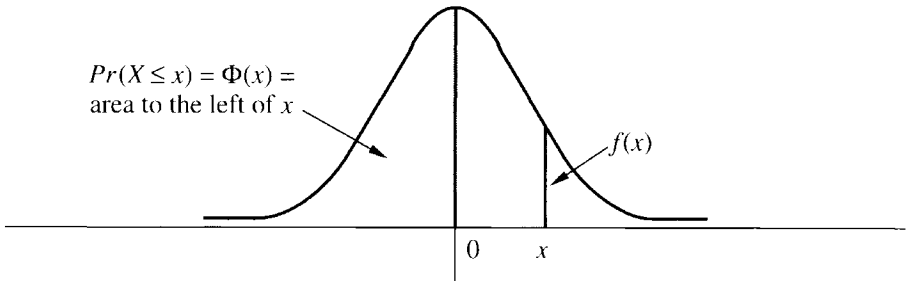


FIGURE 5.9
Cumulative distribution function [$\Phi(x)$] for a standard normal random variable

5.4.1 Using Normal Tables

Under column A in Table 3 of the Appendix, $\Phi(x)$ for various positive values of x for a standard normal distribution are presented. This cumulative distribution function is depicted in Figure 5.10. Notice that the area to the left of 0 is 0.5. Furthermore, the area to the left of x approaches 0 as x becomes small and approaches 1 as x becomes large.

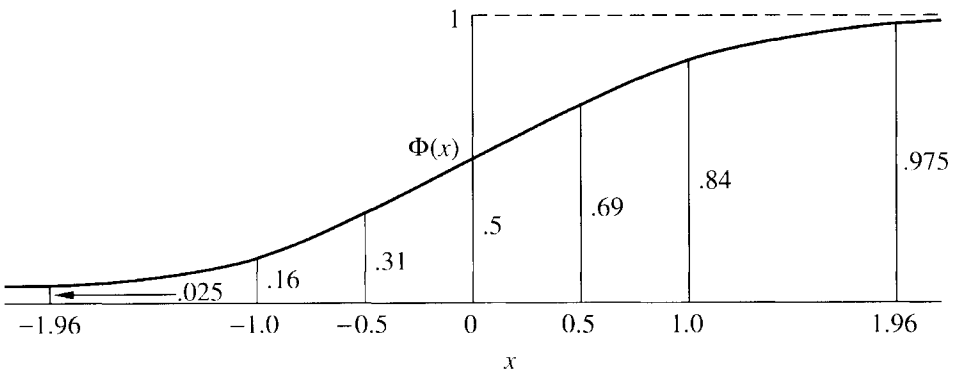


FIGURE 5.10
Cumulative distribution function for a standard normal distribution [$\Phi(x)$]

EXAMPLE 5.11 If $X \sim N(0, 1)$

then find $Pr(X \leq 1.96)$ and $Pr(X \leq 1)$

SOLUTION From Table 3, column A,

$\Phi(1.96) = .975$ and $\Phi(1) = .8413$ ■■■

5.2 Symmetry Properties of the Standard Normal Distribution

From the symmetry properties of the standard normal distribution,

$$\Phi(-x) = Pr(X \leq -x) = Pr(X \geq x) = 1 - Pr(X \leq x) = 1 - \Phi(x)$$

This symmetry property is depicted in Figure 5.11.

The right-hand tail of the standard normal distribution $= Pr(X \geq x)$ is provided in column B of Table 3.

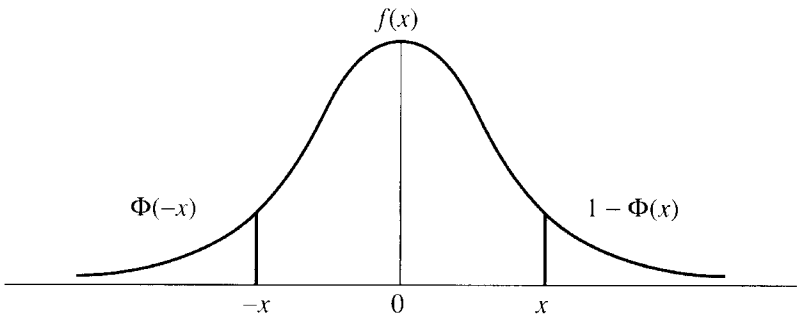


FIGURE 5.11
Illustration of the symmetry properties of the normal distribution

EXAMPLE 5.12 Calculate $Pr(X \leq -1.96)$

if $X \sim N(0, 1)$

SOLUTION $Pr(X \leq -1.96) = Pr(X \geq 1.96) = .0250$ from column B of Table 3 ■■■

Furthermore, for any numbers a, b, we have $\Pr(a \leq X \leq b) = Pr(X \leq b) - Pr(X \leq a)$ and thus we can evaluate $Pr(a \leq X \leq b)$ for any a, b from Table 3.

EXAMPLE 5.13 Compute $Pr(-1 \leq X \leq 1.5)$

if $X \sim N(0, 1)$

SOLUTION
$$\begin{aligned} Pr(-1 \leq X \leq 1.5) &= Pr(X \leq 1.5) - Pr(X \leq -1) \\ &= Pr(X \leq 1.5) - Pr(X \geq 1) = .9332 - .1587 \\ &= .7745 \end{aligned}$$
■■■

EXAMPLE 5.14 **Pulmonary Disease** Forced Vital Capacity (FVC) is a standard measure of pulmonary function and represents the volume of air a person can expel in 6 seconds. A topic of current research interest is to look at potential risk factors, such as cigarette smoking, air pollution, or the type of stove used in the home, that may affect FVC in grade school children. One problem is that pulmonary function is affected by age, sex, and height, and these variables must be corrected for before looking at other risk factors. One way to make these adjustments for a particular child is to find the mean μ and standard deviation σ for children of the same age (in 1-year age groups), sex, and height (in 2-in. height groups) from large national surveys and compute a **standardized FVC**, which is defined as $(x - \mu)/\sigma$, where x is the original FVC. The standardized FVC would then approximately follow an $N(0, 1)$ distribution. Suppose that a child is considered in poor pulmonary health if his or her standardized FVC < -1.5. What percentage of children are in poor pulmonary health?

SOLUTION $Pr(X < -1.5) = Pr(X > 1.5) = .0668$

Thus, about 7% of children are in poor pulmonary health. ■■■

In many instances we will be concerned with tail areas on either side of 0 for a standard normal distribution. For example, the *normal range* for a biological quantity is often defined by a range within x standard deviations of the mean for some specified value of x. The probability of a value falling in this range is given by $Pr(-x \leq X \leq x)$ for a standard normal distribution. This quantity is tabulated in column D of Table 3 for various values of x.

EXAMPLE 5.15 **Pulmonary Disease** Suppose a child is considered to have normal lung growth if his or her standardized FVC is within 1.5 standard deviations of the mean. What proportion of children are within the normal range?

SOLUTION Compute $Pr(-1.5 \leq X \leq 1.5)$. Under 1.50 in Table 3, column D, this quantity is given as .8664. Thus, about 87% of children are considered to have normal lung growth using this definition. ■■■

Finally, in column C of Table 3, the area under the standard normal density from 0 to X is provided, since these areas will occasionally prove useful in work on statistical inference.

EXAMPLE 5.16 Find the area under the standard normal density from 0 to 1.45.

SOLUTION Refer to column C of Table 3 under 1.45. The appropriate area is given by .4265. ■■■

Of course, the areas given in columns A, B, C, and D are redundant in that *all* computations concerning the standard normal distribution could be performed using any one of these columns. In particular, we have seen that $B(x) = 1 - A(x)$. Also, from the symmetry of the normal distribution, we can easily show that $C(x) = A(x) - .5, D(x) = 2 \times C(x) = 2 \times A(x) - 1.0$. However, this redundancy is deliberate, since for some applications one or the other of these columns will be more convenient to use.

The percentiles of a standard normal distribution are often referred to in statistical inference. For this purpose the following definition is introduced:

DEFINITION 5.10

The **(100 × u)th percentile** of a standard normal distribution is denoted by z_u. It is defined by the relationship

$$Pr(X < z_u) = u, \qquad \text{where } X \sim N(0, 1)$$ ■

z_u is depicted graphically in Figure 5.12.

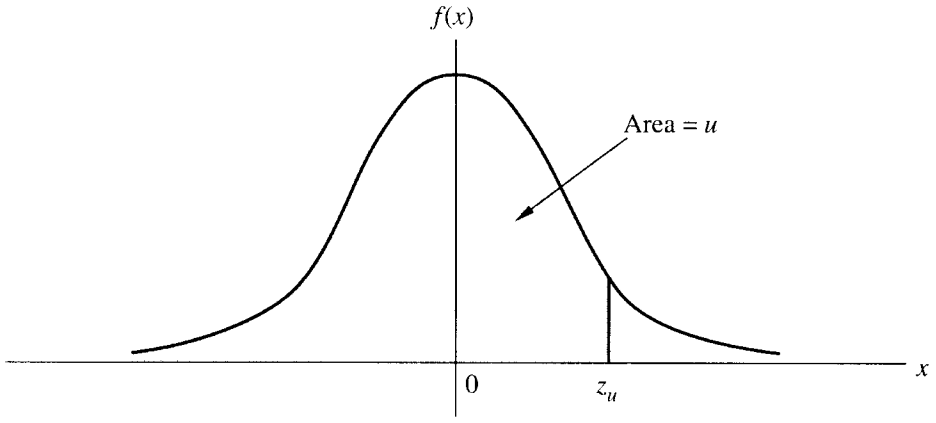


FIGURE 5.12 Graphical display of the (100 × u)th percentile of a standard normal distribution (z_u)

The function z_u is sometimes referred to as the *Inverse Normal Function*. In previous uses of the normal table, we were given a value x and have used the normal tables to evaluate the area to the left of x (i.e., $\Phi[x]$) for a standard normal distribution.

To obtain z_u, we perform this operation in reverse. Thus, to evaluate z_u, we must find the area u in column A of Table 3 and then find the value z_u that corresponds to this area. If $u < 0.5$, then we use the symmetry properties of the normal distribution to obtain $z_u = -z_{1-u}$, where z_{1-u} can be obtained from Table 3.

EXAMPLE 5.17 Compute $z_{.975}$, $z_{.95}$, $z_{.5}$, and $z_{.025}$

SOLUTION From Table 3 we have that

$$\Phi(1.96) = .975$$
$$\Phi(1.645) = .95$$
$$\Phi(0) = .5$$
$$\Phi(-1.96) = 1 - \Phi(1.96) = 1 - .975 = .025$$

Thus,

$$z_{.975} = 1.96$$
$$z_{.95} = 1.645$$
$$z_{.5} = 0$$
$$z_{.025} = -1.96$$

■■■

The percentiles z_u will be frequently used in our work on hypothesis testing in Chapters 7–13.

SECTION 5.5 Conversion from an $N(\mu, \sigma^2)$ Distribution to an $N(0, 1)$ Distribution

EXAMPLE 5.18 **Hypertension** Suppose a borderline hypertensive is defined as a person whose diastolic blood pressure is between 90 and 95 mm Hg inclusive, and the subjects are 35–44-year-old males whose blood pressures are normally distributed with mean 80 and variance 144. What is the probability that a randomly selected person from this population will be a borderline hypertensive? This question can be restated more precisely:

If $X \sim N(80, 144)$

then what is $Pr(90 < X < 95)$

(The solution is given on page 118.)

■■■

More generally, the following question can be asked: If $X \sim N(\mu, \sigma^2)$, then what is $Pr(a < X < b)$ for any a, b? The basic idea is to convert a probability statement about an $N(\mu, \sigma^2)$ distribution to an equivalent probability statement about an $N(0, 1)$ distribution. Consider the random variable $Z = (X - \mu)/\sigma$. We can show that the following relationship holds:

5.3

If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$

then $Z \sim N(0, 1)$.

To see this, compute the expected value and variance of Z. To accomplish this, keep in mind that the expected value and variance have the same properties as the sample mean and variance upon addition of and/or multiplication by a constant. Specifically, for any constant c,

$$E(X + c) = E(X) + c$$
$$E(cX) = cE(X)$$
$$Var(X + c) = Var(X)$$
$$Var(cX) = c^2 Var(X)$$

Therefore, applying these principles,

$$E(Z) = E\left[\frac{X - \mu}{\sigma}\right] = \left(\frac{1}{\sigma}\right)E(X - \mu) = \left(\frac{1}{\sigma}\right)[E(X) - E(\mu)]$$
$$= \left(\frac{1}{\sigma}\right)[E(X) - \mu] = \left(\frac{1}{\sigma}\right)(\mu - \mu) = 0$$
$$Var(Z) = Var\left[\frac{X - \mu}{\sigma}\right] = \left(\frac{1}{\sigma^2}\right)Var(X - \mu)$$
$$= \left(\frac{1}{\sigma^2}\right)Var(X) = \left(\frac{1}{\sigma^2}\right)\sigma^2 = 1$$

Thus, the expected value of Z is 0 and the variance of Z is 1. It is also true that normality is preserved when converting from the random variable X to the random variable $Z = (X - \mu)/\sigma$, but to show this is beyond the scope of this book. Therefore, $Z \sim N(0, 1)$.

We now wish to convert $Pr(a < X < b)$ into a probability statement about Z, since Z is a standard normal random variable and tables are available only for the standard normal distribution. This conversion is given as follows:

5.4 Evaluation of Probabilities for Any Normal Distribution via Standardization

If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$

then $Pr(a < X < b) = Pr\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = \Phi[(b - \mu)/\sigma] - \Phi[(a - \mu)/\sigma]$

Since the Φ function, which is the cumulative distribution function for a standard normal distribution, is given in column A of Table 3 of the Appendix, probabilities for *any* normal distribution can now be evaluated. This procedure is depicted in Figure 5.13.

To see this, note that $a < X < b$ if and only if $(a - \mu)/\sigma < Z < (b - \mu)/\sigma$. To demonstrate this, note that the inequality $a < X < b$ can be written as two inequalities, $a < X$ and $X < b$. If μ is subtracted from both sides of each inequality, we get

$$a - \mu < X - \mu \quad \text{and} \quad X - \mu < b - \mu$$

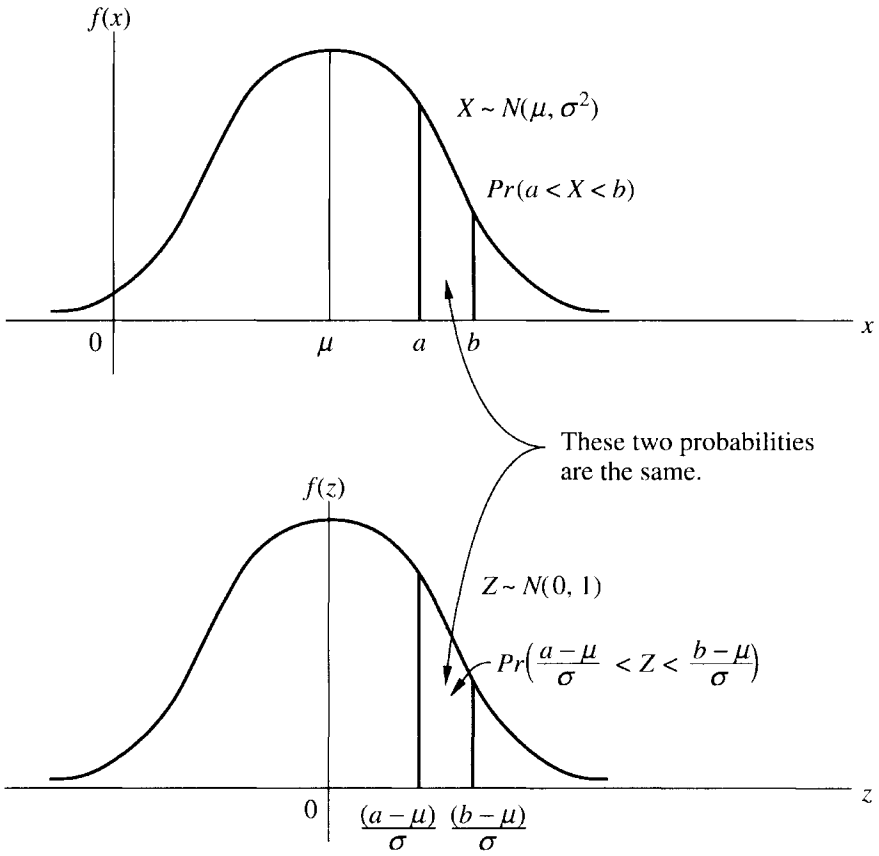


FIGURE 5.13
Evaluation of probabilities for any normal distribution via standardization

Also, if both sides of each inequality are divided by σ, we obtain

$$\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \quad \text{and} \quad \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}$$

or upon rewriting,

$$\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma} \quad \text{or} \quad \frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}$$

Therefore, it follows that

$$Pr(a < X < b) = Pr\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right]$$

However, since $Z \sim N(0, 1)$,

$$Pr(a < X < b) = Pr\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right] = \Phi\left[\frac{b-\mu}{\sigma}\right] - \Phi\left[\frac{a-\mu}{\sigma}\right]$$

This procedure is known as **standardization of a normal variable**.

SOLUTION TO EXAMPLE 5.18 The probability of being a borderline hypertensive among the group of 35–44-year-old males can now be calculated.

$$\begin{aligned} Pr(90 < X < 95) &= Pr\left(\frac{90 - 80}{12} < Z < \frac{95 - 80}{12}\right) \\ &= Pr(0.83 < Z < 1.25) = \Phi(1.25) - \Phi(0.83) \\ &= .8944 - .7967 = .098 \end{aligned}$$

Thus, approximately 9.8% of this population will be borderline hypertensive. ■■■

EXAMPLE 5.19 **Botany** Suppose that tree diameters of a certain species of tree from some defined forest area are assumed to be normally distributed with mean 8 in. and standard deviation 2 in. Find the probability of a tree having an unusually large diameter, which is defined as > 12 in.

SOLUTION We have $X \sim N(8, 4)$ and require

$$\begin{aligned} Pr(X > 12) &= 1 - Pr(X < 12) = 1 - Pr\left(Z < \frac{12 - 8}{2}\right) \\ &= 1 - Pr(Z < 2.0) = 1 - .977 = .023 \end{aligned}$$

Thus, 2.3% of trees from this area have an unusually large diameter. ■■■

The general principle is that for any probability statement concerning normal random variables of the form $Pr(a < X < b)$, the population mean μ is subtracted from each boundary point and divided by the standard deviation σ to obtain an equivalent probability statement for the standard normal random variable Z,

$$Pr[(a - \mu)/\sigma < Z < (b - \mu)/\sigma]$$

The standard normal tables are then used to evaluate this latter probability.

EXAMPLE 5.20 **Cerebrovascular Disease** Diagnosing stroke strictly on the basis of clinical symptoms is difficult. A standard diagnostic test used in clinical medicine to detect stroke in patients is the angiogram. This test has some risks for the patient, and several noninvasive techniques have been developed that are hoped to be as effective as the angiogram. One such method utilizes the measurement of cerebral blood flow (CBF) in the brain, since stroke patients tend to have lower levels of CBF than normal. Assume that in the general population, CBF is normally distributed with mean 75 and standard deviation 17. A patient is classified as being at risk for stroke if his or her CBF is less than 40. What proportion of normal patients will be mistakenly classified as being at risk for stroke?

SOLUTION Let X be the random variable representing CBF. Then $X \sim N(75, 17^2) = N(75, 289)$. We want to find $Pr(X < 40)$. We standardize the limit of 40 so as to use the standard normal distribution. The standardized limit is $(40 - 75)/17 = -2.06$. Thus, if Z represents the standardized normal random variable $= (X - \mu)/\sigma$, then

$$\begin{aligned} Pr(X < 40) &= Pr(Z < -2.06) \\ &= \Phi(-2.06) = 1 - \Phi(2.06) = 1 - .9803 \approx .020 \end{aligned}$$

Thus, about 2.0% of normal patients will be incorrectly classified as being at risk for stroke. ■■■

EXAMPLE 5.21 **Ophthalmology** Glaucoma is a disease of the eye that is manifested by high intraocular pressure. The distribution of intraocular pressure in the general population is approximately normal with mean 16 mm Hg and standard deviation 3 mm Hg. If the normal range for intraocular pressure is considered to be between 12 mm Hg and 20 mm Hg, then what percentage of the general population would fall within this range?

SOLUTION We wish to calculate $Pr(12 \leq X \leq 20)$, where $X \sim N(16, 9)$. The limits 12 and 20 must first be standardized so as to use the standard normal tables to evaluate this probability. The standardized limits are $(12 - 16)/3 = -1.33$ and $(20 - 16)/3 = +1.33$. Thus, we evaluate $Pr(-1.33 \leq Z \leq 1.33)$, where Z follows a standard normal distribution.

$$\begin{aligned} Pr(-1.33 \leq Z \leq 1.33) &= Pr(Z \leq 1.33) - Pr(Z \leq -1.33) \\ &= Pr(Z \leq 1.33) - [1 - Pr(Z \leq 1.33)] \\ &= 2Pr(Z \leq 1.33) - 1 = 2(.9082) - 1 = .816 \end{aligned}$$

Alternatively, $Pr(-1.33 \leq Z \leq 1.33)$ could be evaluated directly from the 1.33 row under column D of Table 3 of the Appendix, yielding a probability of .8165. Thus, 81.6% of the population have intraocular pressures in the normal range. These calculations are depicted in Figure 5.14. ■■■

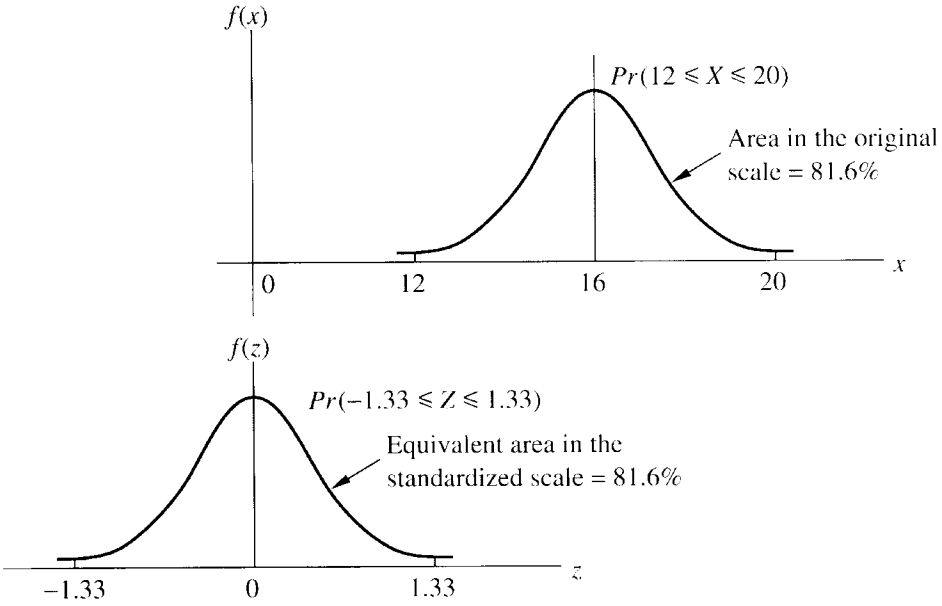


FIGURE 5.14 Calculation of the proportion of people with intraocular pressures in the normal range

SECTION 5.6 Linear Combinations of Random Variables

In work on statistical inference, the use of sums or differences or more complicated linear functions of random variables (either continuous or discrete) will often arise. For this reason, the properties of linear combinations of random variables are important to discuss.

DEFINITION 5.11

A **linear combination** L of the random variables $X_1, \ldots, X_n$ is defined as any function of the form $L = c_1X_1 + \cdots + c_nX_n$. ■

EXAMPLE 5.22 **Renal Disease** Suppose X_1, X_2 represent serum-creatinine levels for two different individuals with end-stage renal disease. Represent the sum, difference, and average of the random variables X_1, X_2 as linear combinations of the random variables X_1, X_2.

SOLUTION The sum is $X_1 + X_2$, where $c_1 = 1$, $c_2 = 1$. The difference is $X_1 - X_2$, where $c_1 = 1$, $c_2 = -1$. The average is $(X_1 + X_2)/2$, where $c_1 = 0.5$, $c_2 = 0.5$. ■■■

It will often be necessary to compute the expected value and variance of linear combinations of random variables. To find the expected value of L, the principle that the expected value of the sum of n random variables is the sum of the n respective expected values is used. Applying this principle,

$$\begin{aligned} E(L) &= E(c_1X_1 + \cdots + c_nX_n) \\ &= E(c_1X_1) + \cdots + E(c_nX_n) = c_1E(X_1) + \cdots + c_nE(X_n) \end{aligned}$$

5.5 Expected Value of Linear Combinations of Random Variables

The expected value of the linear combination $L = \sum_{i=1}^{n} c_iX_i$ is $E(L) = \sum_{i=1}^{n} c_iE(X_i)$.

EXAMPLE 5.23 **Renal Disease** Suppose the expected values of serum creatinine for the two individuals in Example 5.22 are 1.5 and 1.3, respectively. What is the expected value of the average serum-creatinine level of these two individuals?

SOLUTION The expected value of the average serum-creatinine level $= E(0.5X_1 + 0.5X_2) = 0.5E(X_1) + 0.5E(X_2) = 0.75 + 0.65 = 1.4$. ■■■

To compute the variance of linear combinations of random variables, we assume that the random variables are independent. Under this assumption, it can be shown that the variance of the sum of two random variables is the sum of the respective variances. Applying this principle,

$$\begin{aligned} Var(L) &= Var(c_1X_1 + \cdots + c_nX_n) \\ &= Var(c_1X_1) + \cdots + Var(c_nX_n) = c_1^2\,Var(X_1) + \cdots + c_n^2\,Var(X_n) \end{aligned}$$

since

$$Var(c_iX_i) = c_i^2\,Var(X_i)$$

5.6 Variance of Linear Combinations of Random Variables

The variance of the linear combination $L = \sum_{i=1}^{n} c_iX_i$, where $X_1, \ldots, X_n$ are independent is $Var(L) = \sum_{i=1}^{n} c_i^2\,Var(X_i)$.

EXAMPLE 5.24 **Renal Disease** Suppose X_1, X_2 are defined as in Example 5.22. If we know that $Var(X_1) = Var(X_2) = 0.25$, then what is the variance of the average serum-creatinine level over these two people?

SOLUTION We wish to compute $Var(0.5X_1 + 0.5X_2)$. Applying **(5.6)**,

$$Var(0.5X_1 + 0.5X_2) = (0.5)^2 Var(X_1) + (0.5)^2 Var(X_2)$$
$$= 0.25(0.25) + 0.25(0.25) = 0.125$$ ■■■

The results for the expected value and variance of linear combinations in **(5.5)** and **(5.6)** *do not* depend on the assumption of normality. However, linear combinations of normal random variables are often of specific concern. It can be shown that any linear combination of normal random variables is itself normally distributed. This leads to the following important result:

5.7 If $X_1, \ldots, X_n$ are independent normal random variables with expected values $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, and L is any linear combination $= \sum_{i=1}^{n} c_i X_i$, then L is normally distributed with

$$\text{expected value} = E(L) = \sum_{i=1}^{n} c_i \mu_i \quad \text{and} \quad \text{variance} = Var(L) = \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

EXAMPLE 5.25 **Renal Disease** If X_1, X_2 are defined as in Examples 5.22–5.24 and are each normally distributed, then what is the distribution of the average $= 0.5X_1 + 0.5X_2$?

SOLUTION Based on the solutions to Examples 5.23 and 5.24, we know that $E(L) = 1.4$, $Var(L) = 0.125$. Therefore, $(X_1 + X_2)/2 \sim N(1.4, 0.125)$. ■■■

SECTION 5.7 Normal Approximation to the Binomial Distribution

In Chapter 4 the binomial distribution was introduced to assess the probability of k successes in n independent trials, where the probability of success (p) is the same for each trial. If n is large, the binomial distribution is very cumbersome to work with and an approximation is easier to use rather than the exact binomial distribution. The normal distribution is often used to approximate the binomial since it is very easy to work with. The key question is, When will the normal distribution provide an accurate approximation to the binomial?

Suppose a binomial distribution has parameters n and p. If n is large and p is either near 0 or near 1, then the binomial distribution will be very positively or negatively skewed, respectively. See Figure 5.15(a) and (b). Similarly, when n is small, for any p, the distribution will tend to be skewed. See Figure 5.15(c). However, if n is moderately large and p is not too extreme, then the binomial distribution will tend to be symmetric and will be well approximated by a normal distribution. See Figure 5.15(d).

We know from Chapter 4 that the mean and variance of a binomial distribution are np and npq, respectively. A natural approximation to use is a normal distribution with the *same* mean and variance, that is, $N(np, npq)$. Suppose we want to compute $Pr(a \leq X \leq b)$ for some integers a, b, where X is binomially distributed with parameters

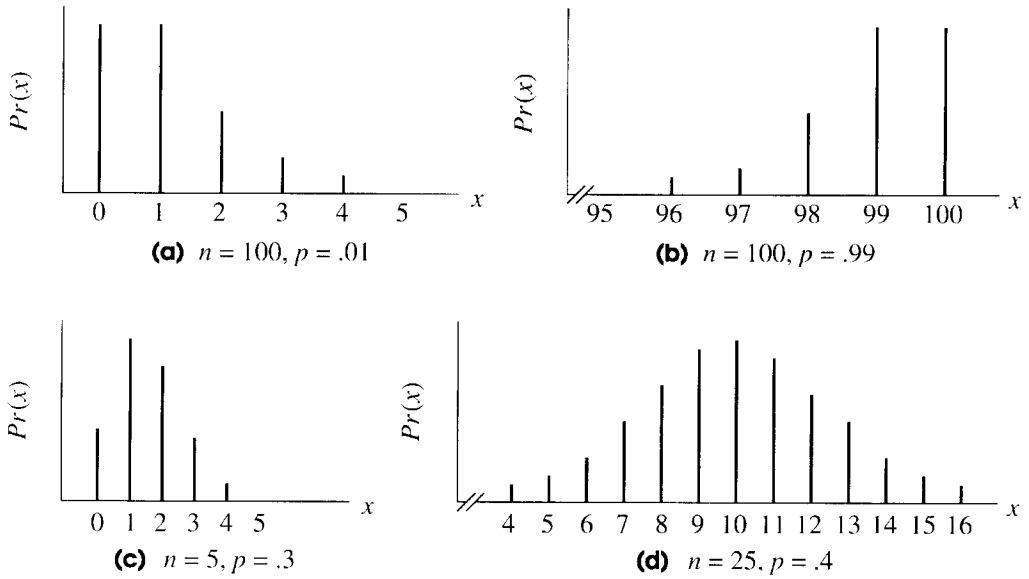


FIGURE 5.15
Symmetry properties of the binomial distribution

n and p. This probability might be approximated by the area under the normal curve from a to b. However, we can show empirically that a better approximation to this probability is given by the area under the normal curve from $a - \frac{1}{2}$ to $b + \frac{1}{2}$. This will generally be the case when any discrete distribution is approximated by the normal distribution. Thus the following rule applies:

5.8 **Normal Approximation to the Binomial Distribution**

If X is a binomial random variable with parameters n and p, then $Pr(a \leq X \leq b)$ is approximated by the area under an $N(np, npq)$ curve from $\left(a - \frac{1}{2}\right)$ to $\left(b + \frac{1}{2}\right)$. This rule implies that for the special case $a = b$, the binomial probability $Pr(X = a)$ is approximated by the area under the normal curve from $\left(a - \frac{1}{2}\right)$ to $\left(a + \frac{1}{2}\right)$. The only exception to this rule is that $Pr(X = 0)$ and $Pr(X = n)$ are approximated by the area under the normal curve to the left of $\frac{1}{2}$ and to the right of $n - \frac{1}{2}$, respectively.

We saw in Equation **(5.7)** that if $X_1, \ldots, X_n$ are independent normal random variables, then any linear combination of these random variables $L = \sum_{i=1}^{n} c_i X_i$ will be normally distributed. In particular, if $c_1 = \cdots = c_n = 1$, then a sum of normal random variables $L = \sum_{i=1}^{n} X_i$ will be normally distributed.

The normal approximation to the binomial distribution is a special case of a very important statistical principle, the central-limit theorem, which is a generalization of Equation **(5.7)**. Under this principle, for large N, a sum of N random variables is approximately normally distributed *even if the individual random variables being summed are not themselves normal.*

DEFINITION 5.12

Let X_i be a random variable that takes on the value 1 with probability p and the value 0 with probability $q = 1 - p$. This type of random variable is defined as a **Bernoulli trial**. This is a special case of a binomial random variable with $n = 1$. ■

We know from the definition of an expected value that $E(X_i) = 1(p) + 0(q) = p$ and that $E(X_i^2) = 1^2(p) + 0^2(q) = p$. Therefore,

$$Var(X_i) = E(X_i^2) - [E(X_i)]^2 = p - p^2 = p(1 - p) = pq$$

Now consider the random variable

$$X = \sum_{i=1}^{n} X_i$$

This random variable simply represents the number of successes among n trials.

EXAMPLE 5.26 Interpret $X_1, \ldots, X_n$ and X in the case of the number of neutrophils among 100 white blood cells (see Example 4.15).

SOLUTION In this case, $n = 100$ and $X_i = 1$ if the i^{th} white blood cell is a neutrophil and $X_i = 0$ if the i^{th} white blood cell is not a neutrophil, where $i = 1, \ldots, 100$. X represents the number of neutrophils among $n = 100$ white blood cells. ■■■

Given Equations **(5.5)** and **(5.6)**, we know that

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = p + p + \cdots + p = np$$

and

$$Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = pq + pq + \cdots + pq = npq$$

Based on the normal approximation to the binomial distribution, we approximate the distribution of X by a normal distribution with mean $= np$ and variance $= npq$. We discuss the central-limit theorem in more detail in section 6.5.3.

EXAMPLE 5.27 Suppose a binomial distribution has parameters $n = 25$, $p = .4$. How can $Pr(7 \leq X \leq 12)$ be approximated?

SOLUTION We have $np = 25(.4) = 10$, $npq = 25(.4)(.6) = 6.0$. Thus, this distribution is approximated by a normal random variable Y with mean 10 and variance 6. We specifically want to compute the area under this normal curve from 6.5 to 12.5. We have

$$\begin{aligned} Pr(6.5 \leq Y \leq 12.5) &= \Phi\left(\frac{12.5 - 10}{\sqrt{6}}\right) - \Phi\left(\frac{6.5 - 10}{\sqrt{6}}\right) \\ &= \Phi(1.02) - \Phi(-1.43) = \Phi(1.02) - [1 - \Phi(1.43)] \\ &= \Phi(1.02) + \Phi(1.43) - 1 = .8463 + .9235 - 1 = .770 \end{aligned}$$

This approximation is depicted in Figure 5.16. ■■■

EXAMPLE 5.28 **Infectious Disease** Suppose we want to compute the probability that between 50 and 75 of 100 white blood cells will be neutrophils, where the probability that any one cell is a neutrophil is .6. These values are chosen as proposed limits to the range of neutrophils in normal people and we wish to predict what proportion of people will be in the normal range according to this definition.

SOLUTION The exact probability is given by

$$\sum_{k=50}^{75} \binom{100}{k} .6^k .4^{100-k}$$

The normal approximation is used to approximate the exact probability. The mean of the binomial distribution in this case is 100(.6) = 60, and the variance is 100(.6)(.4) = 24. Thus, we find the area between 49.5 and 75.5 for an $N(60, 24)$ distribution. This area is

$$\begin{aligned}\Phi\left(\frac{75.5 - 60}{\sqrt{24}}\right) - \Phi\left(\frac{49.5 - 60}{\sqrt{24}}\right) &= \Phi(3.16) - \Phi(-2.14) \\ &= \Phi(3.16) + \Phi(2.14) - 1 \\ &= .9992 + .9840 - 1 = .983\end{aligned}$$

Thus, 98.3% of the people will be normal. ■■■

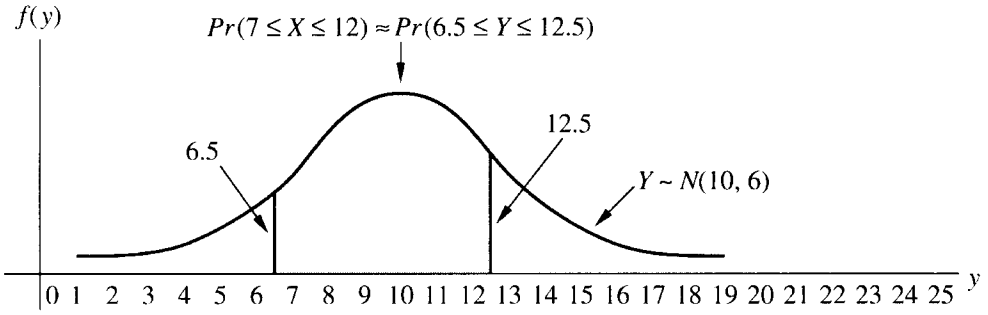


FIGURE 5.16 The approximation of the binomial random variable X with parameters $n = 25$, $p = .4$ by the normal random variable Y with mean 10 and variance 6

EXAMPLE 5.29 **Infectious Disease** Suppose a person is defined as abnormally high if the number of neutrophils is ≥ 76 and abnormally low if the number of neutrophils is ≤ 49. Calculate the proportion of people that are abnormally high and low.

SOLUTION The probability of being abnormally high is given by $Pr(X \geq 76) \approx Pr(Y \geq 75.5)$, where X is a binomial random variable with parameters $n = 100$, $p = .6$ and $Y \sim N((60, 24)$. This latter probability is

$$1 - \Phi\left(\frac{75.5 - 60}{\sqrt{24}}\right) = 1 - \Phi(3.16) = .001$$

Similarly, the probability of being abnormally low is

$$\begin{aligned}Pr(X \leq 49) \approx Pr(Y \leq 49.5) &= \Phi\left(\frac{49.5 - 60}{\sqrt{24}}\right) \\ &= \Phi(-2.14) = 1 - \Phi(2.14) \\ &= 1 - .9840 = .016\end{aligned}$$

Thus, 0.1% of people will be abnormally high and 1.6% will be abnormally low. These probabilities are depicted in Figure 5.17. ■■■

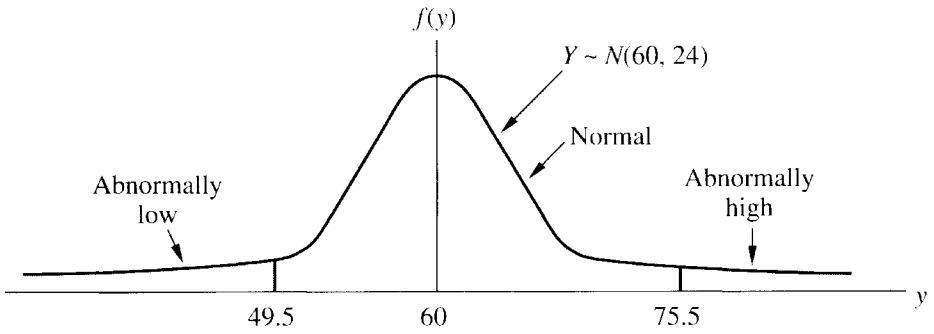


FIGURE 5.17
Normal approximation to the distribution of neutrophils

Under what conditions should this approximation be used?

The normal distribution is used with mean np and variance npq to approximate a binomial distribution with parameters n and p when $npq \geq 5$.

This condition will be satisfied if n is moderately large and p is not too small. To illustrate this condition, the binomial probability distribution for $p = .1$, $n = 10$, 20, 50, and 100 is plotted in Figure 5.18(a) through (d) and $p = .2$, $n = 10$, 20, 50, and 100 is plotted in Figure 5.19(a) through (d) using a Statistical Analysis System (SAS) plotting routine (PROC PLOT).

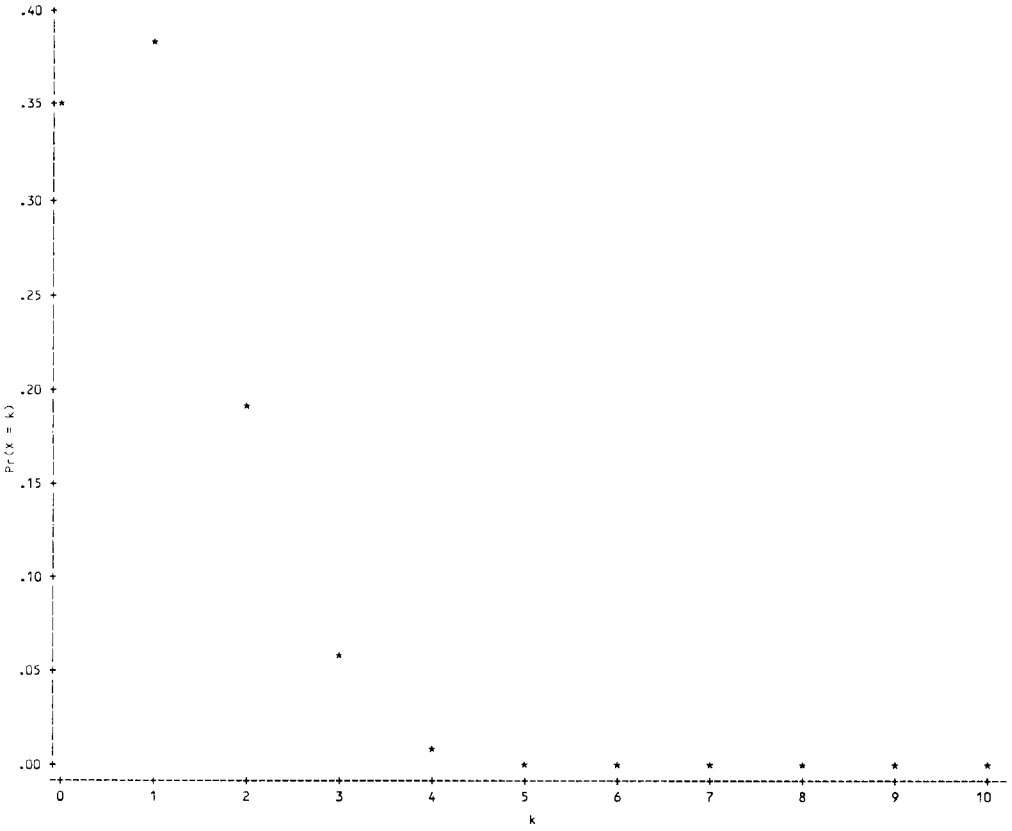
Notice that the normal approximation to the binomial distribution does not fit well in Figure 5.18(a), $n = 10$, $p = .1$ ($npq = 0.9$), or Figure 5.18(b), $n = 20$, $p = .1$ ($npq = 1.8$). The approximation is marginally adequate in Figure 5.18(c), $n = 50$, $p = .1$ ($npq = 4.5$), where the right-hand tail is only slightly longer than the left-hand tail. The approximation is quite good in Figure 5.18(d), $n = 100$, $p = .1$ ($npq = 9.0$), where the distribution appears to be quite symmetric. Similarly, for $p = .2$, although the normal approximation is not good for $n = 10$ [Figure 5.19(a), $npq = 1.6$], it becomes marginally adequate for $n = 20$ [Figure 5.19(b), $npq = 3.2$] and quite good for $n = 50$ [Figure 5.19(c), $npq = 8.0$] and $n = 100$ [Figure 5.19(d), $npq = 16.0$].

Note that the conditions under which the normal approximation to the binomial distribution works well (namely, $npq \geq 5$), which correspond to n moderate and p not too large or too small, are generally *not* the same as the conditions for which the Poisson approximation to the binomial distribution works well [n large (≥ 100) and p very small ($p \leq .01$)]. However, occasionally both of these criteria will be met. In such cases, for example, when $n = 1000$, $p = .01$, the two approximations will yield about the same results. The normal approximation is preferable because it is easier to apply.

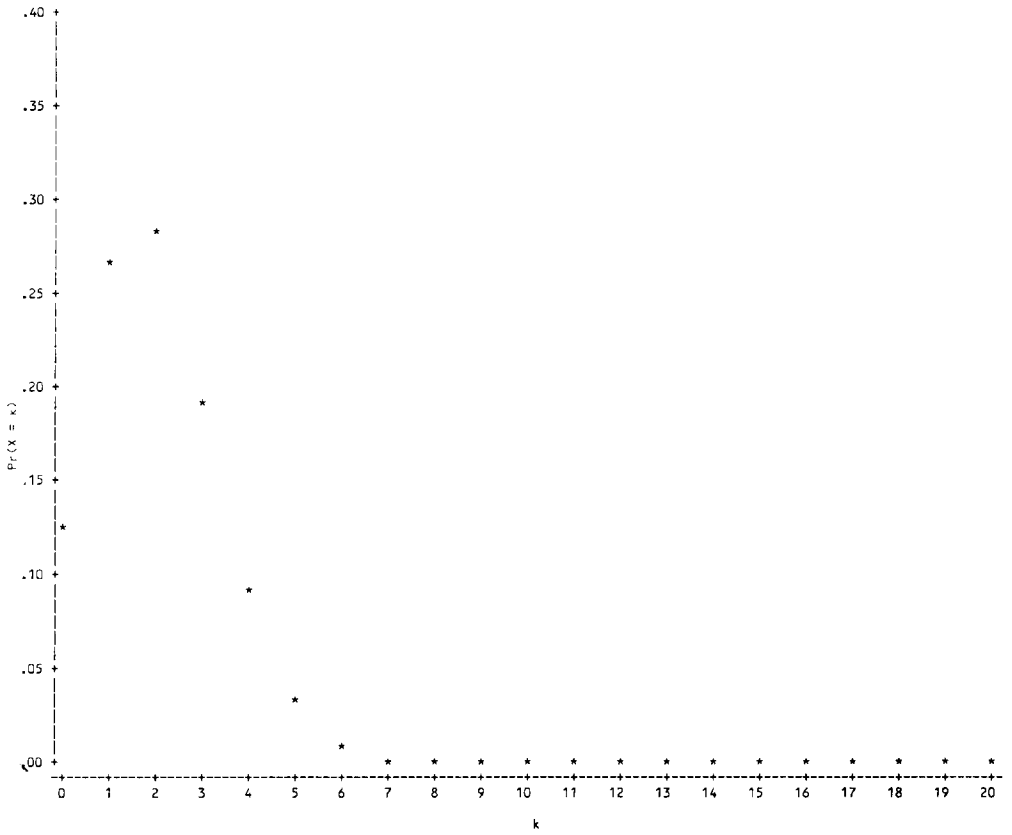
SECTION 5.8 Normal Approximation to the Poisson Distribution

The normal distribution can also be used to approximate discrete distributions other than the binomial distribution, particularly the Poisson distribution. The motivation for this is that the Poisson distribution is cumbersome to use for large values of μ.

FIGURE 5.18(a)(b)
SAS plot of binomial distribution

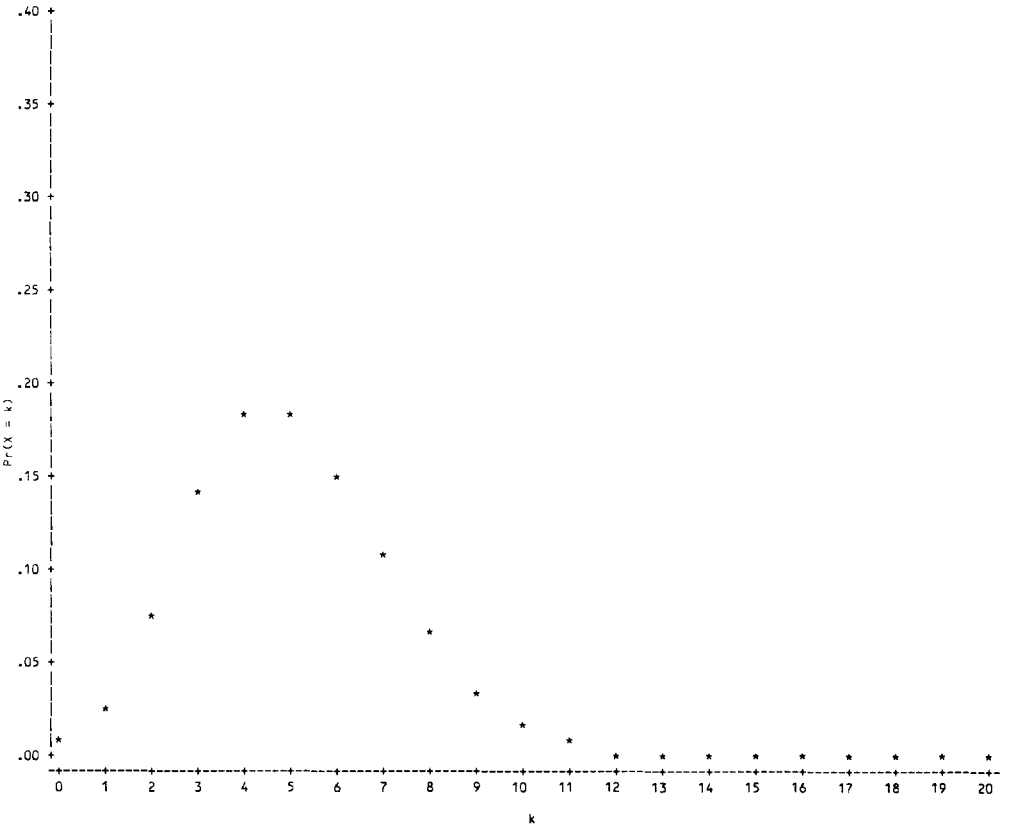


(a) $n = 10$, $p = .1$

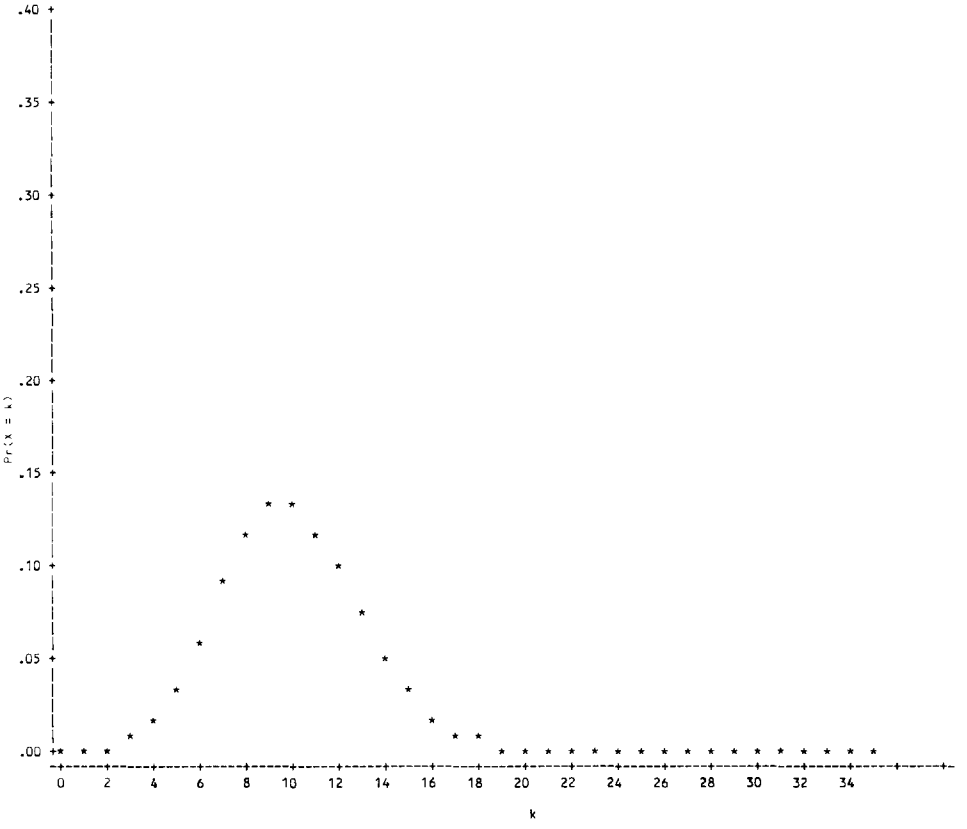


(b) $n = 20$, $p = .1$

FIGURE 5.18(c)(d)
SAS plot of binomial distribution

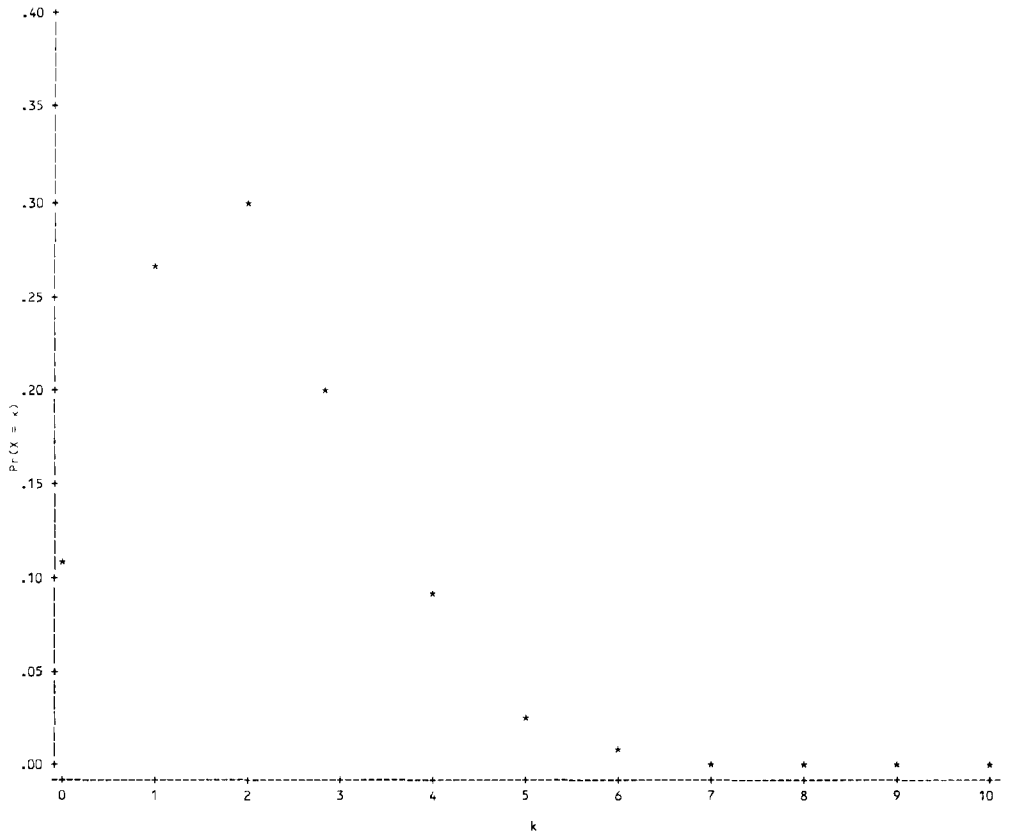


(c) $n = 50$, $p = .1$

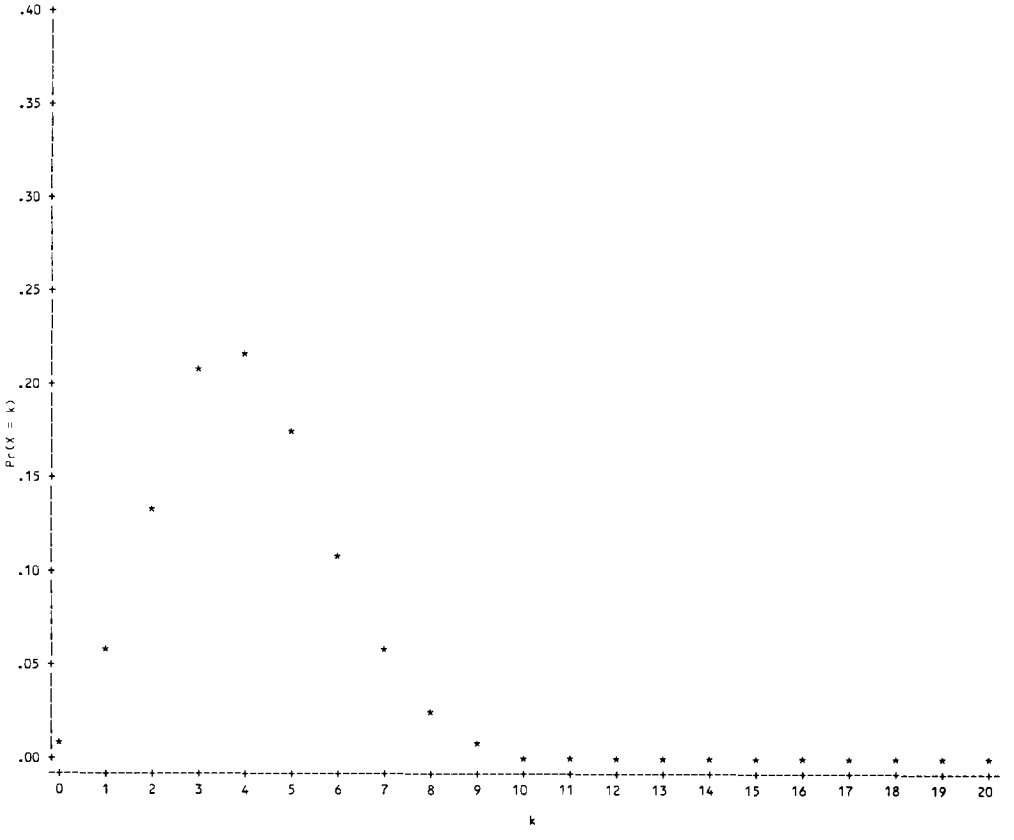


(d) $n = 100$, $p = .1$

FIGURE 5.19(a)(b)
SAS plot of binomial distribution

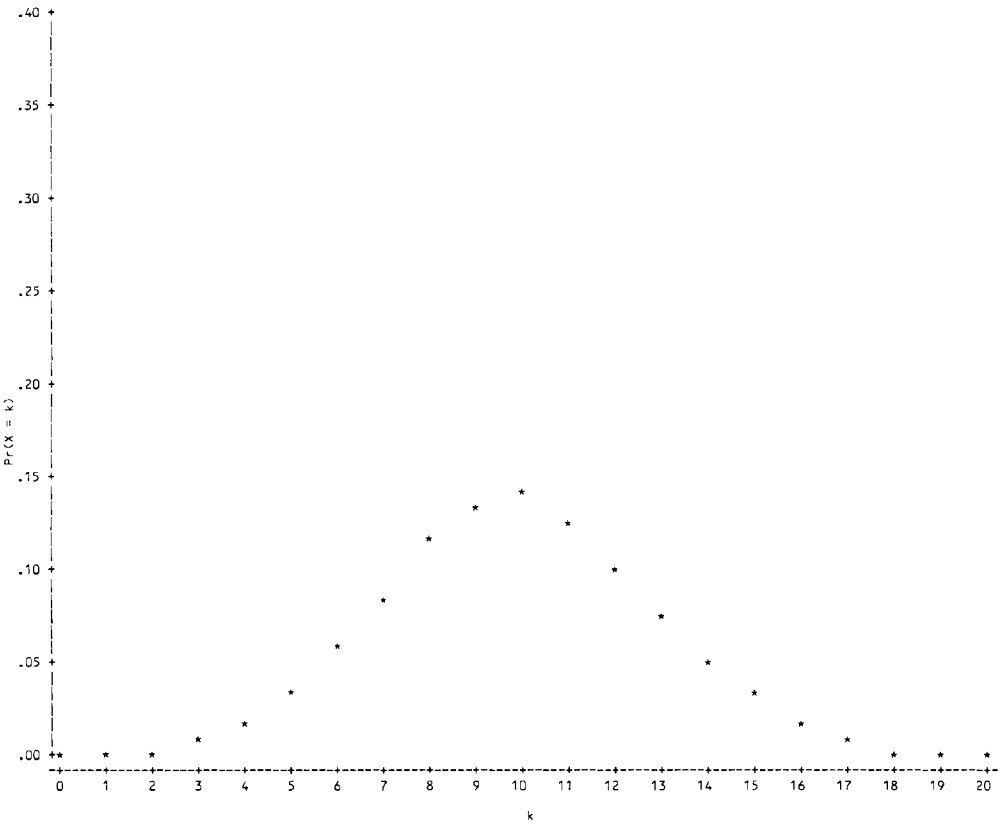


(a) $n = 10$, $p = .2$

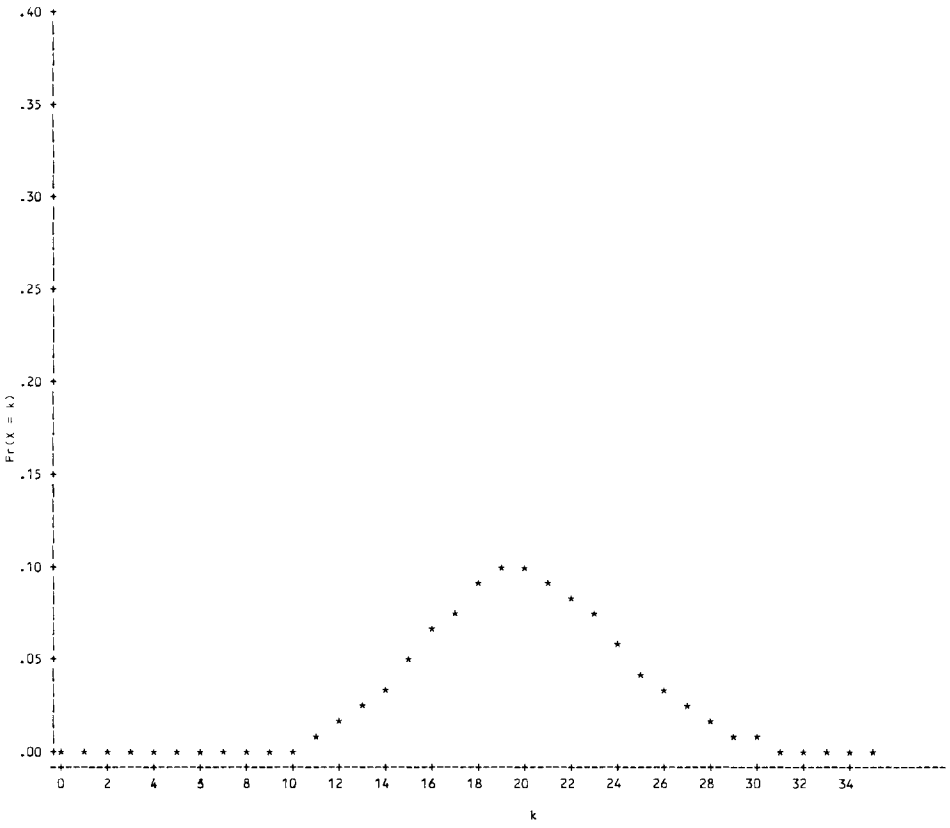


(b) $n = 20$, $p = .2$

FIGURE 5.19(c)(d)
SAS plot of binomial distribution



(c) $n = 50$, $p = .2$



(d) $n = 100$, $p = .2$

The same technique is used as for the binomial distribution; that is, the means and variances of the Poisson distribution and the approximating normal distribution are equated.

5.9 Normal Approximation to the Poisson Distribution

A Poisson distribution with parameter μ is approximated by a normal distribution with mean and variance both equal to μ. $Pr(X = x)$ is approximated by the area under an $N(\mu, \mu)$ density from $x - \frac{1}{2}$ to $x + \frac{1}{2}$ for $x > 0$ or by the area to the left of $\frac{1}{2}$ for $x = 0$. This approximation is used for $\mu \geq 10$.

The Poisson distribution for $\mu = 2, 5, 10$, and 20 is plotted using the SAS plotting program (PROC PLOT) in Figure 5.20(a) through (d), respectively. The normal approximation is clearly inadequate for $\mu = 2$ [Figure 5.20(a)], marginally adequate for $\mu = 5$ [Figure 5.20(b)], and adequate for $\mu = 10$ [Figure 5.20(c)] and $\mu = 20$ [Figure 5.20(d)].

EXAMPLE 5.30 **Bacteriology** Consider again the distribution of the number of bacteria in a petri plate of area A. Assume that the probability of observing x bacteria is given exactly by a Poisson distribution with parameter $\mu = \lambda A$, where $\lambda = 0.1$ and $A = 100$ cm^2. Suppose 20 bacteria are observed in this area. How unusual is this event?

SOLUTION Compute $Pr(X \geq 20) \approx Pr(Y \geq 19.5)$

where $Y \sim N(\lambda A, \lambda A) = N(10, 10)$

We have
$$Pr(Y \geq 19.5) = 1 - Pr(Y \leq 19.5) = 1 - \Phi\left(\frac{19.5 - 10}{\sqrt{10}}\right)$$
$$= 1 - \Phi\left(\frac{9.5}{\sqrt{10}}\right) = 1 - \Phi(3.00)$$
$$= 1 - .9987 = .0013$$

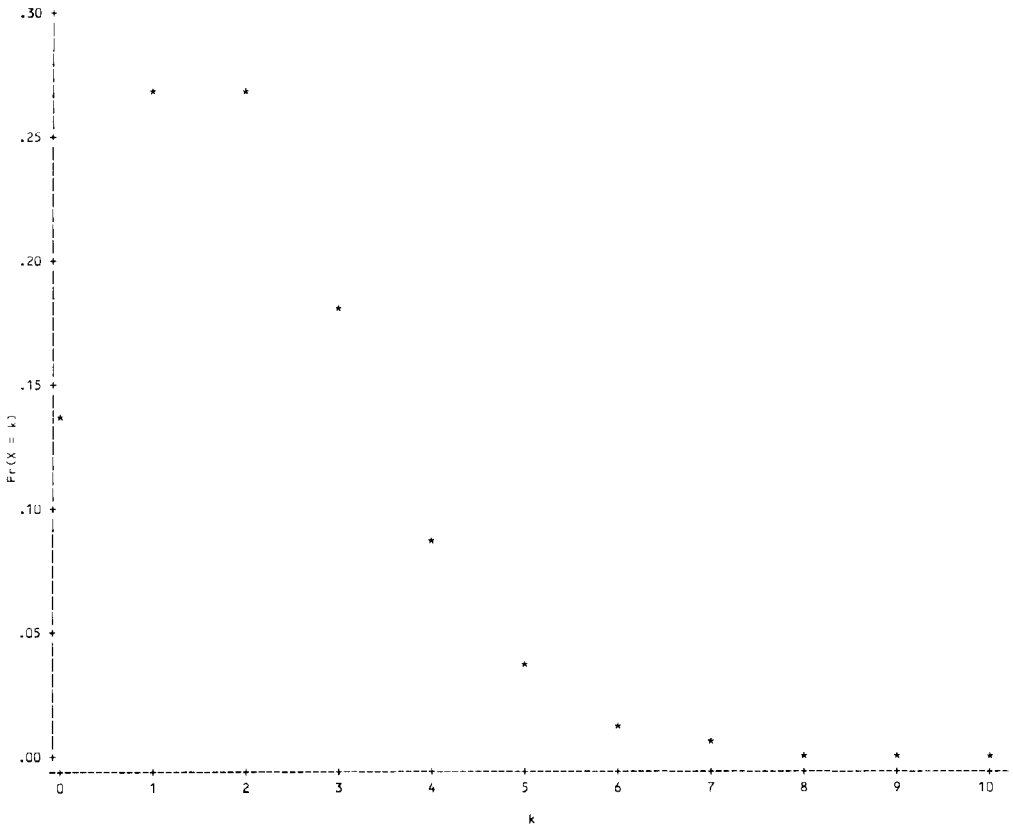
Thus, 20 or more colonies in 100 cm^2 would be expected only 1.3 times in 1000 plates, a rare event indeed. ■■■

SECTION 5.9 Summary

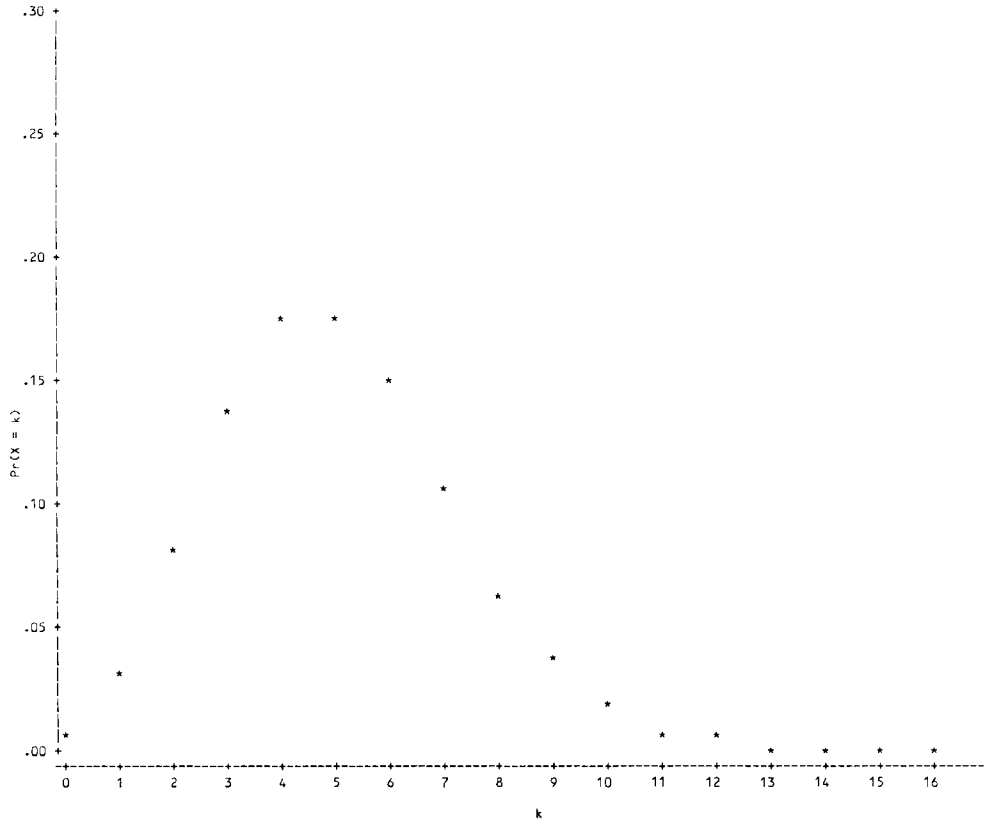
In this chapter continuous random variables were discussed. The concept of a probability density function, which is the analogue to a probability mass function for discrete random variables, was introduced. In addition, generalizations of the concepts of expected value, variance, and cumulative distribution were presented for continuous random variables.

The normal distribution, the most important continuous distribution, was then studied in detail. The normal distribution is often used in statistical work, since many random phenomena follow this probability law, particularly those that can be expressed as a sum of many random variables. It was shown that the normal distribution is

FIGURE 5.20(a)(b)
SAS plot of Poisson distribution

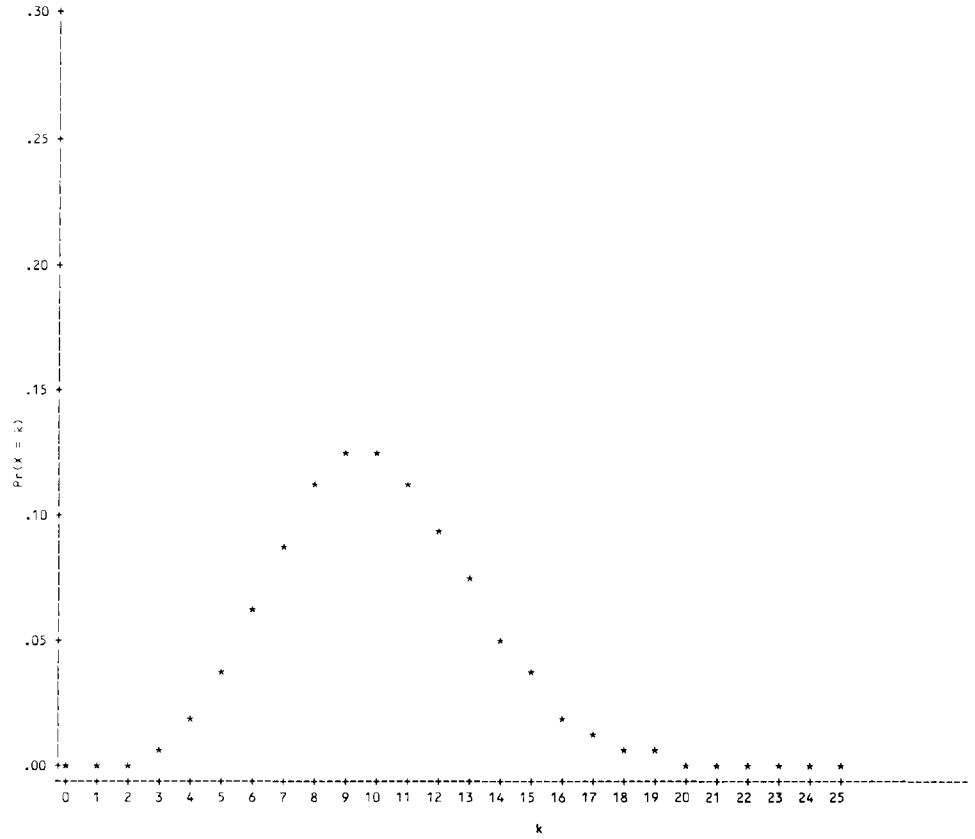


(a) $\mu = 2$

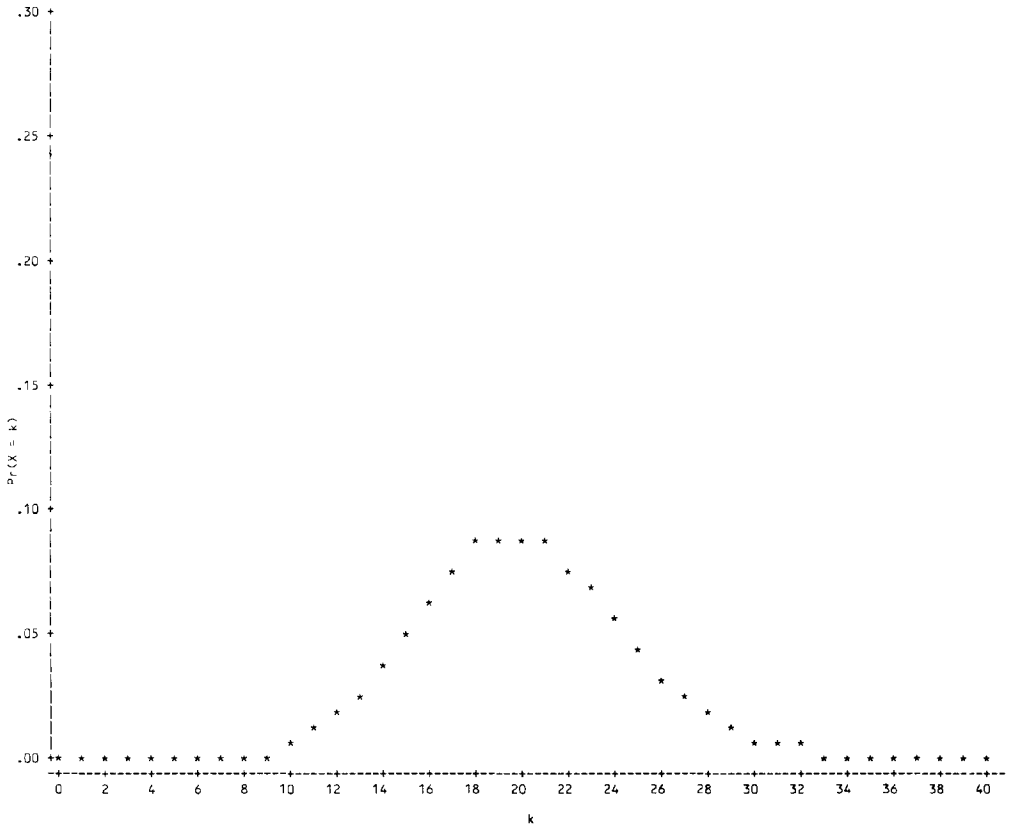


(b) $\mu = 5$

FIGURE 5.20(c)(d)
SAS plot of Poisson distribution



(c) $\mu = 10$



(d) $\mu = 20$

indexed by two parameters, the mean μ and the variance σ^2. Fortunately, all computations concerning any normal random variable can be accomplished using the standard, or unit, normal probability law, which has mean 0 and variance 1. Normal tables were introduced to use when working with the standard normal distribution. Also, since the normal distribution is easy to use, it is often employed to approximate other distributions. In particular, the focus was on the normal approximations to the binomial and Poisson distributions. This is a special case of the central-limit theorem, which is covered in more detail in Chapter 6. Also, to facilitate applications of the central-limit theorem, the properties of linear combinations of random variables were discussed.

In the next three chapters, the normal distribution is used extensively as a foundation for work on statistical inference.

PROBLEMS

Cardiovascular Disease

Since serum cholesterol is related to age and sex, some investigators prefer to express it in terms of z-scores. If X = raw serum cholesterol, then $z = \dfrac{X - \mu}{\sigma}$, where μ is the mean and σ is the standard deviation of serum cholesterol for a given age-sex group. Suppose z is regarded as a standard normal distribution.

* **5.1** What is $Pr(z < 0.5)$?

* **5.2** What is $Pr(z > 0.5)$?

* **5.3** What is $Pr(-1.0 < z < 1.5)$?

Suppose a person is regarded as having high cholesterol if $z > 2.0$ and borderline cholesterol if $1.5 < z < 2.0$.

* **5.4** What proportion of people have high cholesterol?

* **5.5** What proportion of people have borderline cholesterol?

Nutrition

Suppose that total carbohydrate intake in 12–14-year-old males is normally distributed with mean 124 g/1000 cal and standard deviation 20 g/1000 cal.

5.6 What percentage of boys in this age range have carbohydrate intake above 140 g/1000 cal?

5.7 What percentage of boys in this age range have carbohydrate intake below 90 g/1000 cal?

Suppose boys in this age range that live below the poverty level have a mean carbohydrate intake of 121 g/1000 cal with a standard deviation of 19 g/1000 cal.

5.8 Answer Problem 5.6 for boys in this age range and economic environment.

5.9 Answer Problem 5.7 for boys in this age range and economic environment.

Diabetes

A number of clinical characteristics were ascertained in a large group of subjects with insulin-dependent diabetes mellitus (IDDM). Suppose the distribution of percentage of ideal body weight in this group of patients is normal with mean 110 and standard deviation of 13.

5.10 What percentage of subjects with IDDM are above their ideal body weight, i.e., above 100% ideal body weight?

5.11 What percentage of subjects with IDDM are overweight (defined as 10% or more above ideal body weight)?

5.12 What percentage of subjects with IDDM are obese (defined as 20% or more above ideal body weight)?

5.13 What percentage of subjects with IDDM are underweight (defined as 10% or more below ideal body weight)?

5.14 What percentage of subjects with IDDM have normal body weight (within 10% of ideal body weight)?

Pulmonary Disease

Many investigators have studied the relationship between asbestos exposure and death due to chronic obstructive pulmonary disease (COPD).

5.15 Suppose that among workers exposed to asbestos in a shipyard in 1980, 33 died over a 10-year period from COPD, whereas only 24 such deaths could be expected based on statewide mortality rates. Is the number of deaths due to COPD in this group excessive?

5.16 Twelve cases of leukemia are reported in people living in a particular census tract over a 5-year period. Is this number of cases abnormal if only 6.7 cases would be expected on national cancer-incidence rates?

Cardiovascular Disease, Pulmonary Disease

The duration of cigarette smoking has been linked to many diseases, including lung cancer and various forms of heart disease. Suppose we know that among men aged 30–34 who have ever smoked, the mean number of years they smoked is 12.8 with a standard deviation of 5.1 years. For women in this age group, the mean number of years they smoked is 9.3 with a standard deviation of 3.2.

* **5.17** Assuming that the duration of smoking is normally distributed, what proportion of men in this age group have smoked for more than 20 years?

* **5.18** Answer Problem 5.17 for women.

Cancer

Previous census data have indicated that approximately 0.2% of women aged 45–54 will have had cervical cancer at some point in their lives. However, the general feeling is that the rate of cervical cancer has decreased.

5.19 If a new study by mail questionnaire is performed and it is found that 100 out of 100,000 women have had cervical cancer, then is this proportion consistent with the census rate?

Cardiovascular Disease

Serum cholesterol is an important risk factor for coronary disease. We can show that serum cholesterol is approximately normally distributed with mean 219 mg%/mL and standard deviation 50 mg%/mL.

* **5.20** If the clinically desirable range for cholesterol is < 200 mg%/mL, then what proportion of people have clinically desirable levels of cholesterol?

* **5.21** Some investigators feel that only cholesterol levels of over 250 mg%/mL indicate a high-enough risk for heart disease to warrant treatment. What proportion of the population does this group represent?

* **5.22** What proportion of the general population have borderline high-cholesterol levels—that is, > 200, but < 250 mg%/mL?

Nutrition, Cancer

Beta carotene is a substance that is hypothesized to prevent cancer. A dietary survey was undertaken for the purpose of measuring the level of beta carotene intake in the typical American diet. Assume that the distribution of ln carotene is normal with mean 8.34 and standard deviation 1.00. (Units are in ln IU.)

5.23 What percentage of people have dietary carotene levels below 2000 IU? (*Note:* ln 2000 = 7.60.)

5.24 What percentage of people have dietary carotene levels below 1000 IU? (*Note:* ln 1000 = 6.91.)

5.25 Some studies suggest that carotene levels over 10,000 IU may protect against cancer. What percentage of people have a dietary intake of at least 10,000 IU?

Suppose that each person took a carotene supplement pill of dosage 5000 IU in addition to his or her normal diet. Assume that the resulting distribution of ln carotene is normally distributed with mean 9.12 and standard deviation 1.00.

5.26 What percentage of people would have an intake from diet and supplements of at least 10,000 IU?

Hypertension

People are classified as hypertensive if their systolic blood pressure is higher than a specified level for their age group, according to the scheme in Table 5.1.

TABLE 5.1 Mean and standard deviation of systolic blood pressure (mm Hg) in specific age groups

Age group	Mean	Standard deviation	Specified hypertension level
1–14	105.0	5.0	115.0
15–44	125.0	10.0	140.0

Assume that systolic blood pressure is normally distributed with mean and standard deviation given in Table 5.1 for the age groups 1–14 and 15–44, respectively. Define a *family* as a group of 2 people in the age group 1–14 and 2 people in the age group 15–44. A family is classified as hypertensive if *any one* family member is hypertensive.

* **5.27** What proportion of 1–14-year-olds are hypertensive?

* **5.28** What proportion of 15–44-year-olds are hypertensive?

* **5.29** What proportion of families are hypertensive? (Assume that the hypertensive status of different members of the family are independent random variables.)

* **5.30** Suppose an apartment building has 200 families living in it. What is the probability that between 10 and 25 families are hypertensive?

Pulmonary Disease

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. FEV is known to be influenced by age, sex, and cigarette smoking. Assume that in 45–54-year-old nonsmoking males FEV is normally distributed with mean 4.0 liters and standard deviation 0.5 liter.

In comparably aged currently smoking males FEV is normally distributed with mean 3.5 liters and standard deviation 0.6 liter.

5.31 If an FEV of less than 2.5 liters is regarded as showing some functional impairment (occasional breathlessness, inability to climb stairs, etc.), then what is the probability that a currently smoking male has functional impairment?

5.32 Answer Problem 5.31 for a nonsmoking male.

Many people are not functionally impaired now but their pulmonary function usually declines with age and they eventually will be functionally impaired. Assume that the *decline* in FEV over n years is normally distributed with mean $0.03n$ and standard deviation $0.02n$.

5.33 What is the probability that a 45-year-old man with an FEV of 4.0 liters will be functionally impaired by the age of 75?

5.34 Answer Problem 5.33 for a 25-year-old man with an FEV of 4.0 liters.

Infectious Disease

The differential is a standard measurement made during a blood test. It consists of classifying white blood cells into the following 5 categories: (1) basophils, (2) eosinophils, (3) monocytes, (4) lymphocytes, and (5) neutrophils. The usual practice is to look at 100 randomly selected cells under a microscope and count the number of cells within each of the 5 categories. Assume that a normal adult will have the following proportions of cells in each category: basophils, 0.5%; eosinophils, 1.5%; monocytes, 4%; lymphocytes, 34%; and neutrophils, 60%.

* **5.35** An excess of eosinophils is sometimes consistent with a violent allergic reaction. What is the exact probability that a normal adult will have 5 or more eosinophils?

* **5.36** An excess of lymphocytes is consistent with various forms of viral infection, such as hepatitis. What is the probability that a normal adult will have 40 or more lymphocytes?

* **5.37** What is the probability that a normal adult will have 50 or more lymphocytes?

* **5.38** How many lymphocytes would have to appear in the differential before you would feel that the "normal" pattern was violated?

* **5.39** An excess of neutrophils is consistent with several types of bacterial infection. Suppose an adult has x neutrophils. How large would x have to be in order that the probability of a normal adult having x or more neutrophils was $\leq 5\%$?

* **5.40** How large would x have to be in order that the probability of a normal adult having x or more neutrophils was $\leq 1\%$?

Blood Chemistry

In pharmacologic research a variety of clinical chemistry measurements are routinely monitored closely for evidence of side effects of the medication under study. Suppose typical blood-glucose levels are normally distributed with mean 90 mg/dL and standard deviation 38 mg/dL.

5.41 If the normal range is from 65–120 mg/dL, then what percentage of values will fall in the normal range?

5.42 In some studies only values that are at least 1.5 times as high as the upper limit of normal are identified as abnormal. What percentage of values would fall in this range?

5.43 Answer Problem 5.42 for 2.0 times the upper limit of normal.

5.44 Frequently, tests that yield abnormal results are repeated for confirmation. What is the probability that for a normal person a test will be at least 1.5 times as high as the upper limit of normal on two separate occasions?

5.45 Suppose that in a pharmacologic study involving 6000 patients, 75 patients have blood-glucose levels at least 1.5 times the upper limit of normal on one occasion. What is the probability that this result could be due to chance?

Cancer

A treatment trial is proposed to test the efficacy of vitamin E as a preventive agent for cancer. One problem with such a study is how to assess compliance among study participants. A small pilot study is undertaken to establish criteria for compliance with the proposed study agents. In this regard, 10 patients are given 400 IU/day of vitamin E and 10 patients are given similar-sized tablets of placebo over a 3-month period. Their serum vitamin-E levels are measured before and after the 3-month period and the change (3-month − baseline) is shown in Table 5.2.

TABLE 5.2 Change in serum vitamin E(mg/dL) in pilot study

Group	Mean	sd	*n*
Vitamin E	0.80	0.48	10
Placebo	0.05	0.16	10

* **5.46** Suppose a change of 0.30 mg/dL in serum levels is proposed as a test criterion for compliance; that is, a

patient who shows a change of ≥ 0.30 mg/dL is considered a compliant vitamin-E taker. If normality is assumed, what percentage of the vitamin-E group would be expected to show a change of at least 0.30 mg/dL?

* **5.47** Is the measure in Problem 5.46 a measure of sensitivity, specificity, or predictive value?

* **5.48** What percentage of the placebo group would be expected to show a change of not more than 0.30 mg/dL?

* **5.49** Is the measure in Problem 5.48 a measure of sensitivity, specificity, or predictive value?

* **5.50** Suppose a new threshold of change, Δ mg/dL, is proposed for establishing compliance. We wish to use a level of Δ such that the compliance measures in Problems 5.46 and 5.48 for the patients in the vitamin-E and placebo groups are the same. What should Δ be? What would be the compliance in the vitamin-E and placebo groups using this threshold level?

Mental Health

5.51 Refer to Tables 3.2 and 3.3. Suppose a study of Alzheimer's disease is planned in more than one retirement community. How many retired people need to be studied to have a 90% chance of detecting at least 100 people with Alzheimer's disease, assuming that the age-sex-specific prevalence rates and age-sex distribution in Tables 3.2 and 3.3 hold?

5.52 Answer Problem 5.51 for 50 rather than 100 people.

Pulmonary Disease

Refer to the pulmonary function data in the Data Set FEV.DAT on the data disk (see Problem 2.21). We are interested in whether there is a relationship between smoking status and level of pulmonary function. However, FEV is affected by age and sex; also, smoking children tend to be older than nonsmoking children. For these reasons, FEV should be standardized for age and sex. To accomplish this, use the z-score approach outlined above in Problem 5.1, where the z-scores here are defined by age-sex groups.

5.53 Plot the distribution of z-scores for smokers and nonsmokers separately. Do these distributions look normal? Does there appear to be any relationship between smoking and pulmonary function in these data?

5.54 Repeat the analyses in Problem 5.53 for the subgroup of children 10+ years of age (since smoking is very rare prior to this age). Do you reach similar conclusions?

5.55 Repeat the analyses in Problem 5.54 separately for boys and girls. Are your conclusions the same in the two groups?

Note: Formal methods for comparing FEV's between smokers and nonsmokers are discussed in the material on statistical inference in Chapter 8.

Cardiovascular Disease

A clinical trial was conducted to test the efficacy of nifedipine, a new drug for stopping chest pain in patients with angina severe enough to require hospitalization. The duration of the study was 14 days in the hospital unless the patient was withdrawn prematurely from therapy, was discharged from the hospital, or died prior to this time. Patients were randomly assigned to either nifedipine or propranolol and were given the same dosage of each drug in identical capsules at level 1 of therapy. If pain did not cease at this level of therapy, or if pain recurred after a period of pain cessation, then the patient progressed to level 2, whereby the dosage of each drug was increased according to a prespecified schedule. Similarly, if pain continued or recurred at level 2, then the patient progressed to level 3, whereby the dosage of the anginal drug was increased again. Patients randomized to either group were allowed to receive nitrates in any amount that was deemed clinically appropriate to help control pain.

The main objective of the study was to compare the degree of pain relief with nifedipine and propranolol. A secondary objective was to better understand the effects of these agents on other physiologic parameters including heart rate and blood pressure. Data on these latter parameters are given in the Data Set NIFED.DAT (on the data disk); the format of this file is given in Table 5.3.

5.56 Describe the effect of each treatment regimen on changes in heart rate and blood pressure. Do the distribution of changes in these parameters look normal or not?

5.57 Compare graphically the effects of the treatment regimens on heart rate and blood pressure. Do you notice any difference between treatments?
(*Note:* Formal tests for comparing changes in heart rate and blood pressure in the two treatment groups are covered in Chapter 8.)

Hypertension

It is well known that there are racial differences in blood pressure between white and black adults. These differences generally do not exist between white and black children. Since aldosterone levels have been related to blood-pressure levels in adults in previous research, an investigation was performed to look at aldosterone levels among black and white children [1].

TABLE 5.3 Format of NIFED.DAT

Column	Variable	Code
1–2	ID	
4	Treatment group	N = nifedipine/ P = propanolol
6–8	Baseline heart rate[a]	beats/min
10–12	Level 1 heart rate[b]	beats/min
14–16	Level 2 heart rate	beats/min
18–20	Level 3 heart rate	beats/min
22–24	Baseline systolic bp[a]	mm Hg
26–28	Level 1 systolic bp[b]	mm Hg
30–32	Level 2 systolic bp	mm Hg
34–36	Level 3 systolic bp	mm Hg

[a]Immediately prior to randomization
[b]Highest heart rate and systolic bp at baseline and each level of therapy, respectively
Note: Missing values indicate that either:
(1) the patient withdrew from the study prior to entering this level of therapy;
(2) the patient achieved pain relief prior to reaching this level of therapy; or,
(3) the patient encountered this level of therapy, but this particular piece of data was missing.

* **5.58** If the mean plasma-aldosterone level in black children was 230 pmol/L with sd = 203 pmol/L, then what percentage of black children have levels $\leq$ 300 pmol/L if normality is assumed?

* **5.59** If the mean plasma-aldosterone level in white children is 400 pmol/L with sd = 218 pmol/L, then what percentage of white children have levels $\leq$ 300 pmol/L if normality is assumed?

* **5.60** The distribution of plasma-aldosterone concentration in 53 white and 46 black children is shown in Figure 5.21. Does the assumption of normality seem reasonable? Why or why not? (*Hint:* Qualitatively compare the observed number of children who have levels below 300 pmol/L with the expected number in each group under the assumption of normality.)

Hepatic Disease

Suppose we observe 84 alcoholics with cirrhosis of the liver, of whom 29 have hepatomas—that is, liver-cell carcinoma. Suppose we know, based on a large sample, that the risk of hepatoma among alcoholics without cirrhosis of the liver is 24%.

5.61 What is the probability that we observe exactly 29 alcoholics with cirrhosis of the liver who have hepatomas if the true rate of hepatoma among alcoholics (with or without cirrhosis of the liver) is .24?

5.62 What is the probability of observing at least 29 hepatomas among the 84 alcoholics with cirrhosis of the liver under the assumptions in Problem 5.61?

5.63 What is the smallest number of hepatomas that would have to be observed among the group of alcoholics with cirrhosis of the liver in order for the hepatoma experience in this group to be different from the hepatoma experience among alcoholics without cirrhosis of the liver? (*Hint:* Use a 5% probability of getting a result at least as extreme to denote differences between the hepatoma experiences of the 2 groups.)

Hypertension

The Pediatric Task Force Report on Blood Pressure Control in Children [2] reports blood-pressure norms for children by age and sex group. The mean $\pm$ standard deviation for 17-year-old boys for diastolic blood pressure is 63.7 $\pm$ 11.4 mm Hg, based on a large sample.

5.64 One approach for defining elevated blood pressure is to use 90 mm Hg—the standard for elevated adult diastolic blood pressure—as the cutoff. What percentage of 17-year-old boys would have elevated blood pressure using this approach?

5.65 Suppose there are 2000 17-year-old boys in the 11th grade, of whom 25 have elevated blood pressure using the criteria in Problem 5.64. Is this an unusually high number of boys with elevated blood pressure? Why or why not?

Environmental Health

5.66 A study was conducted relating particulate air pollution and daily mortality in Steubenville, Ohio [3]. On average over the last 10 years there have been 3 deaths per day. Suppose that on 90 high-pollution days—days where the total suspended particulates are in the highest quartile among all days—the death rate is 3.2 deaths per day, or 288 deaths observed over the 90 high-pollution days. Are there an unusual number of deaths on high-pollution days?

Refer to the Data Set VALID.DAT (on the data disk) described in Table 2.16.

5.67 Consider the nutrients saturated fat, total fat, and total calories. Plot the distribution of each nutrient for both the diet record (DR) and the food-frequency questionnaire (FFQ). Do you think a normal distribution is appropriate for these nutrients?

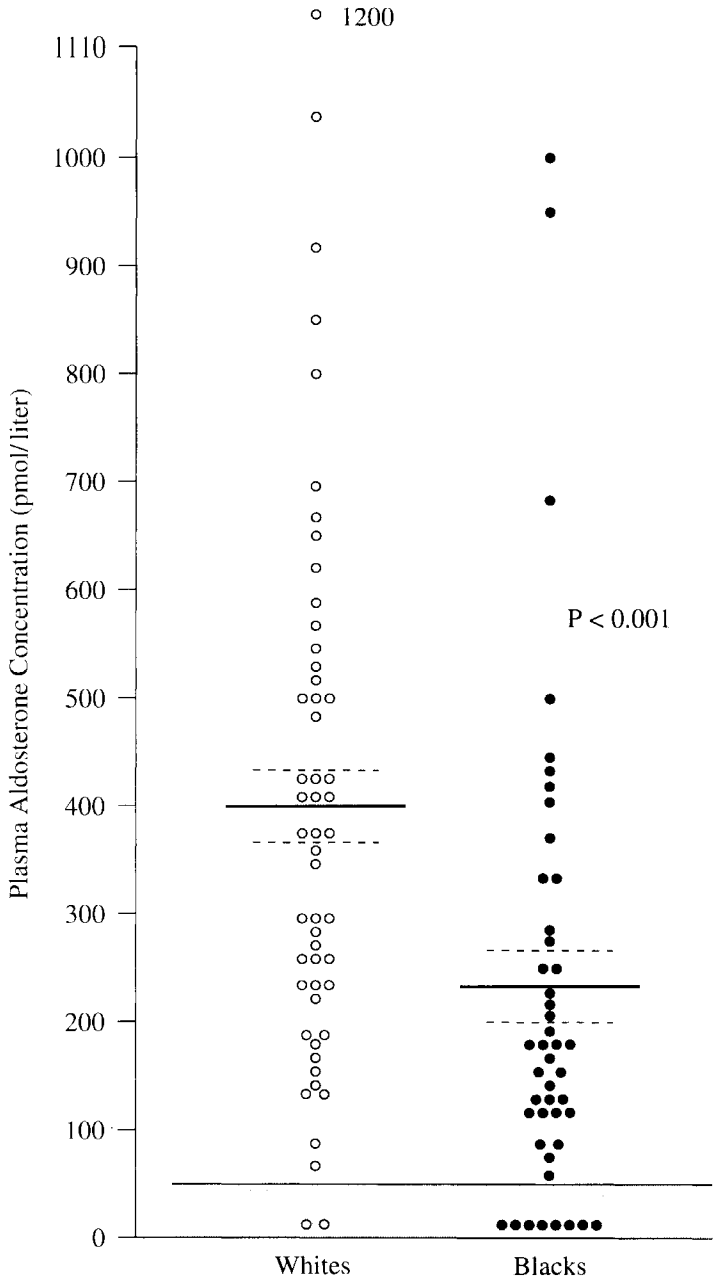


FIGURE 5.21
Plasma aldosterone concentrations in 53 white and 46 black children. Values within the shaded area were undetectable (< 50 pmol per liter). The solid horizontal lines indicate the mean values, and the broken horizontal lines the mean ± se. The concept of standard error (se) is discussed in Chapter 6.

Hint: Compute the observed proportion of women who fall within 1, 1.5, 2.0, and 2.5 standard deviations of the mean. Compare the observed proportions with the expected proportions based on the assumption of normality.

5.68 Answer Problem 5.67 using the ln (nutrient) transformation for each nutrient value. Is the normality assumption more appropriate for log-transformed or untransformed values, or neither?

5.69 A special problem arises for the nutrient alcohol consumption. There is often a large number of nondrinkers (alcohol consumption = 0) and another large group of drinkers with alcohol consumption > 0. The overall distribution of alcohol consumption appears bimodal. Plot the distribution of alcohol consumption for both the DR and the FFQ. Do the distributions appear unimodal or bimodal? Do you think that the normality assumption is appropriate for this nutrient?

TABLE 5.4 Expectation of life and mortality rates, by age, race, and sex: 1973

Age (years)	Expectation of life in years					Mortality rate per 1,000 living at specified age				
	Total	White		Negro and other		Total	White		Negro and other	
		Male	Female	Male	Female		Male	Female	Male	Female
40	34.9	32.2	38.5	28.7	34.4	2.95	3.20	1.82	8.12	4.45
41	34.0	31.3	37.6	27.9	33.6	3.20	3.50	1.99	8.53	4.79
42	33.1	30.4	36.7	27.1	32.7	3.50	3.87	2.18	9.07	5.19
43	32.2	29.5	35.7	26.4	31.9	3.85	4.31	2.42	9.79	5.65
44	31.3	28.6	34.8	25.6	31.1	4.25	4.81	2.68	10.66	6.17
45	30.5	27.8	33.9	24.9	30.3	4.70	5.39	2.97	11.63	6.74
46	29.6	26.9	33.0	24.2	29.5	5.17	6.00	3.27	12.62	7.32
47	28.8	26.1	32.1	23.5	28.7	5.64	6.61	3.57	13.56	7.88
48	27.9	25.3	31.2	22.8	27.9	6.08	7.19	3.85	14.42	8.39
49	27.1	24.4	30.3	22.1	27.1	6.53	7.77	4.12	15.23	8.87
50	26.3	23.6	29.5	21.5	26.4	6.99	8.38	4.41	16.03	9.35
51	25.5	22.8	28.6	20.8	25.6	7.51	9.09	4.74	16.94	9.90
52	24.6	22.0	27.7	20.2	24.9	8.16	9.97	5.14	18.05	10.57
53	23.8	21.2	26.9	19.5	24.1	8.97	11.07	5.64	19.47	11.42
54	23.1	20.5	26.0	18.9	23.4	9.90	12.36	6.22	21.13	12.40
55	22.3	19.7	25.2	18.3	22.7	10.92	13.76	6.86	22.93	13.46
56	21.5	19.0	24.4	17.7	22.0	11.97	15.22	7.52	24.75	14.53
57	20.8	18.3	23.5	17.2	21.3	13.06	16.76	8.19	26.51	15.58
58	20.0	17.6	22.7	16.6	20.6	14.18	18.36	8.84	28.12	16.58
59	19.3	16.9	21.9	16.1	20.0	15.32	20.04	9.49	29.64	17.54
60	18.6	16.2	21.1	15.6	19.3	16.56	21.82	10.21	31.25	18.64
61	17.9	15.6	20.3	15.0	18.7	17.89	23.73	11.02	32.98	19.85
62	17.2	15.0	19.6	14.5	18.0	19.27	25.75	11.88	34.66	20.93
63	16.6	14.3	18.8	14.0	17.4	20.68	27.88	12.81	36.24	21.82
64	15.9	13.7	18.0	13.5	16.8	22.17	30.15	13.82	37.81	22.63
65	15.3	13.2	17.3	13.1	16.2	23.73	32.56	14.94	39.20	23.18
70	12.2	10.4	13.7	10.7	13.2	35.38	47.85	23.63	57.22	40.74
75	9.5	8.1	10.4	9.2	11.3	55.13	72.33	41.66	78.02	55.91
80	7.3	6.3	7.9	7.9	9.4	82.71	107.02	68.61	93.95	65.28
85 and over	5.4	4.7	5.7	6.3	7.3	1,000.00	1,000.00	1,000.00	1,000.00	1,000.00

Source: U.S. National Center for Health Statistics, *Vital Statistics of the United States*, annual.

Occupational Health

Table 5.4 is obtained from the 1975 Statistical Abstract of the United States published by the Census Bureau with the primary data obtained from the National Center for Health Statistics [4]. The right-hand side of the table provides age-race-sex-specific 1-year mortality rates for the United States in 1973. Please note that the entries on the right side of the table are the number of deaths per 1000 individuals; they are not percentages. Suppose we are investigating workers in a nuclear-power plant and wish to ascertain whether the mortality of workers in this plant is higher or lower than expected. On January 1, 1973, we have the following age distribution in the plant as given in Table 5.5:

5.70 Suppose we follow this group of men over a 5-year period from January 1, 1973, to December 31, 1977, and find that 20 of the men have died over this period. Is this an unusual number of deaths? Justify your answer. Please assume that the mortality rate of a 45-year-old, for example, remains constant over the 5-year period. *Hint:* Consider using an approximation to solve this problem.

TABLE 5.5 Age distribution in plant

Age	*n*
45[a]	30
50	80
55	70
60	20
Total	200

[a]We assume for simplicity that 30 of the workers are exactly 45 years old, 80 are exactly 50 years old, 70 are exactly 55 years old, and 20 are exactly 60 years old. We also assume that 80% of the workers are white and 20% are black within each age group.

References

[1] Pratt, J. H., Jones, J. J., Miller, J. Z, Wagner, M. A., & Fineberg, N. S. (1989, October). Racial differences in aldosterone excretion and plasma aldosterone concentrations in children. *New England Journal of Medicine*, *321*(17), 1152–1157.

[2] Report of the Second Task Force on Blood Pressure Control in Children—1987. (1987, January). National Heart, Lung and Blood Institute, Bethesda, Maryland. *Pediatrics*, *79*(1), 1–25.

[3] Schwartz, J., & Dockery, D. W. (1992, January). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, *135*(1), 12–19.

[4] U.S. National Center for Health Statistics (1975). *Vital Statistics of the United States*. Washington, D.C.: Government Printing Office.