*CHAPTER*

*4*

*PROBABILITY DISTRIBUTIONS*

*CHAPTER OVERVIEW*

*Probability distributions of random variables assume powerful roles in statis-*

*tical analyses. Since they show all possible values of a random variable and*

*the probabilities associated with these values, probability distributions may*

*be summarized in ways that enable researchers to easily make objective de-*

*cisions based on samples drawn from the populations that the distributions*

*represent. This chapter introduces frequently used discrete and continuous*

*probability distributions that are used in later chapters to make statistical*

*inferences.*

*TO P I C S*

*4.1 INTRODUCTION*

*4.2 PROBABILITY DISTRIBUTIONS OF DISCRETE VARIABLES*

*4.3 THE BINOMIAL DISTRIBUTION*

*4.4 THE POISSON DISTRIBUTION*

*4.5 CONTINUOUS PROBABILITY DISTRIBUTIONS*

*4.6 THE NORMAL DISTRIBUTION*

*4.7 NORMAL DISTRIBUTION APPLICATIONS*

*4.8 SUMMARY*

*LEARNING OUTCOMES*

*After studying this chapter, the student will*

*1. understand selected discrete distributions and how to use them to calculate*

*probabilities in real-world problems.*

*2. understand selected continuous distributions and how to use them to calculate*

*probabilities in real-world problems.*

*3. be able to explain the similarities and differences between distributions of the*

*discrete type and the continuous type and when the use of each is appropriate.*

*4.1 INTRODUCTION*

*In the preceding chapter we introduced the basic concepts of probability as well as meth-*

*ods for calculating the probability of an event. We build on these concepts in the present*

*chapter and explore ways of calculating the probability of an event under somewhat more*

*complex conditions. In this chapter we shall see that the relationship between the values*

*of a random variable and the probabilities of their occurrence may be summarized by means*

*of a device called a probability distribution. A probability distribution may be expressed*

*in the form of a table, graph, or formula. Knowledge of the probability distribution of a*

*random variable provides the clinician and researcher with a powerful tool for summariz-*

*ing and describing a set of data and for reaching conclusions about a population of data*

*on the basis of a sample of data drawn from the population.*

*4.2 PROBABILITY DISTRIBUTIONS*

*OF DISCRETE VARIABLES*

*Let us begin our discussion of probability distributions by considering the probability*

*distribution of a discrete variable, which we shall define as follows:*

*DEFINITION*

*The probability distribution of a discrete random variable is a table,*

*graph, formula, or other device used to specify all possible values of a*

*discrete random variable along with their respective probabilities.*

*If we let the discrete probability distribution be represented by p1x2, then p1x2 = P1X = x2*

*is the probability of the discrete random variable X to assume a value x.*

*EXAMPLE 4.2.1*

*In an article appearing in the Journal of the American Dietetic Association, Holben et al.*

*(A-1) looked at food security status in families in the Appalachian region of southern Ohio.*

*The purpose of the study was to examine hunger rates of families with children in a local*

*Head Start program in Athens, Ohio. The survey instrument included the 18-question U.S.*

*Household Food Security Survey Module for measuring hunger and food security. In addi-*

*tion, participants were asked how many food assistance programs they had used in the last*

*12 months. Table 4.2.1 shows the number of food assistance programs used by subjects in*

*this sample.*

*We wish to construct the probability distribution of the discrete variable X, where*

*X = number of food assistance programs used by the study subjects.*

*Solution:*

*The values of X are x1 = 1, x 2 = 2, . . . , x 7 = 7, and x8 = 8. We compute*

*the probabilities for these values by dividing their respective frequencies by*

*the total, 297. Thus, for example, p1x 12 = P1X = x 12 = 62>297 = .2088.*

*TABLE 4.2.1 Number of Assistance*

*Programs Utilized by Families with*

*Children in Head Start Programs in*

*Southern Ohio*

*Number of ProgramsFrequency*

*1*

*262*

*47*

*3*

*4*

*5*

*6*

*7*

*839*

*39*

*58*

*37*

*4*

*11*

*Total*

*297*

*Source: David H. Holben, Ph.D. and John P. Holcomb,*

*Ph.D. Used with permission.*

*TABLE 4.2.2 Probability Distribution of*

*Programs Utilized by Families Among*

*the Subjects Described in Example 4.2.1*

*Number of Programs (x )P (X " x )*

*1*

*2*

*3*

*4*

*5*

*6*

*7*

*8.2088*

*.1582*

*.1313*

*.1313*

*.1953*

*.1246*

*.0135*

*.0370*

*Total*

*1.0000*

*We display the results in Table 4.2.2, which is the desired probability*

*distribution.*

*■*

*Alternatively, we can present this probability distribution in the form of a graph, as*

*in Figure 4.2.1. In Figure 4.2.1 the length of each vertical bar indicates the probability*

*for the corresponding value of x.*

*It will be observed in Table 4.2.2 that the values of p1x2 = P1X = x2 are all*

*positive, they are all less than 1, and their sum is equal to 1. These are not phenomena*

*peculiar to this particular example, but are characteristics of all probability distributions*

*of discrete variables. If x1, x 2, x3, . . . , xk are all possible values of the discrete random*

*variable X, then we may then give the following two essential properties of a probability*

*distribution of a discrete variable:*

*(1) 0 … P1X = x2 … 1*

*(2) a P1X = x2 = 1, for all x*

*The reader will also note that each of the probabilities in Table 4.2.2 is the*

*relative frequency of occurrence of the corresponding value of X.*

*With its probability distribution available to us, we can make probability statements*

*regarding the random variable X. We illustrate with some examples.*

*EXAMPLE 4.2.2*

*What is the probability that a randomly selected family used three assistance*

*programs?*

*Solution:*

*We may write the desired probability as p132 = P1X = 32. We see in*

*Table 4.2.2 that the answer is .1313.*

*■*

*EXAMPLE 4.2.3*

*What is the probability that a randomly selected family used either one or two programs?*

*Solution:*

*To answer this question, we use the addition rule for mutually exclusive*

*events. Using probability notation and the results in Table 4.2.2, we write the*

*answer as P11 ´ 22 = P112 + P122 = .2088 + .1582 = .3670.*

*Cumulative Distributions*

*Sometimes it will be more convenient to work*

*with the cumulative probability distribution of a random variable. The cumulative prob-*

*ability distribution for the discrete variable whose probability distribution is given in*

*Table 4.2.2 may be obtained by successively adding the probabilities, P1X = x i2, given*

*in the last column. The cumulative probability for xi is written as F1x i2 = P1X … x i2.*

*It gives the probability that X is less than or equal to a specified value, xi.*

*The resulting cumulative probability distribution is shown in Table 4.2.3. The graph*

*of the cumulative probability distribution is shown in Figure 4.2.2. The graph of a cumu-*

*lative probability distribution is called an ogive. In Figure 4.2.2 the graph of F1x2 con-*

*sists solely of the horizontal lines. The vertical lines only give the graph a connected*

*appearance. The length of each vertical line represents the same probability as that of the*

*corresponding line in Figure 4.2.1. For example, the length of the vertical line at X = 3*

*1.0*

*0.9*

*0.8*

*0.7*

*f (x)*

*0.6*

*0.5*

*0.4*

*0.3*

*0.2*

*0.1*

*0.0*

*1*

*2*

*3*

*4*

*5*

*6*

*7*

*8*

*x (number of programs)*

*FIGURE 4.2.2 Cumulative probability distribu-*

*tion of number of assistance programs among the*

*subjects described in Example 4.2.1.*

*in Figure 4.2.2 represents the same probability as the length of the line erected at X = 3*

*in Figure 4.2.1, or .1313 on the vertical scale.*

*By consulting the cumulative probability distribution we may answer quickly ques-*

*tions like those in the following examples.*

*EXAMPLE 4.2.4*

*What is the probability that a family picked at random used two or fewer assistance*

*programs?*

*Solution:*

*The probability we seek may be found directly in Table 4.2.3 by reading*

*the cumulative probability opposite x = 2, and we see that it is .3670. That*

*is, P1X … 22 = .3670. We also may find the answer by inspecting Figure*

*4.2.2 and determining the height of the graph (as measured on the vertical*

*axis) above the value X = 2.*

*■*

*EXAMPLE 4.2.5*

*What is the probability that a randomly selected family used fewer than four programs?*

*Solution:*

*Since a family that used fewer than four programs used either one, two, or*

*three programs, the answer is the cumulative probability for 3. That is,*

*■*

*P1X 6 42 = P1X … 32 = .4983.*

*EXAMPLE 4.2.6*

*What is the probability that a randomly selected family used five or more programs?*

*Solution:*

*To find the answer we make use of the concept of complementary probabilities.*

*The set of families that used five or more programs is the complement of the*

*set of families that used fewer than five (that is, four or fewer) programs. The*

*sum of the two probabilities associated with these sets is equal to 1. We write*

*this relationship in probability notation as P1X Ú 52 + P1X … 42 = 1.*

*Therefore, P1X Ú 52 = 1 - P1X … 42 = 1 - .6296 = .3704.*

*■*

*EXAMPLE 4.2.7*

*What is the probability that a randomly selected family used between three and five*

*programs, inclusive?*

*Solution:*

*P1X … 52 = .8249 is the probability that a family used between one and*

*five programs, inclusive. To get the probability of between three and*

*five programs, we subtract, from .8249, the probability of two or fewer.*

*Using probability notation we write the answer as P13 … X … 52 =*

*P1X … 52 - P1X … 22 = .8249 - .3670 = .4579.*

*■*

*The probability distribution given in Table 4.2.1 was developed out of actual experience, so*

*to find another variable following this distribution would be coincidental. The probability*

*distributions of many variables of interest, however, can be determined or assumed on*

*the basis of theoretical considerations. In later sections, we study in detail three of these*

*theoretical probability distributions: the binomial, the Poisson, and the normal.*

*Mean and Variance of Discrete Probability Distributions The*

*mean and variance of a discrete probability distribution can easily be found using the*

*formulae below.*

*m = a xp1x2*

*(4.2.1)*

*s = a 1x - m2 p1x2 = a x p1x2 - m*

*2*

*2*

*2*

*2*

*(4.2.2)*

*where p1x2 is the relative frequency of a given random variable X. The standard deviation*

*is simply the positive square root of the variance.*

*EXAMPLE 4.2.8*

*What are the mean, variance, and standard deviation of the distribution from Example 4.2.1?*

*Solution:*

*m = 1121.20882 + 1221.15822 + 1321.13132 + Á + 1821.03702 = 3.5589*

*s2 = 11 - 3.5589221.20882 + 12 - 3.5589221.15822 + 13 - 3.5589221.13132*

*+ Á + 18 - 3.5589221.03702 = 3.8559*

*We therefore can conclude that the mean number of programs utilized was 3.5589 with a*

*variance of 3.8559. The standard deviation is therefore 23.5589 = 1.9637 programs. ■*

*EXERCISES*

*4.2.1In a study by Cross et al. (A-2), patients who were involved in problem gambling treatment were*

*asked about co-occurring drug and alcohol addictions. Let the discrete random variable X represent*

*the number of co-occurring addictive substances used by the subjects. Table 4.2.4 summarizes the*

*frequency distribution for this random variable.*

*(a) Construct a table of the relative frequency and the cumulative frequency for this discrete*

*distribution.*

*(b) Construct a graph of the probability distribution and a graph representing the cumulative*

*probability distribution for these data.*

*4.2.2Refer to Exercise 4.2.1.*

*(a) What is probability that an individual selected at random used five addictive substances?*

*(b) What is the probability that an individual selected at random used fewer than three addictive*

*substances?*

*(c) What is the probability that an individual selected at random used more than six addictive*

*substances?*

*(d) What is the probability that an individual selected at random used between two and five*

*addictive substances, inclusive?*

*4.2.3Refer to Exercise 4.2.1. Find the mean, variance, and standard deviation of this frequency distribution.*

*Table 4.2.4 Number of Co-occurring Addictive Substances*

*Used by Patients in Selected Gambling Treatment Programs*

*Number of Substances Used*

*Frequency*

*0*

*1*

*2*

*3144*

*342*

*142*

*72*

*4*

*5*

*6*

*7*

*839*

*20*

*6*

*9*

*2*

*91*

*Total777*

*4.3 THE BINOMIAL DISTRIBUTION*

*The binomial distribution is one of the most widely encountered probability distributions*

*in applied statistics. The distribution is derived from a process known as a Bernoulli trial,*

*named in honor of the Swiss mathematician James Bernoulli (1654–1705), who made*

*significant contributions in the field of probability, including, in particular, the binomial*

*distribution. When a random process or experiment, called a trial, can result in only one*

*of two mutually exclusive outcomes, such as dead or alive, sick or well, full-term or*

*premature, the trial is called a Bernoulli trial.*

*The Bernoulli Process*

*A sequence of Bernoulli trials forms a Bernoulli*

*process under the following conditions.*

*1. Each trial results in one of two possible, mutually exclusive, outcomes. One of the pos-*

*sible outcomes is denoted (arbitrarily) as a success, and the other is denoted a failure.*

*2. The probability of a success, denoted by p, remains constant from trial to trial. The*

*probability of a failure, 1 - p, is denoted by q.*

*3. The trials are independent; that is, the outcome of any particular trial is not affected*

*by the outcome of any other trial.*

*EXAMPLE 4.3.1*

*We are interested in being able to compute the probability of x successes in n Bernoulli*

*trials. For example, if we examine all birth records from the North Carolina State Center*

*for Health Statistics (A-3) for the calendar year 2001, we find that 85.8 percent of the*

*pregnancies had delivery in week 37 or later. We will refer to this as a full-term birth.*

*With that percentage, we can interpret the probability of a recorded birth in week 37 or*

*later as .858. If we randomly select five birth records from this population, what is the*

*probability that exactly three of the records will be for full-term births?*

*Solution:*

*101*

*Let us designate the occurrence of a record for a full-term birth (F) as a*

*“success,” and hasten to add that a premature birth (P) is not a failure, but*

*medical research indicates that children born in week 36 or sooner are at*

*risk for medical complications. If we are looking for birth records of pre-*

*mature deliveries, these would be designated successes, and birth records*

*of full-term would be designated failures.*

*It will also be convenient to assign the number 1 to a success (record for*

*a full-term birth) and the number 0 to a failure (record of a premature birth).*

*The process that eventually results in a birth record we consider to be*

*a Bernoulli process.*

*Suppose the five birth records selected resulted in this sequence of*

*full-term births:*

*FPFFP*

*In coded form we would write this as*

*10110*

*Since the probability of a success is denoted by p and the probabil-*

*ity of a failure is denoted by q, the probability of the above sequence of*

*outcomes is found by means of the multiplication rule to be*

*P11, 0, 1, 1, 02 = pqppq = q 2 p 3*

*The multiplication rule is appropriate for computing this probability since*

*we are seeking the probability of a full-term, and a premature, and a full-*

*term, and a full-term, and a premature, in that order or, in other words, the*

*joint probability of the five events. For simplicity, commas, rather than inter-*

*section notation, have been used to separate the outcomes of the events in*

*the probability statement.*

*The resulting probability is that of obtaining the specific sequence of out-*

*comes in the order shown. We are not, however, interested in the order of occur-*

*rence of records for full-term and premature births but, instead, as has been*

*stated already, the probability of the occurrence of exactly three records of full-*

*term births out of five randomly selected records. Instead of occurring in the*

*sequence shown above (call it sequence number 1), three successes and two*

*failures could occur in any one of the following additional sequences as well:*

*NumberSequence*

*2*

*3*

*4*

*5*

*6*

*7*

*8*

*9*

*1011100*

*10011*

*11010*

*11001*

*10101*

*01110*

*00111*

*01011*

*01101*

*Each of these sequences has the same probability of occurring, and*

*this probability is equal to q 2 p 3, the probability computed for the first*

*sequence mentioned.*

*When we draw a single sample of size five from the population spec-*

*ified, we obtain only one sequence of successes and failures. The question*

*now becomes, What is the probability of getting sequence number 1 or*

*sequence number 2 . . . or sequence number 10? From the addition rule we*

*know that this probability is equal to the sum of the individual probabili-*

*ties. In the present example we need to sum the 10q 2 p 3’s or, equivalently,*

*multiply q 2 p 3 by 10. We may now answer our original question: What is*

*the probability, in a random sample of size 5, drawn from the specified*

*population, of observing three successes (record of a full-term birth) and*

*two failures (record of a premature birth)? Since in the population,*

*p = .858, q = 11 - p2 = 11 - .8582 = .142 the answer to the question is*

*101.142221.85823 = 101.020221.63162 = .1276*

*■*

*Large Sample Procedure: Use of Combinations*

*We can easily*

*anticipate that, as the size of the sample increases, listing the number of sequences becomes*

*more and more difficult and tedious. What is needed is an easy method of counting the*

*number of sequences. Such a method is provided by means of a counting formula that*

*allows us to determine quickly how many subsets of objects can be formed when we use*

*in the subsets different numbers of the objects that make up the set from which the objects*

*are selected. When the order of the objects in a subset is immaterial, the subset is called*

*a combination of objects. When the order of objects in a subset does matter, we refer to*

*the subset as a permutation of objects. Though permutations of objects are often used in*

*probability theory, they will not be used in our current discussion. If a set consists of n*

*objects, and we wish to form a subset of x objects from these n objects, without regard to*

*the order of the objects in the subset, the result is called a combination. For examples, we*

*define a combination as follows when the combination is formed by taking x objects from*

*a set of n objects.*

*DEFINITION*

*A combination of n objects taken x at a time is an unordered subset of*

*x of the n objects.*

*The number of combinations of n objects that can be formed by taking x of them*

*at a time is given by*

*n!*

*(4.3.1)*

*nCx =*

*x!1n - x2!*

*where x!, read x factorial, is the product of all the whole numbers from x down to 1.*

*That is, x! = x1x - 121x - 22 . . . 112. We note that, by definition, 0! = 1.*

*Let us return to our example in which we have a sample of n = 5 birth records and*

*we are interested in finding the probability that three of them will be for full-term births.*

*TABLE 4.3.1 The Binomial Distribution*

*Number of Successes, xProbability, f (x )*

*0*

*1*

*2n-0 0*

*p*

*nC 0q*

*n -1 1*

*p*

*nC 1q*

*n -2 2*

*p*

*nC 2q*

*o*

*o*

*x*

*nCx q*

*o*

*n-x x*

*p*

*o*

*n*

*nC nq*

*Total*

*n-n n*

*p*

*1*

*The number of sequences in our example is found by Equation 4.3.1 to be*

*5C 3 =*

*5#4#3#2#1*

*120*

*=*

*= 10*

*3#2#1#2#1*

*12*

*In our example we let x = 3, the number of successes, so that n - x = 2, the*

*number of failures. We then may write the probability of obtaining exactly x successes*

*in n trials as*

*f 1x2 = nCx q n-xp x = nCx p xq n-x for x = 0, 1, 2, . . . , n*

*= 0,*

*elsewhere*

*(4.3.2)*

*This expression is called the binomial distribution. In Equation 4.3.2 f 1x2 "*

*P1X = x2, where X is the random variable, the number of successes in n trials. We use f 1x2*

*rather than P1X = x2 because of its compactness and because of its almost universal use.*

*We may present the binomial distribution in tabular form as in Table 4.3.1.*

*We establish the fact that Equation 4.3.2 is a probability distribution by showing*

*the following:*

*1. f 1x2 = 0 for all real values of x. This follows from the fact that n and p are both*

*nonnegative and, hence, nC x , p x, and 11 - p2n-x are all nonnegative and, therefore,*

*their product is greater than or equal to zero.*

*2. ©f 1x2 = 1. This is seen to be true if we recognize that © nCx q n-xp x is equal to*

*311 - p2 + p4n = 1n = 1, the familiar binomial expansion. If the binomial*

*1q + p2n is expanded, we have*

*1q + p2n = q n + nq n-1p 1 +*

*n1n - 12*

*2*

*q n-2p 2 + . . . + nq 1p n-1 + p n*

*If we compare the terms in the expansion, term for term, with the f 1x2 in*

*Table 4.3.1 we see that they are, term for term, equivalent, since*

*f 102 = nC 0q n-0p 0 = q n*

*f 112 = nC1q n-1p 1 = nq n-1p*

*o*

*o*

*o*

*f 1n2 = nCnq n-np n = p n*

*f 122 = nC 2q n-2p 2 =*

*EXAMPLE 4.3.2*

*As another example of the use of the binomial distribution, the data from the North*

*Carolina State Center for Health Statistics (A-3) show that 14 percent of mothers admit-*

*ted to smoking one or more cigarettes per day during pregnancy. If a random sample*

*of size 10 is selected from this population, what is the probability that it will contain*

*exactly four mothers who admitted to smoking during pregnancy?*

*Solution:*

*We take the probability of a mother admitting to smoking to be .14. Using*

*Equation 4.3.2 we find*

*f 142 = 10C 4 1.86261.1424*

*10!*

*1.404567221.00038422*

*4!6!*

*= .0326*

*=*

*■*

*Binomial Table*

*The calculation of a probability using Equation 4.3.2 can be a*

*tedious undertaking if the sample size is large. Fortunately, probabilities for different val-*

*ues of n, p, and x have been tabulated, so that we need only to consult an appropriate*

*table to obtain the desired probability. Table B of the Appendix is one of many such*

*tables available. It gives the probability that X is less than or equal to some specified*

*value. That is, the table gives the cumulative probabilities from x = 0 up through some*

*specified positive number of successes.*

*Let us illustrate the use of the table by using Example 4.3.2, where it was desired*

*to find the probability that x = 4 when n = 10 and p = .14. Drawing on our knowledge*

*of cumulative probability distributions from the previous section, we know that P1x = 42*

*may be found by subtracting P1X … 32 from P1X … 42. If in Table B we locate p = .14*

*for n = 10, we find that P1X … 42 = .9927 and P1X … 32 = .9600. Subtracting the*

*latter from the former gives .9927 - .9600 = .0327, which nearly agrees with our hand*

*calculation (discrepancy due to rounding).*

*Frequently we are interested in determining probabilities, not for specific values*

*of X, but for intervals such as the probability that X is between, say, 5 and 10. Let us*

*illustrate with an example.*

*EXAMPLE 4.3.3*

*Suppose it is known that 10 percent of a certain population is color blind. If a random*

*sample of 25 people is drawn from this population, use Table B in the Appendix to find*

*the probability that:*

*(a) Five or fewer will be color blind.*

*Solution:*

*105*

*This probability is an entry in the table. No addition or subtraction is nec-*

*essary. P1X … 52 = .9666.*

*(b) Six or more will be color blind.*

*Solution:*

*We cannot find this probability directly in the table. To find the answer, we*

*use the concept of complementary probabilities. The probability that six or*

*more are color blind is the complement of the probability that five or fewer*

*are color blind. That is, this set is the complement of the set specified in*

*part a; therefore,*

*P1X Ú 62 = 1 - P1X … 52 = 1 - .9666 = .0334*

*(c) Between six and nine inclusive will be color blind.*

*Solution:*

*We find this by subtracting the probability that X is less than or equal to 5*

*from the probability that X is less than or equal to 9. That is,*

*P16 … X … 92 = P1X … 92 - P 1X … 52 = .9999 - .9666 = .0333*

*(d) Two, three, or four will be color blind.*

*Solution:*

*This is the probability that X is between 2 and 4 inclusive.*

*P12 … X … 42 = P1X … 42 - P1X … 12 = .9020 - .2712 = .6308*

*■*

*Using Table B When p>.5 Table B does not give probabilities for values of*

*p greater than .5. We may obtain probabilities from Table B, however, by restating the*

*problem in terms of the probability of a failure, 1 - p, rather than in terms of the prob-*

*ability of a success, p. As part of the restatement, we must also think in terms of the num-*

*ber of failures, n - x, rather than the number of successes, x. We may summarize this*

*idea as follows:*

*P1X = x ƒ n, p 7 .502 = P1X = n - x ƒ n, 1 - p2*

*(4.3.3)*

*In words, Equation 4.3.3 says, “The probability that X is equal to some specified value given*

*the sample size and a probability of success greater than .5 is equal to the probability that*

*X is equal to n - x given the sample size and the probability of a failure of 1 - p.” For*

*purposes of using the binomial table we treat the probability of a failure as though it were*

*the probability of a success. When p is greater than .5, we may obtain cumulative proba-*

*bilities from Table B by using the following relationship:*

*P1X … x ƒ n, p 7 .502 = P1X Ú n - x ƒ n, 1 - p2*

*(4.3.4)*

*Finally, to use Table B to find the probability that X is greater than or equal to some x*

*when P 7 .5, we use the following relationship:*

*P1X Ú x ƒ n, p 7 .502 = P1X … n - x ƒ n, 1 - p2*

*(4.3.5)*

*EXAMPLE 4.3.4*

*According to a June 2003 poll conducted by the Massachusetts Health Benchmarks*

*project (A-4), approximately 55 percent of residents answered “serious problem” to the*

*question, “Some people think that childhood obesity is a national health problem. What*

*do you think? Is it a very serious problem, somewhat of a problem, not much of a prob-*

*lem, or not a problem at all?” Assuming that the probability of giving this answer to the*

*question is .55 for any Massachusetts resident, use Table B to find the probability that if*

*12 residents are chosen at random:*

*(a) Exactly seven will answer “serious problem.”*

*Solution:*

*We restate the problem as follows: What is the probability that a randomly*

*selected resident gives an answer other than “serious problem” from exactly*

*five residents out of 12, if 45 percent of residents give an answer other than*

*“serious problem.” We find the answer as follows:*

*P1X = 5 ƒ n = 12, p = .452 = P1X … 52 - P1X … 42*

*= .5269 - .3044 = .2225*

*(b) Five or fewer households will answer “serious problem.”*

*Solution:*

*The probability we want is*

*P1X … 5 ƒ n = 12, p = .552 = P1X Ú 12 - 5 ƒ n = 12, p = .452*

*= P1X Ú 7 ƒ n = 12, p = .452*

*= 1 - P1X … 6 ƒ n = 12, p = .452*

*= 1 - .7393 = .2607*

*(c) Eight or more households will answer “serious problem.”*

*Solution:*

*The probability we want is*

*P1X Ú 8 ƒ n = 12, p = .552 = P1X … 4 ƒ n = 12, p = .452 = .3044*

*■*

*Figure 4.3.1 provides a visual representation of the solution to the three parts of*

*Example 4.3.4.*

*The Binomial Parameters The binomial distribution has two parameters,*

*n and p. They are parameters in the sense that they are sufficient to specify a bino-*

*mial distribution. The binomial distribution is really a family of distributions with*

*each possible value of n and p designating a different member of the family. The*

*mean and variance of the binomial distribution are m = np and s2 = np 11 - p2,*

*respectively.*

*Strictly speaking, the binomial distribution is applicable in situations where sam-*

*pling is from an infinite population or from a finite population with replacement. Since*

*in actual practice samples are usually drawn without replacement from finite populations,*

*the question arises as to the appropriateness of the binomial distribution under these*

*circumstances. Whether or not the binomial is appropriate depends on how drastic the*

*effect of these conditions is on the constancy of p from trial to trial. It is generally agreed*

*FIGURE 4.3.1 Schematic representation of solutions to Example 4.3.4 (the relevant*

*numbers of successes and failures in each case are circled).*

*that when n is small relative to N, the binomial model is appropriate. Some writers say*

*that n is small relative to N if N is at least 10 times as large as n.*

*Most statistical software programs allow for the calculation of binomial probabilities*

*with a personal computer. EXCEL, for example, can be used to calculate individual or cumu-*

*lative probabilities for specified values of x, n, and p. Suppose we wish to find the individ-*

*ual probabilities for x = 0 through x = 6 when n = 6 and p = .3. We enter the numbers*

*0 through 6 in Column 1 and proceed as shown in Figure 4.3.2. We may follow a similar*

*procedure to find the cumulative probabilities. For this illustration, we use MINITAB and*

*place the numbers 1 through 6 in Column 1. We proceed as shown in Figure 4.3.3.*

*EXERCISES*

*In each of the following exercises, assume that N is sufficiently large relative to n that the bino-*

*mial distribution may be used to find the desired probabilities.*

*4.3.1*

*Based on data collected by the National Center for Health Statistics and made available to the*

*public in the Sample Adult database (A-5), an estimate of the percentage of adults who have at*

*some point in their life been told they have hypertension is 23.53 percent. If we select a simple*

*random sample of 20 U.S. adults and assume that the probability that each has been told that he*

*or she has hypertension is .24, find the probability that the number of people in the sample who*

*have been told that they have hypertension will be:*

*(a) Exactly three*

*(b) Three or more*

*(c) Fewer than three*

*(d) Between three and seven, inclusive*

*4.3.2Refer to Exercise 4.3.1. How many adults who have been told that they have hypertension would*

*you expect to find in a sample of 20?*

*4.3.3Refer to Exercise 4.3.1. Suppose we select a simple random sample of five adults. Use Equation*

*4.3.2 to find the probability that, in the sample, the number of people who have been told that*

*they have hypertension will be:*

*(a) Zero*

*(b) More than one*

*(c) Between one and three, inclusive*

*(d) Two or fewer*

*(e) Five*

*4.3.4The same survey database cited in exercise 4.3.1 (A-5) shows that 32 percent of U.S. adults indi-*

*cated that they have been tested for HIV at some point in their life. Consider a simple random*

*sample of 15 adults selected at that time. Find the probability that the number of adults who have*

*been tested for HIV in the sample would be:*

*(a) Three*

*(b) Less than five*

*(c) Between five and nine, inclusive*

*(d) More than five, but less than 10*

*(e) Six or more*

*4.3.5Refer to Exercise 4.3.4. Find the mean and variance of the number of people tested for HIV in*

*samples of size 15.*

*4.3.6Refer to Exercise 4.3.4. Suppose we were to take a simple random sample of 25 adults today and*

*find that two have been tested for HIV at some point in their life. Would these results be surpris-*

*ing? Why or why not?*

*4.3.7Coughlin et al. (A-6) estimated the percentage of women living in border counties along the south-*

*ern United States with Mexico (designated counties in California, Arizona, New Mexico, and*

*Texas) who have less than a high school education to be 18.7. Assume the corresponding proba-*

*bility is .19. Suppose we select three women at random. Find the probability that the number with*

*less than a high-school education is:*

*(a) Exactly zero*

*(b) Exactly one*

*(c) More than one*

*(d) Two or fewer*

*(e) Two or three*

*(f) Exactly three*

*4.3.8 In a survey of nursing students pursuing a master’s degree, 75 percent stated that they expect*

*to be promoted to a higher position within one month after receiving the degree. If this per-*

*centage holds for the entire population, find, for a sample of 15, the probability that the num-*

*ber expecting a promotion within a month after receiving their degree is:*

*(a) Six*

*(b) At least seven*

*(c) No more than five*

*(d) Between six and nine, inclusive*

*4.3.9*

*Given the binomial parameters p = .8 and n = 3, show by means of the binomial expansion given*

*in Table 4.3.1 that g f 1x2 = 1.*

*4.4 THE POISSON DISTRIBUTION*

*The next discrete distribution that we consider is the Poisson distribution, named for*

*the French mathematician Simeon Denis Poisson (1781–1840), who is generally cred-*

*ited for publishing its derivation in 1837. This distribution has been used extensively as*

*a probability model in biology and medicine. Haight (1) presents a fairly extensive cat-*

*alog of such applications in Chapter 7 of his book.*

*If x is the number of occurrences of some random event in an interval of time or*

*space (or some volume of matter), the probability that x will occur is given by*

*f 1x2 =*

*e -llx*

*,*

*x!*

*x = 0, 1, 2, . . .*

*(4.4.1)*

*The Greek letter l (lambda) is called the parameter of the distribution and is the aver-*

*age number of occurrences of the random event in the interval (or volume). The symbol*

*e is the constant (to four decimals) 2.7183.*

*It can be shown that f 1x2 Ú 0 for every x and that g x f 1x2 = 1 so that the distri-*

*bution satisfies the requirements for a probability distribution.*

*The Poisson Process We have seen that the binomial distribution results from*

*a set of assumptions about an underlying process yielding a set of numerical observa-*

*tions. Such, also, is the case with the Poisson distribution. The following statements*

*describe what is known as the Poisson process.*

*1. The occurrences of the events are independent. The occurrence of an event in an*

*interval1 of space or time has no effect on the probability of a second occurrence*

*of the event in the same, or any other, interval.*

*2. Theoretically, an infinite number of occurrences of the event must be possible in*

*the interval.*

*3. The probability of the single occurrence of the event in a given interval is propor-*

*tional to the length of the interval.*

*4. In any infinitesimally small portion of the interval, the probability of more than*

*one occurrence of the event is negligible.*

*An interesting feature of the Poisson distribution is the fact that the mean and vari-*

*ance are equal.*

*When to Use the Poisson Model The Poisson distribution is employed*

*as a model when counts are made of events or entities that are distributed at random*

*in space or time. One may suspect that a certain process obeys the Poisson law, and*

*under this assumption probabilities of the occurrence of events or entities within some*

*unit of space or time may be calculated. For example, under the assumptions that the*

*distribution of some parasite among individual host members follows the Poisson law,*

*one may, with knowledge of the parameter l, calculate the probability that a randomly*

*selected individual host will yield x number of parasites. In a later chapter we will*

*learn how to decide whether the assumption that a specified process obeys the Pois-*

*son law is plausible. An additional use of the Poisson distribution in practice occurs*

*when n is large and p is small. In this case, the Poisson distribution can be used to*

*approximate the binomial distribution. In other words,*

*x n-x*

*«*

*nCx p q*

*e -llx*

*,*

*x!*

*x = 0, 1, 2, . . .*

*where l = np.*

*To illustrate the use of the Poisson distribution for computing probabilities, let us*

*consider the following examples.*

*EXAMPLE 4.4.1*

*In a study of drug-induced anaphylaxis among patients taking rocuronium bromide as*

*part of their anesthesia, Laake and Røttingen (A-7) found that the occurrence of anaphy-*

*laxis followed a Poisson model with l = 12 incidents per year in Norway. Find the prob-*

*ability that in the next year, among patients receiving rocuronium, exactly three will*

*experience anaphylaxis.*

*Solution:*

*By Equation 4.4.1, we find the answer to be*

*P1X = 32 =*

*e -12123*

*= .00177*

*3!*

*■*

*EXAMPLE 4.4.2*

*Refer to Example 4.4.1. What is the probability that at least three patients in the next*

*year will experience anaphylaxis if rocuronium is administered with anesthesia?*

*Solution:*

*We can use the concept of complementary events in this case. Since*

*P1X … 22 is the complement of P1X Ú 32, we have*

*P1X Ú 32 = 1 - P1X … 22 = 1 - 3P1X = 02 + P1X = 12 + P1X = 224*

*=1 - c*

*e -12120*

*e -12121*

*e -12122*

*+*

*+*

*d*

*0!*

*1!*

*2!*

*= 1 - 3.00000614 + .00007373 + .000442384*

*= 1 - .00052225*

*= .99947775*

*■*

*In the foregoing examples the probabilities were evaluated directly from the equation.*

*We may, however, use Appendix Table C, which gives cumulative probabilities for*

*various values of l and X.*

*EXAMPLE 4.4.3*

*In the study of a certain aquatic organism, a large number of samples were taken from a*

*pond, and the number of organisms in each sample was counted. The average number*

*of organisms per sample was found to be two. Assuming that the number of organisms*

*follows a Poisson distribution, find the probability that the next sample taken will contain*

*one or fewer organisms.*

*Solution:*

*In Table C we see that when l = 2, the probability that X … 1 is .406.*

*That is, P1X … 1 ƒ 22 = .406.*

*■*

*EXAMPLE 4.4.4*

*Refer to Example 4.4.3. Find the probability that the next sample taken will contain*

*exactly three organisms.*

*Solution:*

*P1X = 3 ƒ 22 = P1X … 32 - P1X … 22 = .857 - .677 = .180*

*EXAMPLE 4.4.5*

*Refer to Example 4.4.3. Find the probability that the next sample taken will contain more*

*than five organisms.*

*Solution:*

*Since the set of more than five organisms does not include five, we are ask-*

*ing for the probability that six or more organisms will be observed. This is*

*obtained by subtracting the probability of observing five or fewer from one.*

*That is,*

*P1X 7 5 ƒ 22 = 1 - P1X … 52 = 1 - .983 = .017*

*■*

*Poisson probabilities are obtainable from most statistical software packages. To illustrate*

*the use of MINITAB for this purpose, suppose we wish to find the individual probabil-*

*ities for x = 0 through x = 6 when l = .7. We enter the values of x in Column 1 and*

*proceed as shown in Figure 4.4.1. We obtain the cumulative probabilities for the same*

*values of x and l as shown in Figure 4.4.2.*

*EXERCISES*

*4.4.1*

*Singh et al. (A-8) looked at the occurrence of retinal capillary hemangioma (RCH) in patients with*

*von Hippel–Lindau (VHL) disease. RCH is a benign vascular tumor of the retina. Using a retro-*

*spective consecutive case series review, the researchers found that the number of RCH tumor*

*incidents followed a Poisson distribution with l = 4 tumors per eye for patients with VHL. Using*

*this model, find the probability that in a randomly selected patient with VHL:*

*(a) There are exactly five occurrences of tumors per eye.*

*(b) There are more than five occurrences of tumors per eye.*

*(c) There are fewer than five occurrences of tumors per eye.*

*(d) There are between five and seven occurrences of tumors per eye, inclusive.*

*4.4.2*

*Tubert-Bitter et al. (A-9) found that the number of serious gastrointestinal reactions reported to the*

*British Committee on Safety of Medicine was 538 for 9,160,000 prescriptions of the anti-inflammatory*

*drug piroxicam. This corresponds to a rate of .058 gastrointestinal reactions per 1000 prescriptions*

*written. Using a Poisson model for probability, with l = .06, find the probability of*

*(a) Exactly one gastrointestinal reaction in 1000 prescriptions*

*(b) Exactly two gastrointestinal reactions in 1000 prescriptions*

*(c) No gastrointestinal reactions in 1000 prescriptions*

*(d) At least one gastrointestinal reaction in 1000 prescriptions*

*4.4.3*

*4.4.4*

*4.4.5*

*If the mean number of serious accidents per year in a large factory (where the number of employ-*

*ees remains constant) is five, find the probability that in the current year there will be:*

*(a) Exactly seven accidents(b) Ten or more accidents*

*(c) No accidents(d) Fewer than five accidents*

*In a study of the effectiveness of an insecticide against a certain insect, a large area of land was*

*sprayed. Later the area was examined for live insects by randomly selecting squares and count-*

*ing the number of live insects per square. Past experience has shown the average number of live*

*insects per square after spraying to be .5. If the number of live insects per square follows a Pois-*

*son distribution, find the probability that a selected square will contain:*

*(a) Exactly one live insect(b) No live insects*

*(c) Exactly four live insects(d) One or more live insects*

*In a certain population an average of 13 new cases of esophageal cancer are diagnosed each year.*

*If the annual incidence of esophageal cancer follows a Poisson distribution, find the probability*

*that in a given year the number of newly diagnosed cases of esophageal cancer will be:*

*(a) Exactly 10(b) At least eight*

*(c) No more than 12(d) Between nine and 15, inclusive*

*(e) Fewer than seven*

*4.5 CONTINUOUS PROBABILITY*

*DISTRIBUTIONS*

*The probability distributions considered thus far, the binomial and the Poisson, are dis-*

*tributions of discrete variables. Let us now consider distributions of continuous random*

*variables. In Chapter 1 we stated that a continuous variable is one that can assume any*

*value within a specified interval of values assumed by the variable. Consequently,*

*between any two values assumed by a continuous variable, there exist an infinite num-*

*ber of values.*

*To help us understand the nature of the distribution of a continuous random vari-*

*able, let us consider the data presented in Table 1.4.1 and Figure 2.3.2. In the table we*

*have 189 values of the random variable, age. The histogram of Figure 2.3.2 was con-*

*structed by locating specified points on a line representing the measurement of interest*

*and erecting a series of rectangles, whose widths were the distances between two spec-*

*ified points on the line, and whose heights represented the number of values of the vari-*

*able falling between the two specified points. The intervals defined by any two consec-*

*utive specified points we called class intervals. As was noted in Chapter 2, subareas of*

*the histogram correspond to the frequencies of occurrence of values of the variable*

*between the horizontal scale boundaries of these subareas. This provides a way whereby*

*the relative frequency of occurrence of values between any two specified points can be*

*calculated: merely determine the proportion of the histogram’s total area falling between*

*the specified points. This can be done more conveniently by consulting the relative fre-*

*quency or cumulative relative frequency columns of Table 2.3.2.*

*Imagine now the situation where the number of values of our random variable is*

*very large and the width of our class intervals is made very small. The resulting his-*

*togram could look like that shown in Figure 4.5.1.*

*If we were to connect the midpoints of the cells of the histogram in Figure 4.5.1*

*to form a frequency polygon, clearly we would have a much smoother figure than the*

*frequency polygon of Figure 2.3.4.*

*In general, as the number of observations, n, approaches infinity, and the width*

*of the class intervals approaches zero, the frequency polygon approaches a smooth curve*

*such as is shown in Figure 4.5.2. Such smooth curves are used to represent graphically*

*the distributions of continuous random variables. This has some important consequences*

*when we deal with probability distributions. First, the total area under the curve is equal*

*to one, as was true with the histogram, and the relative frequency of occurrence of val-*

*ues between any two points on the x-axis is equal to the total area bounded by the*

*curve, the x -axis, and perpendicular lines erected at the two points on the x -axis. See*

*Figure 4.5.3. The probability of any specific value of the random variable is zero. This*

*seems logical, since a specific value is represented by a point on the x -axis and the area*

*above a point is zero.*

*Finding Area Under a Smooth Curve With a histogram, as we have seen,*

*subareas of interest can be found by adding areas represented by the cells. We have no cells*

*in the case of a smooth curve, so we must seek an alternate method of finding subareas. Such*

*a method is provided by the integral calculus. To find the area under a smooth curve between*

*any two points a and b, the density function is integrated from a to b. A density function is a*

*formula used to represent the distribution of a continuous random variable. Integration is the*

*limiting case of summation, but we will not perform any integrations, since the level of math-*

*ematics involved is beyond the scope of this book. As we will see later, for all the continu-*

*ous distributions we will consider, there will be an easier way to find areas under their curves.*

*Although the definition of a probability distribution for a continuous random vari-*

*able has been implied in the foregoing discussion, by way of summary, we present it in*

*a more compact form as follows.*

*DEFINITION*

*A nonnegative function f(x) is called a probability distribution (some-*

*times called a probability density function) of the continuous random*

*variable X if the total area bounded by its curve and the x-axis is*

*equal to 1 and if the subarea under the curve bounded by the curve,*

*the x-axis, and perpendiculars erected at any two points a and b give*

*the probability that X is between the points a and b.*

*Thus, the probability of a continuous random variable to assume values between*

*a and b is denoted by P1a 6 X 6 b2*

*4.6 THE NORMAL DISTRIBUTION*

*We come now to the most important distribution in all of statistics—the normal dis-*

*tribution. The formula for this distribution was first published by Abraham De Moivre*

*(1667–1754) on November 12, 1733. Many other mathematicians figure prominently*

*in the history of the normal distribution, including Carl Friedrich Gauss (1777–1855).*

*The distribution is frequently called the Gaussian distribution in recognition of his*

*contributions.*

*The normal density is given by*

*f 1x2 =*

*1*

*22ps*

*2*

*2*

*e -1x -m2 >2s ,*

*-q 6 x 6 q*

*(4.6.1)*

*In Equation 4.6.1, p and e are the familiar constants, 3.14159 . . . and 2.71828 . . . ,*

*respectively, which are frequently encountered in mathematics. The two parameters of the*

*distribution are m, the mean, and s, the standard deviation. For our purposes we may think*

*of m and s of a normal distribution, respectively, as measures of central tendency and dis-*

*persion as discussed in Chapter 2. Since, however, a normally distributed random variable*

*is continuous and takes on values between - q and + q , its mean and standard deviation*

*may be more rigorously defined; but such definitions cannot be given without using calcu-*

*lus. The graph of the normal distribution produces the familiar bell-shaped curve shown in*

*Figure 4.6.1.*

*Characteristics of the Normal Distribution The following are some*

*important characteristics of the normal distribution.*

*1. It is symmetrical about its mean, m. As is shown in Figure 4.6.1, the curve on*

*either side of m is a mirror image of the other side.*

*2. The mean, the median, and the mode are all equal.*

*3. The total area under the curve above the x-axis is one square unit. This character-*

*istic follows from the fact that the normal distribution is a probability distribution.*

*Because of the symmetry already mentioned, 50 percent of the area is to*

*4. If we erect perpendiculars a distance of 1 standard deviation from the mean in both*

*directions, the area enclosed by these perpendiculars, the x-axis, and the curve will*

*be approximately 68 percent of the total area. If we extend these lateral bound-*

*aries a distance of two standard deviations on either side of the mean, approxi-*

*mately 95 percent of the area will be enclosed, and extending them a distance of*

*three standard deviations will cause approximately 99.7 percent of the total area to*

*be enclosed. These approximate areas are illustrated in Figure 4.6.2.*

*5. The normal distribution is completely determined by the parameters m and s. In*

*other words, a different normal distribution is specified for each different value of*

*m and s. Different values of m shift the graph of the distribution along the x-axis*

*as is shown in Figure 4.6.3. Different values of s determine the degree of flatness*

*or peakedness of the graph of the distribution as is shown in Figure 4.6.4. Because*

*of the characteristics of these two parameters, m is often referred to as a location*

*parameter and s is often referred to as a shape parameter.*

*The Standard Normal Distribution The last-mentioned characteristic of*

*the normal distribution implies that the normal distribution is really a family of distribu-*

*tions in which one member is distinguished from another on the basis of the values of*

*m and s. The most important member of this family is the standard normal distribution*

*or unit normal distribution, as it is sometimes called, because it has a mean of 0 and a*

*standard deviation of 1. It may be obtained from Equation 4.6.1 by creating a random*

*variable.*

*z = 1x - m2>s*

*(4.6.2)*

*The equation for the standard normal distribution is written*

*f 1z2 =*

*1*

*22p*

*2*

*e -z >2,*

*-q 6 z 6 q*

*(4.6.3)*

*The graph of the standard normal distribution is shown in Figure 4.6.5.*

*The z-transformation will prove to be useful in the examples and applications that*

*follow. This value of z denotes, for a value of a random variable, the number of stan-*

*dard deviations that value falls above (#z) or below (&z) the mean, which in this case*

*is 0. For example, a z-transformation that yields a value of z " 1 indicates that the value*

*of x used in the transformation is 1 standard deviation above 0. A value of z " &1*

*indicates that the value of x used in the transformation is 1 standard deviation below 0.*

*This property is illustrated in the examples of Section 4.7.*

*To find the probability that z takes on a value between any two points on the z-axis,*

*say, z 0 and z 1, we must find the area bounded by perpendiculars erected at these points,*

*the curve, and the horizontal axis. As we mentioned previously, areas under the curve of*

*a continuous distribution are found by integrating the function between two values of the*

*variable. In the case of the standard normal, then, to find the area between z 0 and z 1*

*directly, we would need to evaluate the following integral:*

*z1*

*Lz 0 22p*

*1*

*2*

*e -z >2 dz*

*Although a closed-form solution for the integral does not exist, we can use numeri-*

*cal methods of calculus to approximate the desired areas beneath the curve to a*

*desired accuracy. Fortunately, we do not have to concern ourselves with such matters,*

*since there are tables available that provide the results of any integration in which we*

*might be interested. Table D in the Appendix is an example of these tables. In the*

*body of Table D are found the areas under the curve between - q and the values of*

*z shown in the leftmost column of the table. The shaded area of Figure 4.6.6 repre-*

*sents the area listed in the table as being between - q and z 0, where z 0 is the spec-*

*ified value of z.*

*We now illustrate the use of Table D by several examples.*

*EXAMPLE 4.6.1*

*Given the standard normal distribution, find the area under the curve, above the z-axis*

*between z = - q and z = 2.*

*Solution:*

*It will be helpful to draw a picture of the standard normal distribution and*

*shade the desired area, as in Figure 4.6.7. If we locate z = 2 in Table D*

*and read the corresponding entry in the body of the table, we find the*

*desired area to be .9772. We may interpret this area in several ways. We*

*may interpret it as the probability that a z picked at random from the pop-*

*ulation of z’s will have a value between - q and 2. We may also interpret*

*it as the relative frequency of occurrence (or proportion) of values of z*

*between - q and 2, or we may say that 97.72 percent of the z’s have a*

*value between - q and 2.*

*■*

*EXAMPLE 4.6.2*

*What is the probability that a z picked at random from the population of z’s will have a*

*value between -2.55 and +2.55?*

*Solution:*

*Figure 4.6.8 shows the area desired. Table D gives us the area between*

*- q and 2.55, which is found by locating 2.5 in the leftmost column of*

*the table and then moving across until we come to the entry in the column*

*headed by 0.05. We find this area to be .9946. If we look at the picture*

*we draw, we see that this is more area than is desired. We need to sub-*

*tract from .9946 the area to the left of -2.55. Reference to Table D shows*

*that the area to the left of -2.55 is .0054. Thus the desired probability is*

*P1-2.55 6 z 6 2.552 = .9946 - .0054 = .9892*

*Suppose we had been asked to find the probability that z is between -2.55 and 2.55 inclu-*

*sive. The desired probability is expressed as P1-2.55 … z … 2.552. Since, as we noted in*

*Section 4.5, P1z = z 02 = 0,P1-2.55 … z … 2.552 = P1-2.55 6 z 6 2.552 = .9892.*

*EXAMPLE 4.6.3*

*What proportion of z values are between -2.74 and 1.53?*

*Solution:*

*Figure 4.6.9 shows the area desired. We find in Table D that the area between*

*- q and 1.53 is .9370, and the area between - q and -2.74 is .0031. To*

*obtain the desired probability we subtract .0031 from .9370. That is,*

*P 1-2.74 … z … 1.532 = .9370 - .0031 = .9339*

*■*

*EXAMPLE 4.6.4*

*Given the standard normal distribution, find P 1z Ú 2.712.*

*Solution:*

*The area desired is shown in Figure 4.6.10. We obtain the area to the right*

*of z = 2.71 by subtracting the area between - q and 2.71 from 1. Thus,*

*P 1z Ú 2.712 = 1 - P 1z … 2.712*

*= 1 - .9966*

*= .0034*

*EXAMPLE 4.6.5*

*Given the standard normal distribution, find P1.84 … z … 2.452.*

*Solution:*

*The area we are looking for is shown in Figure 4.6.11. We first obtain the*

*area between - q and 2.45 and from that subtract the area between - q*

*and .84. In other words,*

*P1.84 … z … 2.452 = P1z … 2.452 - P1z … .842*

*= .9929 - .7995*

*= .1934*

*0*

*.84*

*2.45*

*z*

*FIGURE 4.6.11 Standard normal curve showing*

*P 1.84 … z … 2.452.*

*■*

*EXERCISES*

*Given the standard normal distribution find:*

*4.6.1The area under the curve between z = 0 and z = 1.43*

*4.6.2The probability that a z picked at random will have a value between z = -2.87 and z = 2.64*

*4.6.3P1z Ú .552*

*4.6.4 P1z Ú -.552*

*4.6.54.6.6 P1z 6 2.332*

*4.6.7P 1z 6 -2.332*

*P1-1.96 … z … 1.962*

*4.6.8 P1-2.58 … z … 2.582*

*4.6.9P1-1.65 … z … 1.652*

*4.6.10 P1z = .742*

*Given the following probabilities, find z1:*

*4.6.11*

*4.6.13*

*4.6.15*

*P1z … z 12 = .0055*

*P1z 7 z 12 = .0384*

*4.6.12 P1-2.67 … z … z 12 = .9718*

*4.6.14 P1z 1 … z … 2.982 = .1117*

*P1-z 1 … z … z 12 = .8132*

*4.7 NORMAL DISTRIBUTION APPLICATIONS*

*Although its importance in the field of statistics is indisputable, one should realize that*

*the normal distribution is not a law that is adhered to by all measurable characteris-*

*tics occurring in nature. It is true, however, that many of these characteristics are*

*approximately normally distributed. Consequently, even though no variable encoun-*

*tered in practice is precisely normally distributed, the normal distribution can be used*

*to model the distribution of many variables that are of interest. Using the normal dis-*

*tribution as a model allows us to make useful probability statements about some vari-*

*ables much more conveniently than would be the case if some more complicated model*

*had to be used.*

*Human stature and human intelligence are frequently cited as examples of vari-*

*ables that are approximately normally distributed. On the other hand, many distributions*

*relevant to the health field cannot be described adequately by a normal distribution.*

*Whenever it is known that a random variable is approximately normally distributed, or*

*when, in the absence of complete knowledge, it is considered reasonable to make this*

*assumption, the statistician is aided tremendously in his or her efforts to solve practical*

*problems relative to this variable. Bear in mind, however, that “normal” in this context*

*refers to the statistical properties of a set of data and in no way connotes normality in*

*the sense of health or medical condition.*

*There are several other reasons why the normal distribution is so important in sta-*

*tistics, and these will be considered in due time. For now, let us see how we may answer*

*simple probability questions about random variables when we know, or are willing to*

*assume, that they are, at least, approximately normally distributed.*

*EXAMPLE 4.7.1*

*The Uptimer is a custom-made lightweight battery-operated activity monitor that records*

*the amount of time an individual spends in the upright position. In a study of children*

*ages 8 to 15 years, Eldridge et al. (A-10) studied 529 normally developing children who*

*each wore the Uptimer continuously for a 24-hour period that included a typical school*

*day. The researchers found that the amount of time children spent in the upright position*

*followed a normal distribution with a mean of 5.4 hours and standard deviation of 1.3*

*hours. Assume that this finding applies to all children 8 to 15 years of age. Find the prob-*

*ability that a child selected at random spends less than 3 hours in the upright position in*

*a 24-hour period.*

*Solution:*

*First let us draw a picture of the distribution and shade the area correspon-*

*ding to the probability of interest. This has been done in Figure 4.7.1.*

*s = 1.3*

*3.0*

*m = 5.4*

*x*

*FIGURE 4.7.1 Normal distribution to approximate*

*distribution of amount of time children spent in upright*

*position (mean and standard deviation estimated).*

*FIGURE 4.7.2 Normal distribution of time spent*

*upright 1x2 and the standard normal distribution 1z2.*

*If our distribution were the standard normal distribution with a mean*

*of 0 and a standard deviation of 1, we could make use of Table D and find*

*the probability with little effort. Fortunately, it is possible for any normal*

*distribution to be transformed easily to the standard normal. What we do*

*is transform all values of X to corresponding values of z. This means that*

*the mean of X must become 0, the mean of z. In Figure 4.7.2 both distri-*

*butions are shown. We must determine what value of z, say, z 0, corresponds*

*to an x of 3.0. This is done using formula 4.6.2, z = 1x - m2>s, which*

*transforms any value of x in any normal distribution to the corresponding*

*value of z in the standard normal distribution. For the present example we*

*have*

*z =*

*3.0 - 5.4*

*= -1.85*

*1.3*

*The value of z 0 we seek, then, is -1.85.*

*■*

*Let us examine these relationships more closely. It is seen that the distance from the*

*mean, 5.4, to the x-value of interest, 3.0, is 3.0 - 5.4 = -2.4, which is a distance of*

*1.85 standard deviations. When we transform x values to z values, the distance of the z*

*value of interest from its mean, 0, is equal to the distance of the corresponding x value*

*from its mean, 5.4, in standard deviation units. We have seen that this latter distance is*

*1.85 standard deviations. In the z distribution a standard deviation is equal to 1, and con-*

*sequently the point on the z scale located a distance of 1.85 standard deviations below*

*0 is z = -1.85, the result obtained by employing the formula. By consulting Table D,*

*we find that the area to the left of z = -1.85 is .0322. We may summarize this discus-*

*sion as follows:*

*P1x 6 3.02 = Pa z 6*

*3.0 - 5.4*

*b = P1z 6 -1.852 = .0322*

*1.3*

*To answer the original question, we say that the probability is .0322 that a randomly*

*selected child will have uptime of less than 3.0 hours.*

*EXAMPLE 4.7.2*

*Diskin et al. (A-11) studied common breath metabolites such as ammonia, acetone, iso-*

*prene, ethanol, and acetaldehyde in five subjects over a period of 30 days. Each day,*

*breath samples were taken and analyzed in the early morning on arrival at the labora-*

*tory. For subject A, a 27-year-old female, the ammonia concentration in parts per billion*

*(ppb) followed a normal distribution over 30 days with mean 491 and standard devia-*

*tion 119. What is the probability that on a random day, the subject’s ammonia concen-*

*tration is between 292 and 649 ppb?*

*Solution:*

*In Figure 4.7.3 are shown the distribution of ammonia concentrations and*

*the z distribution to which we transform the original values to determine*

*the desired probabilities. We find the z value corresponding to an x of*

*292 by*

*Similarly, for x = 649 we have*

*z =*

*649 - 491*

*= 1.33*

*119*

*From Table D we find the area between - q and -1.67 to be .0475 and the*

*area between - q and 1.33 to be .9082. The area desired is the difference*

*between these, .9082 - .0475 = .8607. To summarize,*

*292 - 491*

*649 - 491*

*… z …*

*b*

*119*

*119*

*= P1-1.67 … z … 1.332*

*= P1- q … z … 1.332 - P1- q … z … -1.672*

*= .9082 - .0475*

*= .8607*

*P1292 … x … 6492 = Pa*

*The probability asked for in our original question, then, is .8607.*

*■*

*EXAMPLE 4.7.3*

*In a population of 10,000 of the children described in Example 4.7.1, how many would*

*you expect to be upright more than 8.5 hours?*

*Solution:*

*We first find the probability that one child selected at random from the pop-*

*ulation would be upright more than 8.5 hours. That is,*

*P1x Ú 8.52 = Paz Ú*

*8.5 - 5.4*

*b = P1z Ú 2.382 = 1 - .9913 = .0087*

*1.3*

*Out of 10,000 people we would expect 10,0001.00872 = 87 to spend more*

*than 8.5 hours upright.*

*■*

*We may use MINITAB to calculate cumulative standard normal probabilities. Suppose*

*we wish to find the cumulative probabilities for the following values of z: -3, -2, -1,*

*0, 1, 2, and 3. We enter the values of z into Column 1 and proceed as shown in Fig-*

*ure 4.7.4.*

*The preceding two sections focused extensively on the normal distribution, the*

*most important and most frequently used continuous probability distribution. Though*

*much of what will be covered in the next several chapters uses this distribution, it is not*

*the only important continuous probability distribution. We will be introducing several*

*other continuous distributions later in the text, namely the t-distribution, the chi-square*

*distribution, and the F-distribution. The details of these distributions will be discussed*

*in the chapters in which we need them for inferential tests.*

*EXERCISES*

*4.7.1For another subject (a 29-year-old male) in the study by Diskin et al. (A-11), acetone levels were*

*normally distributed with a mean of 870 and a standard deviation of 211 ppb. Find the probability*

*that on a given day the subject’s acetone level is:*

*(a) Between 600 and 1000 ppb*

*(b) Over 900 ppb*

*(c) Under 500 ppb*

*(d) Between 900 and 1100 ppb*

*4.7.2In the study of fingerprints, an important quantitative characteristic is the total ridge count for the*

*10 fingers of an individual. Suppose that the total ridge counts of individuals in a certain popula-*

*tion are approximately normally distributed with a mean of 140 and a standard deviation of 50.*

*Find the probability that an individual picked at random from this population will have a ridge*

*count of:*

*(a) 200 or more*

*(b) Less than 100*

*(c) Between 100 and 200*

*(d) Between 200 and 250*

*(e) In a population of 10,000 people how many would you expect to have a ridge count of 200*

*or more?*

*4.7.3One of the variables collected in the North Carolina Birth Registry data (A-3) is pounds gained*

*during pregnancy. According to data from the entire registry for 2001, the number of pounds gained*

*during pregnancy was approximately normally distributed with a mean of 30.23 pounds and a stan-*

*dard deviation of 13.84 pounds. Calculate the probability that a randomly selected mother in North*

*Carolina in 2001 gained:*

*(a) Less than 15 pounds during pregnancy*

*(b) More than 40 pounds*

*(c) Between 14 and 40 pounds*

*(d) Less than 10 pounds*

*(e) Between 10 and 20 pounds*

*4.7.4Suppose the average length of stay in a chronic disease hospital of a certain type of patient is*

*60 days with a standard deviation of 15. If it is reasonable to assume an approximately normal*

*distribution of lengths of stay, find the probability that a randomly selected patient from this group*

*will have a length of stay:*

*(a) Greater than 50 days*

*(b) Less than 30 days*

*(c) Between 30 and 60 days*

*(d) Greater than 90 days*

*4.7.5If the total cholesterol values for a certain population are approximately normally distributed with a*

*mean of 200 mg/100 ml and a standard deviation of 20 mg/100 ml, find the probability that an indi-*

*vidual picked at random from this population will have a cholesterol value:*

*(a) Between 180 and 200 mg/100 ml*

*(b) Greater than 225 mg/100 ml*

*(c) Less than 150 mg/100 ml*

*(d) Between 190 and 210 mg/100 ml*

*4.7.6 Given a normally distributed population with a mean of 75 and a variance of 625, find:*

*(a) P150 … x … 1002 (b) P1x 7 902*

*(c) P1x 6 602*

*(d) P1x Ú 852*

*(e) P130 … x … 1102*

*4.7.7 The weights of a certain population of young adult females are approximately normally distrib-*

*uted with a mean of 132 pounds and a standard deviation of 15. Find the probability that a sub-*

*ject selected at random from this population will weigh:*

*(a) More than 155 pounds*

*(b) 100 pounds or less*

*(c) Between 105 and 145 pounds*

*4.8 SUMMARY*

*In this chapter the concepts of probability described in the preceding chapter are further*

*developed. The concepts of discrete and continuous random variables and their proba-*

*bility distributions are discussed. In particular, two discrete probability distributions, the*

*binomial and the Poisson, and one continuous probability distribution, the normal, are*

*examined in considerable detail. We have seen how these theoretical distributions allow*

*us to make probability statements about certain random variables that are of interest to*

*the health professional.*