
SPECIAL DISTRIBUTIONS

Chapter 5

- | | | | |
|-----|--|------|------------------------------------|
| 5.1 | Introduction | 5.7 | The Gamma Distributions |
| 5.2 | The Bernoulli and Binomial Distributions | 5.8 | The Beta Distributions |
| 5.3 | The Hypergeometric Distributions | 5.9 | The Multinomial Distributions |
| 5.4 | The Poisson Distributions | 5.10 | The Bivariate Normal Distributions |
| 5.5 | The Negative Binomial Distributions | 5.11 | Supplementary Exercises |
| 5.6 | The Normal Distributions | | |

5.1 Introduction

In this chapter, we shall define and discuss several special families of distributions that are widely used in applications of probability and statistics. The distributions that will be presented here include discrete and continuous distributions of univariate, bivariate, and multivariate types. The discrete univariate distributions are the families of Bernoulli, binomial, hypergeometric, Poisson, negative binomial, and geometric distributions. The continuous univariate distributions are the families of normal, lognormal, gamma, exponential, and beta distributions. Other continuous univariate distributions (introduced in exercises and examples) are the families of Weibull and Pareto distributions. Also discussed is the multinomial family of multivariate discrete distributions, and the bivariate normal family of bivariate continuous distributions.

We shall briefly describe how each of these families of distributions arise in applied problems and show why each might be an appropriate probability model for some experiment. For each family, we shall present the form of the p.f. or the p.d.f. and discuss some of the basic properties of the distributions in the family.

The list of distributions presented in this chapter, or in this entire text for that matter, is not intended to be exhaustive. These distributions are known to be useful in a wide variety of applied problems. In many real-world problems, however, one will need to consider other distributions not mentioned here. The tools that we develop for use with these distributions can be generalized for use with other distributions. Our purpose in providing in-depth presentations of the most popular distributions here is to give the reader a feel for how to use probability to model the variation and uncertainty in applied problems as well as some of the tools that get used during probability modeling.

5.2 The Bernoulli and Binomial Distributions

The simplest type of experiment has only two possible outcomes, call them 0 and 1. If X equals the outcome from such an experiment, then X has the simplest type of nondegenerate distribution, which is a member of the family of Bernoulli distributions. If n independent random variables X_1, \dots, X_n all have the same

Bernoulli distribution, then their sum is equal to the number of the X_i 's that equal 1, and the distribution of the sum is a member of the binomial family.

The Bernoulli Distributions

Example 5.2.1

A Clinical Trial. The treatment given to a particular patient in a clinical trial can either succeed or fail. Let $X = 0$ if the treatment fails, and let $X = 1$ if the treatment succeeds. All that is needed to specify the distribution of X is the value $p = \Pr(X = 1)$ (or, equivalently, $1 - p = \Pr(X = 0)$). Each different p corresponds to a different distribution for X . The collection of all such distributions corresponding to all $0 \leq p \leq 1$ form the family of Bernoulli distributions. ◀

An experiment of a particularly simple type is one in which there are only two possible outcomes, such as head or tail, success or failure, defective or nondefective, patient recovers or does not recover. It is convenient to designate the two possible outcomes of such an experiment as 0 and 1, as in Example 5.2.1. The following recap of Definition 3.1.5 can then be applied to every experiment of this type.

Definition 5.2.1

Bernoulli Distribution. A random variable X has the *Bernoulli distribution with parameter p* ($0 \leq p \leq 1$) if X can take only the values 0 and 1 and the probabilities are

$$\Pr(X = 1) = p \quad \text{and} \quad \Pr(X = 0) = 1 - p. \quad (5.2.1)$$

The p.f. of X can be written as follows:

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & \text{for } x = 0, 1, \\ 0 & \text{otherwise.} \end{cases} \quad (5.2.2)$$

To verify that this p.f. $f(x|p)$ actually does represent the Bernoulli distribution specified by the probabilities (5.2.1), it is simply necessary to note that $f(1|p) = p$ and $f(0|p) = 1 - p$.

If X has the Bernoulli distribution with parameter p , then X^2 and X are the same random variable. It follows that

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p,$$

$$E(X^2) = E(X) = p,$$

and

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = p(1 - p).$$

Furthermore, the m.g.f. of X is

$$\psi(t) = E(e^{tX}) = pe^t + (1 - p) \quad \text{for } -\infty < t < \infty.$$

Definition 5.2.2

Bernoulli Trials/Process. If the random variables in a finite or infinite sequence X_1, X_2, \dots are i.i.d., and if each random variable X_i has the Bernoulli distribution with parameter p , then it is said that X_1, X_2, \dots are *Bernoulli trials with parameter p* . An infinite sequence of Bernoulli trials is also called a *Bernoulli process*.

Example 5.2.2

Tossing a Coin. Suppose that a fair coin is tossed repeatedly. Let $X_i = 1$ if a head is obtained on the i th toss, and let $X_i = 0$ if a tail is obtained ($i = 1, 2, \dots$). Then the random variables X_1, X_2, \dots are Bernoulli trials with parameter $p = 1/2$. ◀

Example 5.2.3

Defective Parts. Suppose that 10 percent of the items produced by a certain machine are defective and the parts are independent of each other. We will sample n items at random and inspect them. Let $X_i = 1$ if the i th item is defective, and let $X_i = 0$ if it is nondefective ($i = 1, \dots, n$). Then the variables X_1, \dots, X_n form n Bernoulli trials with parameter $p = 1/10$. ◀

Example 5.2.4

Clinical Trials. In the many clinical trial examples in earlier chapters (Example 4.7.8, for instance), the random variables X_1, X_2, \dots , indicating whether each patient is a success, were conditionally Bernoulli trials with parameter p given $P = p$, where P is the unknown proportion of patients in a very large population who recover. ◀

The Binomial Distributions

Example 5.2.5

Defective Parts. In Example 5.2.3, let $X = X_1 + \dots + X_{10}$, which equals the number of defective parts among the 10 sampled parts. What is the distribution of X ? ◀

As derived after Example 3.1.9, the distribution of X in Example 5.2.5 is the binomial distribution with parameters 10 and 1/10. We repeat the general definition of binomial distributions here.

Definition 5.2.3

Binomial Distribution. A random variable X has the *binomial distribution with parameters n and p* if X has a discrete distribution for which the p.f. is as follows:

$$f(x|n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \dots, n, \\ 0 & \text{otherwise.} \end{cases} \quad (5.2.3)$$

In this distribution, n must be a positive integer, and p must lie in the interval $0 \leq p \leq 1$.

Probabilities for various binomial distributions can be obtained from the table given at the end of this book and from many statistical software programs.

The binomial distributions are of fundamental importance in probability and statistics because of the following result, which was derived in Sec. 3.1 and which we restate here in the terminology of this chapter.

Theorem 5.2.1

If the random variables X_1, \dots, X_n form n Bernoulli trials with parameter p , and if $X = X_1 + \dots + X_n$, then X has the binomial distribution with parameters n and p . ■

When X is represented as the sum of n Bernoulli trials as in Theorem 5.2.1, the values of the mean, variance, and m.g.f. of X can be derived very easily. These values, which were already obtained in Example 4.2.5 and on pages 231 and 238, are

$$E(X) = \sum_{i=1}^n E(X_i) = np,$$

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) = np(1-p),$$

and

$$\psi(t) = E(e^{tX}) = \prod_{i=1}^n E(e^{tX_i}) = (pe^t + 1 - p)^n. \quad (5.2.4)$$

The reader can use the m.g.f. in Eq. (5.2.4) to establish the following simple extension of Theorem 4.4.6.

Theorem
5.2.2

If X_1, \dots, X_k are independent random variables, and if X_i has the binomial distribution with parameters n_i and p ($i = 1, \dots, k$), then the sum $X_1 + \dots + X_k$ has the binomial distribution with parameters $n = n_1 + \dots + n_k$ and p . ■

Theorem 5.2.2 also follows easily if we represent each X_i as the sum of n_i Bernoulli trials with parameter p . If $n = n_1 + \dots + n_k$, and if all n trials are independent, then the sum $X_1 + \dots + X_k$ will simply be the sum of n Bernoulli trials with parameter p . Hence, this sum must have the binomial distribution with parameters n and p .

Example
5.2.6

Castaneda v. Partida. Courts have used the binomial distributions to calculate probabilities of jury compositions from populations with known racial and ethnic compositions. In the case of *Castaneda v. Partida*, 430 U.S. 482 (1977), a local population was 79.1 percent Mexican American. During a 2.5-year period, there were 220 persons called to serve on grand juries, but only 100 were Mexican Americans. The claim was made that this was evidence of discrimination against Mexican Americans in the grand jury selection process. The court did a calculation under the assumption that grand jurors were drawn at random and independently from the population each with probability 0.791 of being Mexican American. Since the claim was that 100 was too small a number of Mexican Americans, the court calculated the probability that a binomial random variable X with parameters 220 and 0.791 would be 100 or less. The probability is very small (less than 10^{-25}). Is this evidence of discrimination against Mexican Americans? The small probability was calculated under the assumption that X had the binomial distribution with parameters 220 and 0.791, which means that the court was assuming that there was no discrimination against Mexican Americans when performing the calculation. In other words, the small probability is the conditional probability of observing $X \leq 100$ given that there is no discrimination. What should be more interesting to the court is the reverse conditional probability, namely, the probability that there is no discrimination given that $X = 100$ (or given $X \leq 100$). This sounds like a case for Bayes' theorem. After we introduce the beta distributions in Sec. 5.8, we shall show how to use Bayes' theorem to calculate this probability (Examples 5.8.3 and 5.8.4). ◀

Note: Bernoulli and Binomial Distributions. Every random variable that takes only the two values 0 and 1 must have a Bernoulli distribution. However, not every sum of Bernoulli random variables has a binomial distribution. There are two conditions needed to apply Theorem 5.2.1. The Bernoulli random variables must be mutually independent, and they must all have the same parameter. If either of these conditions fails, the distribution of the sum will not be a binomial distribution. When the court did a binomial calculation in Example 5.2.6, it was defining “no discrimination” to mean that jurors were selected independently and with the same probability 0.791 of being Mexican American. If the court had defined “no discrimination” some other way, they would have needed to do a different, presumably more complicated, probability calculation.

We conclude this section with an example that shows how Bernoulli and binomial calculations can improve efficiency when data collection is costly.

Example
5.2.7

Group Testing. Military and other large organizations are often faced with the need to test large numbers of members for rare diseases. Suppose that each test requires

a small amount of blood, and it is guaranteed to detect the disease if it is anywhere in the blood. Suppose that 1000 people need to be tested for a disease that affects $1/5$ of 1 percent of all people. Let $X_j = 1$ if person j has the disease and $X_j = 0$ if not, for $j = 1, \dots, 1000$. We model the X_j as i.i.d. Bernoulli random variables with parameter 0.002 for $j = 1, \dots, 1000$. The most naïve approach would be to perform 1000 tests to see who has the disease. But if the tests are costly, there may be a more economical way to test. For example, one could divide the 1000 people into 10 groups of size 100 each. For each group, take a portion of the blood sample from each of the 100 people in the group and combine them into one sample. Then test each of the 10 combined samples. If none of the 10 combined samples has the disease, then nobody has the disease, and we needed only 10 tests instead of 1000. If only one of the combined samples has the disease, then we can test those 100 people separately, and we needed only 110 tests.

In general, let $Z_{1,i}$ be the number of people in group i who have the disease for $i = 1, \dots, 10$. Then each $Z_{1,i}$ has the binomial distribution with parameters 100 and 0.002. Let $Y_{1,i} = 1$ if $Z_{1,i} > 0$ and $Y_{1,i} = 0$ if $Z_{1,i} = 0$. Then each $Y_{1,i}$ has the Bernoulli distribution with parameter

$$\Pr(Z_{1,i} > 0) = 1 - \Pr(Z_{1,i} = 0) = 1 - 0.998^{100} = 0.181,$$

and they are independent. Then $Y_1 = \sum_{i=1}^{10} Y_{1,i}$ is the number of groups whose members we have to test individually. Also, Y_1 has the binomial distribution with parameters 10 and 0.181. The number of people that we need to test individually is $100Y_1$. The mean of $100Y_1$ is $100 \times 10 \times 0.181 = 181$. So, the expected total number of tests is $10 + 181 = 191$, rather than 1000. One can compute the entire distribution of the total number of tests, $100Y_1 + 10$. The maximum number of tests needed by this group testing procedure is 1010, which would be the case if all 10 groups had at least one person with the disease, but this has probability 3.84×10^{-8} . In all other cases, group testing requires fewer than 1000 tests.

There are multiple-stage versions of group testing in which each of the groups that tests positive is split further into subgroups which are each tested together. If each of those subgroups is sufficiently large, they can be further subdivided into smaller sub-subgroups, etc. Finally, only the final-stage subgroups that have a positive result are tested individually. This can further reduce the expected number of tests. For example, consider the following two-stage version of the procedure described earlier. We could divide each of the 10 groups of 100 people into 10 subgroups of 10 people each. Following the above notation, let $Z_{2,i,k}$ be the number of people in subgroup k of group i who have the disease, for $i = 1, \dots, 10$ and $k = 1, \dots, 10$. Then each $Z_{2,i,k}$ has the binomial distribution with parameters 10 and 0.002. Let $Y_{2,i,k} = 1$ if $Z_{2,i,k} > 0$ and $Y_{2,i,k} = 0$ otherwise. Notice that $Y_{2,i,k} = 0$ for $k = 1, \dots, 10$ for every i such that $Y_{1,i} = 0$. So, we only need to test individuals in those subgroups such that $Y_{2,i,k} = 1$. Each $Y_{2,i,k}$ has the Bernoulli distribution with parameter

$$\Pr(Z_{2,i,k} > 0) = 1 - \Pr(Z_{2,i,k} = 0) = 1 - 0.998^{10} = 0.0198,$$

and they are independent. Then $Y_2 = \sum_{i=1}^{10} \sum_{k=1}^{10} Y_{2,i,k}$ is the number of groups whose members we have to test individually. Also, Y_2 has the binomial distribution with parameters 100 and 0.0198. The number of people that we need to test individually is $10Y_2$. The mean of $10Y_2$ is $10 \times 100 \times 0.0198 = 19.82$. The number of subgroups that we need to test in the second stage is Y_1 , whose mean is 1.81. So, the expected total number of tests is $10 + 1.81 + 19.82 = 31.63$, which is even smaller than the 191 for the one-stage procedure described earlier. ◀

Summary

A random variable X has the Bernoulli distribution with parameter p if the p.f. of X is $f(x|p) = p^x(1-p)^{1-x}$ for $x = 0, 1$ and 0 otherwise. If X_1, \dots, X_n are i.i.d. random variables all having the Bernoulli distribution with parameter p , then we refer to X_1, \dots, X_n as Bernoulli trials, and $X = \sum_{i=1}^n X_i$ has the binomial distribution with parameters n and p . Also, X is the number of successes in the n Bernoulli trials, where success on trial i corresponds to $X_i = 1$ and failure corresponds to $X_i = 0$.

Exercises

1. Suppose that X is a random variable such that $E(X^k) = 1/3$ for $k = 1, 2, \dots$. Assuming that there cannot be more than one distribution with this same sequence of moments (see Exercise 14), determine the distribution of X .

2. Suppose that a random variable X can take only the two values a and b with the following probabilities:

$$\Pr(X = a) = p \quad \text{and} \quad \Pr(X = b) = 1 - p.$$

Express the p.f. of X in a form similar to that given in Eq. (5.2.2).

3. Suppose that a fair coin (probability of heads equals $1/2$) is tossed independently 10 times. Use the table of the binomial distribution given at the end of this book to find the probability that strictly more heads are obtained than tails.

4. Suppose that the probability that a certain experiment will be successful is 0.4, and let X denote the number of successes that are obtained in 15 independent performances of the experiment. Use the table of the binomial distribution given at the end of this book to determine the value of $\Pr(6 \leq X \leq 9)$.

5. A coin for which the probability of heads is 0.6 is tossed nine times. Use the table of the binomial distribution given at the end of this book to find the probability of obtaining an even number of heads.

6. Three men A , B , and C shoot at a target. Suppose that A shoots three times and the probability that he will hit the target on any given shot is $1/8$, B shoots five times and the probability that he will hit the target on any given shot is $1/4$, and C shoots twice and the probability that he will hit the target on any given shot is $1/2$. What is the expected number of times that the target will be hit?

7. Under the conditions of Exercise 6, assume also that all shots at the target are independent. What is the variance of the number of times that the target will be hit?

8. A certain electronic system contains 10 components. Suppose that the probability that each individual component will fail is 0.2 and that the components fail inde-

pendently of each other. Given that at least one of the components has failed, what is the probability that at least two of the components have failed?

9. Suppose that the random variables X_1, \dots, X_n form n Bernoulli trials with parameter p . Determine the conditional probability that $X_1 = 1$, given that

$$\sum_{i=1}^n X_i = k \quad (k = 1, \dots, n).$$

10. The probability that each specific child in a given family will inherit a certain disease is p . If it is known that at least one child in a family of n children has inherited the disease, what is the expected number of children in the family who have inherited the disease?

11. For $0 \leq p \leq 1$, and $n = 2, 3, \dots$, determine the value of

$$\sum_{x=2}^n x(x-1) \binom{n}{x} p^x (1-p)^{n-x}.$$

12. If a random variable X has a discrete distribution for which the p.f. is $f(x)$, then the value of x for which $f(x)$ is maximum is called the *mode* of the distribution. If this same maximum $f(x)$ is attained at more than one value of x , then all such values of x are called *modes* of the distribution. Find the mode or modes of the binomial distribution with parameters n and p . *Hint*: Study the ratio $f(x+1|n, p)/f(x|n, p)$.

13. In a clinical trial with two treatment groups, the probability of success in one treatment group is 0.5, and the probability of success in the other is 0.6. Suppose that there are five patients in each group. Assume that the outcomes of all patients are independent. Calculate the probability that the first group will have at least as many successes as the second group.

14. In Exercise 1, we assumed that there could be at most one distribution with moments $E(X^k) = 1/3$ for $k = 1, 2, \dots$. In this exercise, we shall prove that there can be only one such distribution. Prove the following

facts and show that they imply that at most one distribution has the given moments.

- a. $\Pr(|X| \leq 1) = 1$. (If not, show that $\lim_{k \rightarrow \infty} E(X^{2k}) = \infty$.)
- b. $\Pr(X^2 \in \{0, 1\}) = 1$. (If not, prove that $E(X^4) < E(X^2)$.)
- c. $\Pr(X = -1) = 0$. (If not, prove that $E(X) < E(X^2)$.)

15. In Example 5.2.7, suppose that we use the two-stage version described at the end of the example. What is the maximum number of tests that could possibly be needed

by this version? What is the probability that the maximum number of tests would be required?

16. For the 1000 people in Example 5.2.7, suppose that we use the following three-stage group testing procedure. First, divide the 1000 people into five groups of size 200 each. For each group that tests positive, further divide it into five subgroups of size 40 each. For each subgroup that tests positive, further divide it into five sub-subgroups of size 8 each. For each sub-subgroup that tests positive, test all eight people. Find the expected number and maximum number of tests.

5.3 The Hypergeometric Distributions

In this section, we consider dependent Bernoulli random variables. A common source of dependent Bernoulli random variables is sampling without replacement from a finite population. Suppose that a finite population consists of a known number of successes and failures. If we sample a fixed number of units from that population, the number of successes in our sample will have a distribution that is a member of the family of hypergeometric distributions.

Definition and Examples

Example 5.3.1

Sampling without Replacement. Suppose that a box contains A red balls and B blue balls. Suppose also that $n \geq 0$ balls are selected at random from the box without replacement, and let X denote the number of red balls that are obtained. Clearly, we must have $n \leq A + B$ or we would run out of balls. Also, if $n = 0$, then $X = 0$ because there are no balls, red or blue, drawn. For cases with $n \geq 1$, we can let $X_i = 1$ if the i th ball drawn is red and $X_i = 0$ if not. Then each X_i has a Bernoulli distribution, but X_1, \dots, X_n are not independent in general. To see this, assume that both $A > 0$ and $B > 0$ as well as $n \geq 2$. We will now show that $\Pr(X_2 = 1|X_1 = 0) \neq \Pr(X_2 = 1|X_1 = 1)$. If $X_1 = 1$, then when the second ball is drawn there are only $A - 1$ red balls remaining out of a total of $A + B - 1$ available balls. Hence, $\Pr(X_2 = 1|X_1 = 1) = (A - 1)/(A + B - 1)$. By the same reasoning,

$$\Pr(X_2 = 1|X_1 = 0) = \frac{A}{A + B - 1} > \frac{A - 1}{A + B - 1}.$$

Hence, X_2 is not independent of X_1 , and we should not expect X to have a binomial distribution. ◀

The problem described in Example 5.3.1 is a template for all cases of sampling without replacement from a finite population with only two types of objects. Anything that we learn about the random variable X in Example 5.3.1 will apply to every case of sampling without replacement from finite populations with only two types of objects. First, we derive the distribution of X .

Theorem 5.3.1 Probability Function. The distribution of X in Example 5.3.1 has the p.f.

$$f(x|A, B, n) = \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}}, \quad (5.3.1)$$

for

$$\max\{0, n - B\} \leq x \leq \min\{n, A\}, \quad (5.3.2)$$

and $f(x|A, B, n) = 0$ otherwise.

Proof Clearly, the value of X can neither exceed n nor exceed A . Therefore, it must be true that $X \leq \min\{n, A\}$. Similarly, because the number of blue balls $n - X$ that are drawn cannot exceed B , the value of X must be at least $n - B$. Because the value of X cannot be less than 0, it must be true that $X \geq \max\{0, n - B\}$. Hence, the value of X must be an integer in the interval in (5.3.2).

We shall now find the p.f. of X using combinatorial arguments from Sec. 1.8. The degenerate cases, those with A , B , and/or n equal to 0, are easy to prove because $\binom{k}{0} = 1$ for all nonnegative k , including $k = 0$. For the cases in which all of A , B , and n are strictly positive, there are $\binom{A+B}{n}$ ways to choose n balls out of the $A + B$ available balls, and all of these choices are equally likely. For each integer x in the interval (5.3.2), there are $\binom{A}{x}$ ways to choose x red balls, and for each such choice there are $\binom{B}{n-x}$ ways to choose $n - x$ blue balls. Hence, the probability of obtaining exactly x red balls out of n is given by Eq. (5.3.1). Furthermore, $f(x|A, B, n)$ must be 0 for all other values of x , because all other values are impossible. ■

Definition 5.3.1 Hypergeometric Distribution. Let A , B , and n be nonnegative integers with $n \leq A + B$. If a random variable X has a discrete distribution with p.f. as in Eqs. (5.3.1) and (5.3.2), then it is said that X has the *hypergeometric distribution with parameters A , B , and n* .

Example 5.3.2 Sampling without Replacement from an Observed Data Set. Consider the patients in the clinical trial whose results are tabulated in Table 2.1. We might need to reexamine a subset of the patients in the placebo group. Suppose that we need to sample 11 distinct patients from the 34 patients in that group. What is the distribution of the number of successes (no relapse) that we obtain in the subsample? Let X stand for the number of successes in the subsample. Table 2.1 indicates that there are 10 successes and 24 failures in the placebo group. According to the definition of the hypergeometric distribution, X has the hypergeometric distribution with parameters $A = 10$, $B = 24$, and $n = 11$. In particular, the possible values of X are the integers from 0 to 10. Even though we sample 11 patients, we cannot observe 11 successes, since only 10 successes are available. ◀

The Mean and Variance for a Hypergeometric Distribution

Theorem 5.3.2 Mean and Variance. Let X have a hypergeometric distribution with strictly positive parameters A , B , and n . Then

$$E(X) = \frac{nA}{A+B}, \quad (5.3.3)$$

$$\text{Var}(X) = \frac{nAB}{(A+B)^2} \cdot \frac{A+B-n}{A+B-1}. \quad (5.3.4)$$

Proof Assume that X is as defined in Example 5.3.1, the number of red balls drawn when n balls are selected at random without replacement from a box containing A red balls and B blue balls. For $i = 1, \dots, n$, let $X_i = 1$ if the i th ball that is selected is red, and let $X_i = 0$ if the i th ball is blue. As explained in Example 4.2.4, we can imagine that the n balls are selected from the box by first arranging all the balls in the box in some random order and then selecting the first n balls from this arrangement. It can be seen from this interpretation that, for $i = 1, \dots, n$,

$$\Pr(X_i = 1) = \frac{A}{A+B} \quad \text{and} \quad \Pr(X_i = 0) = \frac{B}{A+B}.$$

Therefore, for $i = 1, \dots, n$,

$$E(X_i) = \frac{A}{A+B} \quad \text{and} \quad \text{Var}(X_i) = \frac{AB}{(A+B)^2}. \quad (5.3.5)$$

Since $X = X_1 + \dots + X_n$, the mean of X is the sum of the means of the X_i 's, namely, Eq. (5.3.3).

Next, use Theorem 4.6.7 to write

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j). \quad (5.3.6)$$

Because of the symmetry among the random variables X_1, \dots, X_n , every term $\text{Cov}(X_i, X_j)$ in the final summation in Eq. (5.3.6) will have the same value as $\text{Cov}(X_1, X_2)$. Since there are $\binom{n}{2}$ terms in this summation, it follows from Eqs. (5.3.5) and (5.3.6) that

$$\text{Var}(X) = \frac{nAB}{(A+B)^2} + n(n-1) \text{Cov}(X_1, X_2). \quad (5.3.7)$$

We could compute $\text{Cov}(X_1, X_2)$ directly, but it is simpler to argue as follows. If $n = A + B$, then $\Pr(X = A) = 1$ because *all* the balls in the box will be selected without replacement. Thus, for $n = A + B$, X is a constant random variable and $\text{Var}(X) = 0$. Setting Eq. (5.3.7) to 0 and solving for $\text{Cov}(X_1, X_2)$ gives

$$\text{Cov}(X_1, X_2) = -\frac{AB}{(A+B)^2(A+B-1)}.$$

Plugging this value back into Eq. (5.3.7) gives Eq. (5.3.4). ■

Comparison of Sampling Methods

If we had sampled *with* replacement in Example 5.3.1, the number of red balls would have the binomial distribution with parameters n and $A/(A+B)$. In that case, the mean number of red balls would still be $nA/(A+B)$, but the variance would be different. To see how the variances from sampling with and without replacement are related, let $T = A + B$ denote the total number of balls in the box, and let $p = A/T$ denote the proportion of red balls in the box. Then Eq. (5.3.4) can be rewritten as follows:

$$\text{Var}(X) = np(1-p) \frac{T-n}{T-1}. \quad (5.3.8)$$

The variance $np(1-p)$ of the binomial distribution is the variance of the number of red balls when sampling with replacement. The factor $\alpha = (T-n)/(T-1)$ in Eq. (5.3.8) therefore represents the reduction in $\text{Var}(X)$ caused by sampling without replacement from a finite population. This α is called the *finite population correction* in the theory of sampling from finite populations without replacement.

If $n = 1$, the value of this factor α is 1, because there is no distinction between sampling with replacement and sampling without replacement when only one ball is being selected. If $n = T$, then (as previously mentioned) $\alpha = 0$ and $\text{Var}(X) = 0$. For values of n between 1 and T , the value of α will be between 0 and 1.

For each fixed sample size n , it can be seen that $\alpha \rightarrow 1$ as $T \rightarrow \infty$. This limit reflects the fact that when the population size T is very large compared to the sample size n , there is very little difference between sampling with replacement and sampling without replacement. Theorem 5.3.4 expresses this idea more formally. The proof relies on the following result which gets used several times in this text.

Theorem 5.3.3 Let a_n and c_n be sequences of real numbers such that a_n converges to 0, and $c_n a_n^2$ converges to 0. Then

$$\lim_{n \rightarrow \infty} (1 + a_n)^{c_n} e^{-a_n c_n} = 1.$$

In particular, if $a_n c_n$ converges to b , then $(1 + a_n)^{c_n}$ converges to e^b . ■

The proof of Theorem 5.3.3 is left to the reader in Exercise 11.

Theorem 5.3.4 Closeness of Binomial and Hypergeometric Distributions. Let $0 < p < 1$, and let n be a positive integer. Let Y have the binomial distribution with parameters n and p . For each positive integer T , let A_T and B_T be integers such that $\lim_{T \rightarrow \infty} A_T = \infty$, $\lim_{T \rightarrow \infty} B_T = \infty$, and $\lim_{T \rightarrow \infty} A_T/(A_T + B_T) = p$. Let X_T have the hypergeometric distribution with parameters A_T , B_T , and n . For each fixed n and each $x = 0, \dots, n$,

$$\lim_{T \rightarrow \infty} \frac{\Pr(Y = x)}{\Pr(X_T = x)} = 1. \quad (5.3.9)$$

Proof Once A_T and B_T are both larger than n , the formula in (5.3.1) is $\Pr(X_T = x)$ for all $x = 0, \dots, n$. So, for large T , we have

$$\Pr(X_T = x) = \binom{n}{x} \frac{A_T! B_T! (A_T + B_T - n)!}{(A_T - x)! (B_T - n + x)! (A_T + B_T)!}.$$

Apply Stirling's formula (Theorem 1.7.5) to each of the six factorials in the second factor above. A little manipulation gives that

$$\lim_{T \rightarrow \infty} \frac{\binom{n}{x} A_T^{A_T+1/2} B_T^{B_T+1/2} (A_T + B_T - n)^{A_T+B_T-n+1/2}}{\Pr(X_T = x) (A_T - x)^{A_T-x+1/2} (B_T - n + x)^{B_T-n+x+1/2} (A_T + B_T)^{A_T+B_T+1/2}} \quad (5.3.10)$$

equals 1. Each of the following limits follows from Theorem 5.3.3:

$$\begin{aligned} \lim_{T \rightarrow \infty} \left(\frac{A_T}{A_T - x} \right)^{A_T-x+1/2} &= e^x \\ \lim_{T \rightarrow \infty} \left(\frac{B_T}{B_T - n + x} \right)^{B_T-n+x+1/2} &= e^{n-x} \\ \lim_{T \rightarrow \infty} \left(\frac{A_T + B_T - n}{A_T + B_T} \right)^{A_T+B_T-n+1/2} &= e^{-n}. \end{aligned}$$

Inserting these limits in (5.3.10) yields

$$\lim_{T \rightarrow \infty} \frac{\binom{n}{x} A_T^x B_T^{n-x}}{\Pr(X_T = x)(A_T + B_T)^n} = 1. \quad (5.3.11)$$

Since $A_T/(A_T + B_T)$ converges to p , we have

$$\lim_{T \rightarrow \infty} \frac{A_T^x B_T^{n-x}}{(A_T + B_T)^n} = p^x (1 - p)^{n-x}. \quad (5.3.12)$$

Together, (5.3.11) and (5.3.12) imply that

$$\lim_{T \rightarrow \infty} \frac{\binom{n}{x} p^x (1 - p)^{n-x}}{\Pr(X_T = x)} = 1.$$

The numerator of this last expression is $\Pr(Y = x)$; hence, (5.3.9) holds. ■

In words, Theorem 5.3.4 says that if the sample size n represents a negligible fraction of the total population $A + B$, then the hypergeometric distribution with parameters A , B , and n will be very nearly the same as the binomial distribution with parameters n and $p = A/(A + B)$.

Example 5.3.3

Population of Unknown Composition. The hypergeometric distribution can arise as a conditional distribution when sampling is done without replacement from a finite population of unknown composition. The simplest example would be to modify Example 5.3.1 so that we still know the value of $T = A + B$ but no longer know A and B . That is, we know how many balls are in the box, but we don't know how many are red or blue. This makes $P = A/T$, the proportion of red balls, unknown. Let $h(p)$ be the p.f. of P . Here P is a random variable whose possible values are $0, 1/T, \dots, (T-1)/T, 1$. Conditional on $P = p$, we can behave as if we know that $A = pT$ and $B = (1 - p)T$, and then the conditional distribution of X (the number of red balls in a sample of size n) is the hypergeometric distribution with parameters pT , $(1 - p)T$, and n .

Suppose now that T is so large that the difference is essentially negligible between this hypergeometric distribution and the binomial distribution with parameters n and p . In this case, it is no longer necessary that we assume that T is known. This is the situation that we had in mind (in Examples 3.4.10 and 3.6.7, as well as their many variations and other examples) when we referred to P as the proportion of successes among all patients who might receive a treatment or the proportion of defectives among all parts produced by a machine. We think of T as essentially infinite so that conditional on the proportion A/T , which we call P , the individual draws become independent Bernoulli trials. If either A or T (or both) is unknown, it makes sense that $P = A/T$ will be unknown. In the augmented experiment described on page 61, in which P can be computed from the experimental outcome, we have that P is a random variable. ◀

Note: Essentially Infinite Populations. The case in which T is essentially infinite in Example 5.3.3 is the motivation for using the binomial distributions as models for numbers of successes in samples from very large finite populations. Look at Example 5.2.6, for instance. The number of Mexican Americans available to be sampled for grand jury duty is finite, but it is huge relative to the number (220) of grand jurors selected during the 2.5-year period. Technically, it is impossible that the individual grand jurors are selected independently, but the difference is too small for even the best defense attorney to make anything out of it. In the future, we will often model Bernoulli random variables as independent when we imagine selecting them

at random without replacement from a huge finite population. We shall be relying on Theorem 5.3.4 in these cases without explicitly saying so.

Extending the Definition of Binomial Coefficients

There is an extension of the definition of a binomial coefficient given in Sec. 1.8 that allows a simplification of the expression for the p.f. of the hypergeometric distribution. For all positive integers r and m , where $r \leq m$, the binomial coefficient $\binom{m}{r}$ was defined to be

$$\binom{m}{r} = \frac{m!}{r!(m-r)!}. \quad (5.3.13)$$

It can be seen that the value of $\binom{m}{r}$ specified by Eq. (5.3.13) can also be written in the form

$$\binom{m}{r} = \frac{m(m-1) \cdots (m-r+1)}{r!}. \quad (5.3.14)$$

For every real number m that is not necessarily a positive integer and every positive integer r , the value of the right side of Eq. (5.3.14) is a well-defined number. Therefore, for every real number m and every positive integer r , we can extend the definition of the binomial coefficient $\binom{m}{r}$ by defining its value as that given by Eq. (5.3.14).

The value of the binomial coefficient $\binom{m}{r}$ can be obtained from this definition for all positive integers r and m . If $r \leq m$, the value of $\binom{m}{r}$ is given by Eq. (5.3.13). If $r > m$, one of the factors in the numerator of (5.3.14) will be 0 and $\binom{m}{r} = 0$. Finally, for every real number m , we shall define the value of $\binom{m}{0}$ to be $\binom{m}{0} = 1$.

When this extended definition of a binomial coefficient is used, it can be seen that the value of $\binom{A}{x} \binom{B}{n-x}$ is 0 for every integer x such that either $x > A$ or $n - x > B$. Therefore, we can write the p.f. of the hypergeometric distribution with parameters A , B , and n as follows:

$$f(x|A, B, n) = \begin{cases} \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} & \text{for } x = 0, 1, \dots, n, \\ 0 & \text{otherwise.} \end{cases} \quad (5.3.15)$$

It then follows from Eq. (5.3.14) that $f(x|A, B, n) > 0$ if and only if x is an integer in the interval (5.3.2).

Summary

We introduced the family of hypergeometric distributions. Suppose that n units are drawn at random without replacement from a finite population consisting of T units of which A are successes and $B = T - A$ are failures. Let X stand for the number of successes in the sample. Then the distribution of X is the hypergeometric distribution with parameters A , B , and n . We saw that the distinction between sampling from a finite population with and without replacement is negligible when the size of the population is huge relative to the size of the sample. We also generalized the binomial coefficient notation so that $\binom{m}{r}$ is defined for all real numbers m and all positive integers r .

Exercises

1. In Example 5.3.2, compute the probability that all 10 success patients appear in the subsample of size 11 from the Placebo group.
2. Suppose that a box contains five red balls and ten blue balls. If seven balls are selected at random without replacement, what is the probability that at least three red balls will be obtained?
3. Suppose that seven balls are selected at random without replacement from a box containing five red balls and ten blue balls. If \bar{X} denotes the proportion of red balls in the sample, what are the mean and the variance of \bar{X} ?
4. If a random variable X has the hypergeometric distribution with parameters $A = 8$, $B = 20$, and n , for what value of n will $\text{Var}(X)$ be a maximum?
5. Suppose that n students are selected at random without replacement from a class containing T students, of whom A are boys and $T - A$ are girls. Let X denote the number of boys that are obtained. For what sample size n will $\text{Var}(X)$ be a maximum?
6. Suppose that X_1 and X_2 are independent random variables, that X_1 has the binomial distribution with parameters n_1 and p , and that X_2 has the binomial distribution with parameters n_2 and p , where p is the same for both X_1 and X_2 . For each fixed value of k ($k = 1, 2, \dots, n_1 + n_2$), prove that the conditional distribution of X_1 given that

$X_1 + X_2 = k$ is hypergeometric with parameters n_1 , n_2 , and k .

7. Suppose that in a large lot containing T manufactured items, 30 percent of the items are defective and 70 percent are nondefective. Also, suppose that ten items are selected at random without replacement from the lot. Determine (a) an exact expression for the probability that not more than one defective item will be obtained and (b) an approximate expression for this probability based on the binomial distribution.

8. Consider a group of T persons, and let a_1, \dots, a_T denote the heights of these T persons. Suppose that n persons are selected from this group at random without replacement, and let X denote the sum of the heights of these n persons. Determine the mean and variance of X .

9. Find the value of $\binom{3/2}{4}$.

10. Show that for all positive integers n and k ,

$$\binom{-n}{k} = (-1)^k \binom{n+k-1}{k}.$$

11. Prove Theorem 5.3.3. *Hint:* Prove that

$$\lim_{n \rightarrow \infty} c_n \log(1 + a_n) - a_n c_n = 0$$

by applying Taylor's theorem with remainder (see Exercise 13 in Sec. 4.2) to the function $f(x) = \log(1 + x)$ around $x = 0$.

5.4 The Poisson Distributions

Many experiments consist of observing the occurrence times of random arrivals. Examples include arrivals of customers for service, arrivals of calls at a switchboard, occurrences of floods and other natural and man-made disasters, and so forth. The family of Poisson distributions is used to model the number of such arrivals that occur in a fixed time period. Poisson distributions are also useful approximations to binomial distributions with very small success probabilities.

Definition and Properties of the Poisson Distributions

Example 5.4.1

Customer Arrivals. A store owner believes that customers arrive at his store at a rate of 4.5 customers per hour on average. He wants to find the distribution of the actual number X of customers who will arrive during a particular one-hour period later in the day. He models customer arrivals in different time periods as independent of each other. As a first approximation, he divides the one-hour period into 3600 seconds and thinks of the arrival rate as being $4.5/3600 = 0.00125$ per second. He then says that during each second either 0 or 1 customers will arrive, and the probability of an arrival during any single second is 0.00125. He then tries to use the binomial distribution with

parameters $n = 3600$ and $p = 0.00125$ for the distribution of the number of customers who arrive during the one-hour period later in the day.

He starts calculating f , the p.f. of this binomial distribution, and quickly discovers how cumbersome the calculations are. However, he realizes that the successive values of $f(x)$ are closely related to each other because $f(x)$ changes in a systematic way as x increases. So he computes

$$\frac{f(x+1)}{f(x)} = \frac{\binom{n}{x+1} p^{x+1} (1-p)^{n-x-1}}{\binom{n}{x} p^x (1-p)^{n-x}} = \frac{(n-x)p}{(x+1)(1-p)} \approx \frac{np}{x+1},$$

where the reasoning for the approximation at the end is as follows: For the first 30 or so values of x , $n-x$ is essentially the same as n and dividing by $1-p$ has almost no effect because p is so small. For example, for $x = 30$, the actual value is 0.1441, while the approximation is 0.1452. This approximation suggests defining $\lambda = np$ and approximating $f(x+1) \approx f(x)\lambda/(x+1)$ for all the values of x that matter. That is,

$$\begin{aligned} f(1) &= f(0)\lambda, \\ f(2) &= f(1)\frac{\lambda}{2} = f(0)\frac{\lambda^2}{2}, \\ f(3) &= f(2)\frac{\lambda}{3} = f(0)\frac{\lambda^3}{6}, \\ &\vdots \end{aligned}$$

Continuing the pattern for all x yields $f(x) = f(0)\lambda^x/x!$ for all x . To obtain a p.f. for X , he would need to make sure that $\sum_{x=0}^{\infty} f(x) = 1$. This is easily achieved by setting

$$f(0) = \frac{1}{\sum_{x=0}^{\infty} \lambda^x/x!} = e^{-\lambda},$$

where the last equality follows from the following well-known calculus result:

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}, \quad (5.4.1)$$

for all $\lambda > 0$. Hence, $f(x) = e^{-\lambda}\lambda^x/x!$ for $x = 0, 1, \dots$ and $f(x) = 0$ otherwise is a p.f. ◀

The approximation formula for the p.f. of a binomial distribution at the end of Example 5.4.1 is actually a useful p.f. that can model many phenomena of types similar to the arrivals of customers.

Definition 5.4.1 *Poisson Distribution.* Let $\lambda > 0$. A random variable X has the *Poisson distribution with mean λ* if the p.f. of X is as follows:

$$f(x|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases} \quad (5.4.2)$$

At the end of Example 5.4.1, we proved that the function in Eq. (5.4.2) is indeed a p.f. In order to justify the phrase “with mean λ ” in the definition of the distribution, we need to prove that the mean is indeed λ .

Theorem 5.4.1 *Mean.* The mean of the distribution with p.f. equal to (5.4.2) is λ .

Proof If X has the distribution with p.f. $f(x|\lambda)$, then $E(X)$ is given by the following infinite series:

$$E(X) = \sum_{x=0}^{\infty} xf(x|\lambda).$$

Since the term corresponding to $x = 0$ in this series is 0, we can omit this term and can begin the summation with the term for $x = 1$. Therefore,

$$E(X) = \sum_{x=1}^{\infty} xf(x|\lambda) = \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!}.$$

If we now let $y = x - 1$ in this summation, we obtain

$$E(X) = \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda} \lambda^y}{y!}.$$

The sum of the series in this equation is the sum of $f(y|\lambda)$, which equals 1. Hence, $E(X) = \lambda$. ■

Example
5.4.2

Customer Arrivals. In Example 5.4.1, the store owner was approximating the binomial distribution with parameters 3600 and 0.00125 with a distribution that we now know as the Poisson distribution with mean $\lambda = 3600 \times 0.00125 = 4.5$. For $x = 0, \dots, 9$, Table 5.1 has the binomial and corresponding Poisson probabilities.

The division of the one-hour period into 3600 seconds was somewhat arbitrary. The owner could have divided the hour into 7200 half-seconds or 14400 quarter-seconds, etc. Regardless of how finely the time is divided, the product of the number of time intervals and the rate in customers per time interval will always be 4.5 because they are all based on a rate of 4.5 customers per hour. Perhaps the store owner would do better simply modeling the number X of arrivals as a Poisson random variable with mean 4.5, rather than choosing an arbitrarily sized time interval to accommodate a tedious binomial calculation. The disadvantage to the Poisson model for X is that there is positive probability that a Poisson random variable will be arbitrarily large, whereas a binomial random variable with parameters n and p can never exceed n . However, the probability is essentially 0 that a Poisson random variable with mean 4.5 will exceed 19. ◀

Table 5.1 Binomial and Poisson probabilities in Example 5.4.2

	x				
	0	1	2	3	4
Binomial	0.01108	0.04991	0.11241	0.16874	0.18991
Poisson	0.01111	0.04999	0.11248	0.16872	0.18981
	x				
	5	6	7	8	9
Binomial	0.17094	0.12819	0.08237	0.04630	0.02313
Poisson	0.17083	0.12812	0.08237	0.04633	0.02317

Theorem 5.4.2 Variance. The variance of the Poisson distribution with mean λ is also λ .

Proof The variance can be found by a technique similar to the one used in the proof of Theorem 5.4.1 to find the mean. We begin by considering the following expectation:

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)f(x|\lambda) = \sum_{x=2}^{\infty} x(x-1)f(x|\lambda) \\ &= \sum_{x=2}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!}. \end{aligned}$$

If we let $y = x - 2$, we obtain

$$E[X(X-1)] = \lambda^2 \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda^2. \quad (5.4.3)$$

Since $E[X(X-1)] = E(X^2) - E(X) = E(X^2) - \lambda$, it follows from Eq. (5.4.3) that $E(X^2) = \lambda^2 + \lambda$. Therefore,

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \lambda. \quad (5.4.4)$$

Hence, the variance is also equal to λ . ■

Theorem 5.4.3 Moment Generating Function. The m.g.f. of the Poisson distribution with mean λ is

$$\psi(t) = e^{\lambda(e^t-1)}, \quad (5.4.5)$$

for all real t .

Proof For every value of t ($-\infty < t < \infty$),

$$\psi(t) = E(e^{tX}) = \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}.$$

It follows from Eq. (5.4.1) that, for $-\infty < t < \infty$,

$$\psi(t) = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}. \quad \blacksquare$$

The mean and the variance, as well as all other moments, can be determined from the m.g.f. given in Eq. (5.4.5). We shall not derive the values of any other moments here, but we shall use the m.g.f. to derive the following property of Poisson distributions.

Theorem 5.4.4 If the random variables X_1, \dots, X_k are independent and if X_i has the Poisson distribution with mean λ_i ($i = 1, \dots, k$), then the sum $X_1 + \dots + X_k$ has the Poisson distribution with mean $\lambda_1 + \dots + \lambda_k$.

Proof Let $\psi_i(t)$ denote the m.g.f. of X_i for $i = 1, \dots, k$, and let $\psi(t)$ denote the m.g.f. of the sum $X_1 + \dots + X_k$. Since X_1, \dots, X_k are independent, it follows that, for $-\infty < t < \infty$,

$$\psi(t) = \prod_{i=1}^k \psi_i(t) = \prod_{i=1}^k e^{\lambda_i(e^t-1)} = e^{(\lambda_1 + \dots + \lambda_k)(e^t-1)}.$$

It can be seen from Eq. (5.4.5) that this m.g.f. $\psi(t)$ is the m.g.f. of the Poisson distribution with mean $\lambda_1 + \cdots + \lambda_k$. Hence, the distribution of $X_1 + \cdots + X_k$ must be as stated in the theorem. ■

A table of probabilities for Poisson distributions with various values of the mean λ is given at the end of this book.

**Example
5.4.3**

Customer Arrivals. Suppose that the store owner in Examples 5.4.1 and 5.4.2 is interested not only in the number of customers that arrive in the one-hour period, but also in how many customers arrive in the next hour after that period. Let Y be the number of customers that arrive in the second hour. By the reasoning at the end of Example 5.4.2, the owner might model Y as a Poisson random variable with mean 4.5. He would also say that X and Y are independent because he has been assuming that arrivals in disjoint time intervals are independent. According to Theorem 5.4.4, $X + Y$ would have the Poisson distribution with mean $4.5 + 4.5 = 9$. What is the probability that at least 12 customers will arrive in the entire two-hour period? We can use the table of Poisson probabilities in the back of this book by looking in the $\lambda = 9$ column. Either add up the numbers corresponding to $k = 0, \dots, 11$ and subtract the total from 1, or add up those from $k = 12$ to the end. Either way, the result is $\Pr(X \geq 12) = 0.1970$. ◀

The Poisson Approximation to Binomial Distributions

In Examples 5.4.1 and 5.4.2, we illustrated how close the Poisson distribution with mean 4.5 is to the binomial distribution with parameters 3600 and 0.00125. We shall now demonstrate a general version of that result, namely, that when the value of n is large and the value of p is close to 0, the binomial distribution with parameters n and p can be approximated by the Poisson distribution with mean np .

**Theorem
5.4.5**

Closeness of Binomial and Poisson Distributions. For each integer n and each $0 < p < 1$, let $f(x|n, p)$ denote the p.f. of the binomial distribution with parameters n and p . Let $f(x|\lambda)$ denote the p.f. of the Poisson distribution with mean λ . Let $\{p_n\}_{n=1}^\infty$ be a sequence of numbers between 0 and 1 such that $\lim_{n \rightarrow \infty} np_n = \lambda$. Then

$$\lim_{n \rightarrow \infty} f(x|n, p_n) = f(x|\lambda),$$

for all $x = 0, 1, \dots$

Proof We begin by writing

$$f(x|n, p_n) = \frac{n(n-1) \cdots (n-x+1)}{x!} p_n^x (1-p_n)^{n-x}.$$

Next, let $\lambda_n = np_n$ so that $\lim_{n \rightarrow \infty} \lambda_n = \lambda$. Then $f(x|n, p_n)$ can be rewritten in the following form:

$$f(x|n, p_n) = \frac{\lambda_n^x}{x!} \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n} \left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-x}. \quad (5.4.6)$$

For each $x \geq 0$,

$$\lim_{n \rightarrow \infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n} \left(1 - \frac{\lambda_n}{n}\right)^{-x} = 1.$$

Furthermore, it follows from Theorem 5.3.3 that

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda_n}{n}\right)^n = e^{-\lambda}. \quad (5.4.7)$$

It now follows from Eq. (5.4.6) that for every $x \geq 0$,

$$\lim_{n \rightarrow \infty} f(x|n, p_n) = \frac{e^{-\lambda} \lambda^x}{x!} = f(x|\lambda). \quad \blacksquare$$

**Example
5.4.4**

Approximating a Probability. Suppose that in a large population the proportion of people who have a certain disease is 0.01. We shall determine the probability that in a random group of 200 people at least four people will have the disease.

In this example, we can assume that the exact distribution of the number of people having the disease among the 200 people in the random group is the binomial distribution with parameters $n = 200$ and $p = 0.01$. Therefore, this distribution can be approximated by the Poisson distribution for which the mean is $\lambda = np = 2$. If X denotes a random variable having this Poisson distribution, then it can be found from the table of the Poisson distribution at the end of this book that $\Pr(X \geq 4) = 0.1428$. Hence, the probability that at least four people will have the disease is approximately 0.1428. The actual value is 0.1420. ◀

Theorem 5.4.5 says that if n is large and p is small so that np is close to λ , then the binomial distribution with parameters n and p is close to the Poisson distribution with mean λ . Recall Theorem 5.3.4, which says that if A and B are large compared to n and if $A/(A + B)$ is close to p , then the hypergeometric distribution with parameters A , B , and n is close to the binomial distribution with parameters n and p . These two results can be combined into the following theorem, whose proof is left to Exercise 17.

**Theorem
5.4.6**

Closeness of Hypergeometric and Poisson Distributions. Let $\lambda > 0$. Let Y have the Poisson distribution with mean λ . For each positive integer T , let A_T , B_T , and n_T be integers such that $\lim_{T \rightarrow \infty} A_T = \infty$, $\lim_{T \rightarrow \infty} B_T = \infty$, $\lim_{T \rightarrow \infty} n_T = \infty$, and $\lim_{T \rightarrow \infty} n_T A_T / (A_T + B_T) = \lambda$. Let X_T have the hypergeometric distribution with parameters A_T , B_T , and n_T . For each fixed $x = 0, 1, \dots$,

$$\lim_{T \rightarrow \infty} \frac{\Pr(Y = x)}{\Pr(X_T = x)} = 1. \quad \blacksquare$$

Poisson Processes

**Example
5.4.5**

Customer Arrivals. In Example 5.4.3, the store owner believes that the number of customers that arrive in each one-hour period has the Poisson distribution with mean 4.5. What if the owner is interested in a half-hour period or a 4-hour and 15-minute period? Is it safe to assume that the number of customers that arrive in a half-hour period has the Poisson distribution with mean 2.25? ◀

In order to be sure that all of the distributions for the various numbers of arrivals in Example 5.4.5 are consistent with each other, the store owner needs to think about the overall process of customer arrivals, not just a few isolated time periods. The following definition gives a model for the overall process of arrivals that will allow the store owner to construct distributions for all the counts of customer arrivals that interest him as well as other useful things.

Definition
5.4.2

Poisson Process. A *Poisson process* with rate λ per unit time is a process that satisfies the following two properties:

- i. The number of arrivals in every fixed interval of time of length t has the Poisson distribution with mean λt .
- ii. The numbers of arrivals in every collection of disjoint time intervals are independent.

The answer to the question at the end of Example 5.4.5 will be “yes” if the store owner makes the assumption that customers arrive according to a Poisson process with rate 4.5 per hour. Here is another example.

Example
5.4.6

Radioactive Particles. Suppose that radioactive particles strike a certain target in accordance with a Poisson process at an average rate of three particles per minute. We shall determine the probability that 10 or more particles will strike the target in a particular two-minute period.

In a Poisson process, the number of particles striking the target in any particular one-minute period has the Poisson distribution with mean λ . Since the mean number of strikes in any one-minute period is 3, it follows that $\lambda = 3$ in this example. Therefore, the number of strikes X in any two-minute period will have the Poisson distribution with mean 6. It can be found from the table of the Poisson distribution at the end of this book that $\Pr(X \geq 10) = 0.0838$. ◀

Note: Generality of Poisson Processes. Although we have introduced Poisson processes in terms of counts of arrivals during time intervals, Poisson processes are actually more general. For example, a Poisson process can be used to model occurrences in space as well as time. A Poisson process could be used to model telephone calls arriving at a switchboard, atomic particles emitted from a radioactive source, diseased trees in a forest, or defects on the surface of a manufactured product. The reason for the popularity of the Poisson process model is twofold. First, the model is computationally convenient. Second, there is a mathematical justification for the model if one makes three plausible assumptions about how the phenomena occur. We shall present the three assumptions in some detail after another example.

Example
5.4.7

Cryptosporidium in Drinking Water. *Cryptosporidium* is a genus of protozoa that occurs as small oocysts and can cause painful sickness and even death when ingested. Occasionally, oocysts are detected in public drinking water supplies. A concentration as low as one oocyst per five liters can be enough to trigger a boil-water advisory. In April 1993, many thousands of people became ill during a cryptosporidiosis outbreak in Milwaukee, Wisconsin. Different water systems have different systems for monitoring protozoa occurrence in drinking water. One problem with monitoring systems is that detection technology is not always very sensitive. One popular technique is to push a large amount of water through a very fine filter and then treat the material captured on the filter in a way that identifies *Cryptosporidium* oocysts. The number of oocysts is then counted and recorded. Even if there is an oocyst on the filter, the probability can be as low as 0.1 that it will get counted.

Suppose that, in a particular water supply, oocysts occur according to a Poisson process with rate λ oocysts per liter. Suppose that the filtering system is capable of capturing all oocysts in a sample, but that the counting system has probability p of actually observing each oocyst that is on the filter. Assume that the counting system observes or misses each oocyst on the filter independently. What is the distribution of the number of counted oocysts from t liters of filtered water?

Let Y be the number of oocysts in the t liters (all of which make it onto the filter). Then Y has the Poisson distribution with mean λt . Let $X_i = 1$ if the i th oocyst on the filter gets counted, and $X_i = 0$ if not. Let X be the counted number of oocysts so that $X = X_1 + \cdots + X_y$ if $Y = y$. Conditional on $Y = y$, we have assumed that the X_i are independent Bernoulli random variables with parameter p , so X has the binomial distribution with parameters y and p conditional on $Y = y$. We want the marginal distribution of X . This can be found using the law of total probability for random variables (3.6.11). For $x = 0, 1, \dots$,

$$\begin{aligned}
 f_1(x) &= \sum_{y=0}^{\infty} g_1(x|y) f_2(y) \\
 &= \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} e^{-\lambda t} \frac{(\lambda t)^y}{y!} \\
 &= e^{-\lambda t} \frac{(p\lambda t)^x}{x!} \sum_{y=x}^{\infty} \frac{[\lambda t(1-p)]^{y-x}}{(y-x)!} \\
 &= e^{-\lambda t} \frac{(p\lambda t)^x}{x!} \sum_{u=0}^{\infty} \frac{[\lambda t(1-p)]^u}{u!} \\
 &= e^{-\lambda t} \frac{(p\lambda t)^x}{x!} e^{\lambda t(1-p)} = e^{-p\lambda t} \frac{(p\lambda t)^x}{x!}.
 \end{aligned}$$

This is easily recognized as the p.f. of the Poisson distribution with mean $p\lambda t$. The effect of losing a fraction $1-p$ of the oocyst count is merely to lower the rate of the Poisson process from λ per liter to $p\lambda$ per liter.

Suppose that $\lambda = 0.2$ and $p = 0.1$. How much water must we filter in order for there to be probability at least 0.9 that we will count at least one oocyst? The probability of counting at least one oocyst is 1 minus the probability of counting none, which is $e^{-p\lambda t} = e^{-0.02t}$. So, we need t large enough so that $1 - e^{-0.02t} \geq 0.9$, that is, $t \geq 115$. A typical procedure is to test 100 liters, which would have probability $1 - e^{-0.02 \times 100} = 0.86$ of detecting at least one oocyst. ◀



Assumptions Underlying the Poisson Process Model

In what follows, we shall refer to time intervals, but the assumptions can be used equally well for subregions of two- or three-dimensional regions or sublengths of a linear distance. Indeed, a Poisson process can be used to model occurrences in any region that can be subdivided into arbitrarily small pieces. There are three assumptions that lead to the Poisson process model.

The first assumption is that the numbers of occurrences in any collection of *disjoint* intervals of time must be mutually independent. For example, even though an unusually large number of telephone calls are received at a switchboard during a particular interval, the probability that at least one call will be received during a forthcoming interval remains unchanged. Similarly, even though no call has been received at the switchboard for an unusually long interval, the probability that a call will be received during the next short interval remains unchanged.

The second assumption is that the probability of an occurrence during each very short interval of time must be approximately proportional to the length of that interval. To express this condition more formally, we shall use the standard

mathematical notation in which $o(t)$ denotes any function of t having the property that

$$\lim_{t \rightarrow 0} \frac{o(t)}{t} = 0. \quad (5.4.8)$$

According to (5.4.8), $o(t)$ must be a function that approaches 0 as $t \rightarrow 0$, and, furthermore, this function must approach 0 at a rate faster than t itself. An example of such a function is $o(t) = t^\alpha$, where $\alpha > 1$. It can be verified that this function satisfies Eq. (5.4.8). The second assumption can now be expressed as follows: There exists a constant $\lambda > 0$ such that for every time interval of length t , the probability of at least one occurrence during that interval has the form $\lambda t + o(t)$. Thus, for every very small value of t , the probability of at least one occurrence during an interval of length t is equal to λt plus a quantity having a smaller order of magnitude.

One of the consequences of the second assumption is that the process being observed must be *stationary* over the entire period of observation; that is, the probability of an occurrence must be the same over the entire period. There can be neither busy intervals, during which we know in advance that occurrences are likely to be more frequent, nor quiet intervals, during which we know in advance that occurrences are likely to be less frequent. This condition is reflected in the fact that the same constant λ expresses the probability of an occurrence in every interval over the entire period of observation. The second assumption can be relaxed at the cost of more complicated mathematics, but we shall not do so here.

The third assumption is that, for each very short interval of time, the probability that there will be two or more occurrences in that interval must have a smaller order of magnitude than the probability that there will be just one occurrence. In symbols, the probability of two or more occurrences in a time interval of length t must be $o(t)$. Thus, the probability of two or more occurrences in a small interval must be negligible in comparison with the probability of one occurrence in that interval. Of course, it follows from the second assumption that the probability of one occurrence in that same interval will itself be negligible in comparison with the probability of no occurrences.

Under the preceding three assumptions, it can be shown that the process will satisfy the definition of a Poisson process with rate λ . See Exercise 16 in this section for one method of proof.



Summary

Poisson distributions are used to model data that arrive as counts. A Poisson process with rate λ is a model for random occurrences that have a constant expected rate λ per unit time (or per unit area). We must assume that occurrences in disjoint time intervals (or disjoint areas) are independent and that two or more occurrences cannot happen at the same time (or place). The number of occurrences in an interval of length (or area of size) t has the Poisson distribution with mean $t\lambda$. If n is large and p is small, then the binomial distribution with parameters n and p is approximately the same as the Poisson distribution with mean np .

Exercises

1. In Example 5.4.7, with $\lambda = 0.2$ and $p = 0.1$, compute the probability that we would detect at least two oocysts after filtering 100 liters of water.

2. Suppose that on a given weekend the number of accidents at a certain intersection has the Poisson distribution with mean 0.7. What is the probability that there will be at least three accidents at the intersection during the weekend?

3. Suppose that the number of defects on a bolt of cloth produced by a certain process has the Poisson distribution with mean 0.4. If a random sample of five bolts of cloth is inspected, what is the probability that the total number of defects on the five bolts will be at least 6?

4. Suppose that in a certain book there are on the average λ misprints per page and that misprints occurred according to a Poisson process. What is the probability that a particular page will contain no misprints?

5. Suppose that a book with n pages contains on the average λ misprints per page. What is the probability that there will be at least m pages which contain more than k misprints?

6. Suppose that a certain type of magnetic tape contains on the average three defects per 1000 feet. What is the probability that a roll of tape 1200 feet long contains no defects?

7. Suppose that on the average a certain store serves 15 customers per hour. What is the probability that the store will serve more than 20 customers in a particular two-hour period?

8. Suppose that X_1 and X_2 are independent random variables and that X_i has the Poisson distribution with mean λ_i ($i = 1, 2$). For each fixed value of k ($k = 1, 2, \dots$), determine the conditional distribution of X_1 given that $X_1 + X_2 = k$.

9. Suppose that the total number of items produced by a certain machine has the Poisson distribution with mean λ , all items are produced independently of one another, and the probability that any given item produced by the machine will be defective is p . Determine the marginal distribution of the number of defective items produced by the machine.

10. For the problem described in Exercise 9, let X denote the number of defective items produced by the machine, and let Y denote the number of nondefective items produced by the machine. Show that X and Y are independent random variables.

11. The mode of a discrete distribution was defined in Exercise 12 of Sec. 5.2. Determine the mode or modes of the Poisson distribution with mean λ .

12. Suppose that the proportion of colorblind people in a certain population is 0.005. What is the probability that there will not be more than one colorblind person in a randomly chosen group of 600 people?

13. The probability of triplets in human births is approximately 0.001. What is the probability that there will be exactly one set of triplets among 700 births in a large hospital?

14. An airline sells 200 tickets for a certain flight on an airplane that has only 198 seats because, on the average, 1 percent of purchasers of airline tickets do not appear for the departure of their flight. Determine the probability that everyone who appears for the departure of this flight will have a seat.

15. Suppose that internet users access a particular Web site according to a Poisson process with rate λ per hour, but λ is unknown. The Web site maintainer believes that λ has a continuous distribution with p.d.f.

$$f(\lambda) = \begin{cases} 2e^{-2\lambda} & \text{for } \lambda > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let X be the number of users who access the Web site during a one-hour period. If $X = 1$ is observed, find the conditional p.d.f. of λ given $X = 1$.

16. In this exercise, we shall prove that the three assumptions underlying the Poisson process model do indeed imply that occurrences happen according to a Poisson process. What we need to show is that, for each t , the number of occurrences during a time interval of length t has the Poisson distribution with mean λt . Let X stand for the number of occurrences during a particular time interval of length t . Feel free to use the following extension of Eq. (5.4.7): For all real a ,

$$\lim_{u \rightarrow 0} (1 + au + o(u))^{1/u} = e^a, \quad (5.4.9)$$

- a.** For each positive integer n , divide the time interval into n disjoint subintervals of length t/n each. For $i = 1, \dots, n$, let $Y_i = 1$ if exactly one arrival occurs in the i th subinterval, and let A_i be the event that two or more occurrences occur during the i th subinterval. Let $W_n = \sum_{i=1}^n Y_i$. For each nonnegative integer k , show that we can write $\Pr(X = k) = \Pr(W_n = k) + \Pr(B)$, where B is a subset of $\bigcup_{i=1}^n A_i$.
- b.** Show that $\lim_{n \rightarrow \infty} \Pr(\bigcup_{i=1}^n A_i) = 0$. *Hint:* Show that $\Pr(\bigcap_{i=1}^n A_i^c) = (1 + o(u))^{1/u}$ where $u = 1/n$.
- c.** Show that $\lim_{n \rightarrow \infty} \Pr(W_n = k) = e^{-\lambda} (\lambda t)^k / k!$. *Hint:* $\lim_{n \rightarrow \infty} n! / [n^k (n - k)!] = 1$.
- d.** Show that X has the Poisson distribution with mean λt .

17. Prove Theorem 5.4.6. One approach is to adapt the proof of Theorem 5.3.4 by replacing n by n_T in that proof. The steps of the proof that are significantly different are the following. (i) You will need to show that $B_T - n_T$ goes to ∞ . (ii) The three limits that depend on Theorem 5.3.3 need to be rewritten as ratios converging to 1. For example, the second one is rewritten as

$$\lim_{T \rightarrow \infty} \left(\frac{B_T}{B_T - n_T + x} \right)^{B_T - n_T + x + 1/2} e^{-n_T + x} = 1.$$

You'll need a couple more such limits as well. (iii) Instead of (5.3.12), prove that

$$\lim_{T \rightarrow \infty} \frac{n_T^x A_T^x B_T^{n_T - x}}{(A_T + B_T)^{n_T}} = \lambda^x e^{-\lambda}.$$

18. Let A_T , B_T , and n_T be sequences, all three of which go to ∞ as $T \rightarrow \infty$. Prove that $\lim_{T \rightarrow \infty} n_T A_T / (A_T + B_T) = \lambda$ if and only if $\lim_{T \rightarrow \infty} n_T A_T / B_T = \lambda$.

5.5 The Negative Binomial Distributions

Earlier we learned that, in n Bernoulli trials with probability of success p , the number of successes has the binomial distribution with parameters n and p . Instead of counting successes in a fixed number of trials, it is often necessary to observe the trials until we see a fixed number of successes. For example, while monitoring a piece of equipment to see when it needs maintenance, we might let it run until it produces a fixed number of errors and then repair it. The number of failures until a fixed number of successes has a distribution in the family of negative binomial distributions.

Definition and Interpretation

Example 5.5.1

Defective Parts. Suppose that a machine produces parts that can be either good or defective. Let $X_i = 1$ if the i th part is defective and $X_i = 0$ otherwise. Assume that the parts are good or defective independently of each other with $\Pr(X_i = 1) = p$ for all i . An inspector observes the parts produced by this machine until she sees four defectives. Let X be the number of good parts observed by the time that the fourth defective is observed. What is the distribution of X ? ◀

The problem described in Example 5.5.1 is typical of a general situation in which a sequence of Bernoulli trials can be observed. Suppose that an infinite sequence of Bernoulli trials is available. Call the two possible outcomes success and failure, with p being the probability of success. In this section, we shall study the distribution of the total number of failures that will occur before exactly r successes have been obtained, where r is a fixed positive integer.

Theorem 5.5.1

Sampling until a Fixed Number of Successes. Suppose that an infinite sequence of Bernoulli trials with probability of success p are available. The number X of failures that occur before the r th success has the following p.d.f.:

$$f(x|r, p) = \begin{cases} \binom{r+x-1}{x} p^r (1-p)^x & \text{for } x = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases} \quad (5.5.1)$$

Proof For $n = r, r+1, \dots$, we shall let A_n denote the event that the total number of trials required to obtain exactly r successes is n . As explained in Example 2.2.8, the event A_n will occur if and only if exactly $r-1$ successes occur among the first $n-1$

trials and the r th success is obtained on the n th trial. Since all trials are independent, it follows that

$$\Pr(A_n) = \binom{n-1}{r-1} p^{r-1} (1-p)^{(n-1)-(r-1)} \cdot p = \binom{n-1}{r-1} p^r (1-p)^{n-r}. \quad (5.5.2)$$

For each value of x ($x = 0, 1, 2, \dots$), the event that exactly x failures are obtained before the r th success is obtained is the same as the event that the total number of trials required to obtain r successes is $r + x$. In other words, if X denotes the number of failures that will occur before the r th success is obtained, then $\Pr(X = x) = \Pr(A_{r+x})$. Eq. (5.5.1) now follows from Eq. (5.5.2). ■

Definition 5.5.1 *Negative Binomial Distribution.* A random variable X has the *negative binomial distribution with parameters r and p* ($r = 1, 2, \dots$ and $0 < p < 1$) if X has a discrete distribution for which the p.f. $f(x|r, p)$ is as specified by Eq. (5.5.1).

Example 5.5.2 *Defective Parts.* Example 5.5.1 is worded so that defective parts are successes and good parts are failures. The distribution of the number X of good parts observed by the time of the fourth defective is the negative binomial distribution with parameters 4 and p . ◀

The Geometric Distributions

The most common special case of a negative binomial random variable is one for which $r = 1$. This would be the number of failures until the first success.

Definition 5.5.2 *Geometric Distribution.* A random variable X has the *geometric distribution with parameter p* ($0 < p < 1$) if X has a discrete distribution for which the p.f. $f(x|1, p)$ is as follows:

$$f(x|1, p) = \begin{cases} p(1-p)^x & \text{for } x = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases} \quad (5.5.3)$$

Example 5.5.3 *Triples in the Lottery.* A common daily lottery game involves the drawing of three digits from 0 to 9 independently with replacement and independently from day to day. Lottery watchers often get excited when all three digits are the same, an event called *triples*. If p is the probability of obtaining triples, and if X is the number of days without triples before the first triple is observed, then X has the geometric distribution with parameter p . In this case, it is easy to see that $p = 0.01$, since there are 10 different triples among the 1000 equally likely daily numbers. ◀

The relationship between geometric and negative binomial distributions goes beyond the fact that the geometric distributions are special cases of negative binomial distributions.

Theorem 5.5.2 If X_1, \dots, X_r are i.i.d. random variables and if each X_i has the geometric distribution with parameter p , then the sum $X_1 + \dots + X_r$ has the negative binomial distribution with parameters r and p .

Proof Consider an infinite sequence of Bernoulli trials with success probability p . Let X_1 denote the number of failures that occur before the first success is obtained; then X_1 will have the geometric distribution with parameter p .

Now continue observing the Bernoulli trials after the first success. For $j = 2, 3, \dots$, let X_j denote the number of failures that occur after $j - 1$ successes have

been obtained but before the j th success is obtained. Since all the trials are independent and the probability of obtaining a success on each trial is p , it follows that each random variable X_j will have the geometric distribution with parameter p and that the random variables X_1, X_2, \dots will be independent. Furthermore, for $r = 1, 2, \dots$, the sum $X_1 + \dots + X_r$ will be equal to the total number of failures that occur before exactly r successes have been obtained. Therefore, this sum will have the negative binomial distribution with parameters r and p . ■

Properties of Negative Binomial and Geometric Distributions

Theorem 5.5.3 **Moment Generating Function.** If X has the negative binomial distribution with parameters r and p , then the m.g.f. of X is as follows:

$$\psi(t) = \left(\frac{p}{1 - (1-p)e^t} \right)^r \quad \text{for } t < \log \left(\frac{1}{1-p} \right). \quad (5.5.4)$$

The m.g.f. of the geometric distribution with parameter p is the special case of Eq. (5.5.4) with $r = 1$.

Proof Let X_1, \dots, X_r be a random sample of r geometric random variables each with parameter p . We shall find the m.g.f. of X_1 and then apply Theorems 4.4.4 and 5.5.2 to find the m.g.f. of the negative binomial distribution with parameters r and p .

The m.g.f. $\psi_1(t)$ of X_1 is

$$\psi_1(t) = E(e^{tX_1}) = p \sum_{x=0}^{\infty} [(1-p)e^t]^x. \quad (5.5.5)$$

The infinite series in Eq. (5.5.5) will have a finite sum for every value of t such that $0 < (1-p)e^t < 1$, that is, for $t < \log(1/[1-p])$. It is known from elementary calculus that for every number α ($0 < \alpha < 1$),

$$\sum_{x=0}^{\infty} \alpha^x = \frac{1}{1-\alpha}.$$

Therefore, for $t < \log(1/[1-p])$, the m.g.f. of the geometric distribution with parameter p is

$$\psi_1(t) = \frac{p}{1 - (1-p)e^t}. \quad (5.5.6)$$

Each of X_1, \dots, X_r has the same m.g.f., namely, ψ_1 . According to Theorem 4.4.4, the m.g.f. of $X = X_1 + \dots + X_r$ is $\psi(t) = [\psi_1(t)]^r$. Theorem 5.5.2 says that X has the negative binomial distribution with parameters r and p , and hence the m.g.f. of X is $[\psi_1(t)]^r$, which is the same as Eq. (5.5.4). ■

Theorem 5.5.4 **Mean and Variance.** If X has the negative binomial distribution with parameters r and p , the mean and the variance of X must be

$$E(X) = \frac{r(1-p)}{p} \quad \text{and} \quad \text{Var}(X) = \frac{r(1-p)}{p^2}. \quad (5.5.7)$$

The mean and variance of the geometric distribution with parameter p are the special case of Eq. (5.5.7) with $r = 1$.

Proof Let X_1 have the geometric distribution with parameter p . We will find the mean and variance by differentiating the m.g.f. Eq. (5.5.5):

$$E(X_1) = \psi'_1(0) = \frac{1-p}{p}, \quad (5.5.8)$$

$$\text{Var}(X_1) = \psi''_1(0) - [\psi'_1(0)]^2 = \frac{1-p}{p^2}. \quad (5.5.9)$$

If X has the negative binomial distribution with parameters r and p , represent it as the sum $X = X_1 + \cdots + X_r$ of r independent random variables, each having the same distribution as X_1 . Eq. (5.5.7) now follows from Eqs. (5.5.8) and (5.5.9). ■

**Example
5.5.4**

Triples in the Lottery. In Example 5.5.3, the number X of daily draws without a triple until we see a triple has the geometric distribution with parameter $p = 0.01$. The total number of days until we see the first triple is then $X + 1$. So, the expected number of days until we observe triples is $E(X) + 1 = 100$.

Now suppose that a lottery player has been waiting 120 days for triples to occur. Such a player might conclude from the preceding calculation that triples are “due.” The most straightforward way to address such a claim would be to start by calculating the conditional distribution of X given that $X \geq 120$. ◀

The next result says that the lottery player at the end of Example 5.5.4 couldn't be farther from correct. Regardless of how long he has waited for triples, the time remaining until triples occur has the same geometric distribution (and the same mean) as it had when he started waiting. The proof is simple and is left as Exercise 8.

**Theorem
5.5.5**

Memoryless Property of Geometric Distributions. Let X have the geometric distribution with parameter p , and let $k \geq 0$. Then for every integer $t \geq 0$,

$$\Pr(X = k + t | X \geq k) = \Pr(X = t). \quad \blacksquare$$

The intuition behind Theorem 5.5.5 is the following: Think of X as the number of failures until the first success in a sequence of Bernoulli trials. Let Y be the number of failures starting with the $k + 1$ st trial until the next success. Then Y has the same distribution as X and is independent of the first k trials. Hence, conditioning on anything that happened on the first k trials, such as no successes yet, doesn't affect the distribution of Y —it is still the same geometric distribution. A formal proof can be given in Exercise 8. In Exercise 13, you can prove that the geometric distributions are the only discrete distributions that have the memoryless property.

**Example
5.5.5**

Triples in the Lottery. In Example 5.5.4, after the first 120 non-triples, the process essentially starts over again and we still have to wait a geometric amount of time until the first triple.

At the beginning of the experiment, the expected number of failures (non-triples) that will occur before the first success (triples) is $(1-p)/p$, as given by Eq. (5.5.8). If it is known that failures were obtained on the first 120 trials, then the conditional expected total number of failures before the first success (given the 120 failures on the first 120 trials) is simply $120 + (1-p)/p$. ◀

Extension of Definition of Negative Binomial Distribution

By using the definition of binomial coefficients given in Eq. (5.3.14), the function $f(x|r, p)$ can be regarded as the p.f. of a discrete distribution for each number $r > 0$ (not necessarily an integer) and each number p in the interval $0 < p < 1$. In other words, it can be verified that for $r > 0$ and $0 < p < 1$,

$$\sum_{x=0}^{\infty} \binom{r+x-1}{x} p^r (1-p)^x = 1. \quad (5.5.10)$$

Summary

If we observe a sequence of independent Bernoulli trials with success probability p , the number of failures until the r th success has the negative binomial distribution with parameters r and p . The special case of $r = 1$ is the geometric distribution with parameter p . The sum of independent negative binomial random variables with the same second parameter p has a negative binomial distribution.

Exercises

1. Consider a daily lottery as described in Example 5.5.4.
 - a. Compute the probability that two particular days in a row will both have triples.
 - b. Suppose that we observe triples on a particular day. Compute the conditional probability that we observe triples again the next day.
2. Suppose that a sequence of independent tosses are made with a coin for which the probability of obtaining a head on each given toss is $1/30$.
 - a. What is the expected number of tails that will be obtained before five heads have been obtained?
 - b. What is the variance of the number of tails that will be obtained before five heads have been obtained?
3. Consider the sequence of coin tosses described in Exercise 2.
 - a. What is the expected number of tosses that will be required in order to obtain five heads?
 - b. What is the variance of the number of tosses that will be required in order to obtain five heads?
4. Suppose that two players A and B are trying to throw a basketball through a hoop. The probability that player A will succeed on any given throw is p , and he throws until he has succeeded r times. The probability that player B will succeed on any given throw is mp , where m is a given integer ($m = 2, 3, \dots$) such that $mp < 1$, and she throws until she has succeeded mr times.
 - a. For which player is the expected number of throws smaller?
 - b. For which player is the variance of the number of throws smaller?
5. Suppose that the random variables X_1, \dots, X_k are independent and that X_i has the negative binomial distribution with parameters r_i and p ($i = 1 \dots k$). Prove that the sum $X_1 + \dots + X_k$ has the negative binomial distribution with parameters $r = r_1 + \dots + r_k$ and p .
6. Suppose that X has the geometric distribution with parameter p . Determine the probability that the value of X will be one of the even integers $0, 2, 4, \dots$.
7. Suppose that X has the geometric distribution with parameter p . Show that for every nonnegative integer k , $\Pr(X \geq k) = (1-p)^k$.
8. Prove Theorem 5.5.5.
9. Suppose that an electronic system contains n components that function independently of each other, and suppose that these components are connected in series, as defined in Exercise 5 of Sec. 3.7. Suppose also that each component will function properly for a certain number of periods and then will fail. Finally, suppose that for $i = 1, \dots, n$, the number of periods for which component i will function properly is a discrete random variable having

a geometric distribution with parameter p_i . Determine the distribution of the number of periods for which the system will function properly.

10. Let $f(x|r, p)$ denote the p.f. of the negative binomial distribution with parameters r and p , and let $f(x|\lambda)$ denote the p.f. of the Poisson distribution with mean λ , as defined by Eq. (5.4.2). Suppose $r \rightarrow \infty$ and $p \rightarrow 1$ in such a way that the value of $r(1 - p)$ remains constant and is equal to λ throughout the process. Show that for each fixed nonnegative integer x ,

$$f(x|r, p) \rightarrow f(x|\lambda).$$

11. Prove that the p.f. of the negative binomial distribution can be written in the following alternative form:

$$f(x|r, p) = \begin{cases} \binom{r-1}{x} p^r (-[1 - p])^x & \text{for } x = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

Hint: Use Exercise 10 in Sec. 5.3.

12. Suppose that a machine produces parts that are defective with probability P , but P is unknown. Suppose that

P has a continuous distribution with p.d.f.

$$f(p) = \begin{cases} 10(1 - p)^9 & \text{if } 0 < p < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional on $P = p$, assume that all parts are independent of each other. Let X be the number of nondefective parts observed until the first defective part. If we observe $X = 12$, compute the conditional p.d.f. of P given $X = 12$.

13. Let F be the c.d.f. of a discrete distribution that has the memoryless property stated in Theorem 5.5.5. Define $\ell(x) = \log[1 - F(x - 1)]$ for $x = 1, 2, \dots$

a. Show that, for all integers $t, h > 0$,

$$1 - F(h - 1) = \frac{1 - F(t + h - 1)}{1 - F(t - 1)}.$$

b. Prove that $\ell(t + h) = \ell(t) + \ell(h)$ for all integers $t, h > 0$.

c. Prove that $\ell(t) = t\ell(1)$ for every integer $t > 0$.

d. Prove that F must be the c.d.f. of a geometric distribution.

5.6 The Normal Distributions

The most widely used model for random variables with continuous distributions is the family of normal distributions. These distributions are the first ones we shall see whose p.d.f.'s cannot be integrated in closed form, and hence tables of the c.d.f. or computer programs are necessary in order to compute probabilities and quantiles for normal distributions.

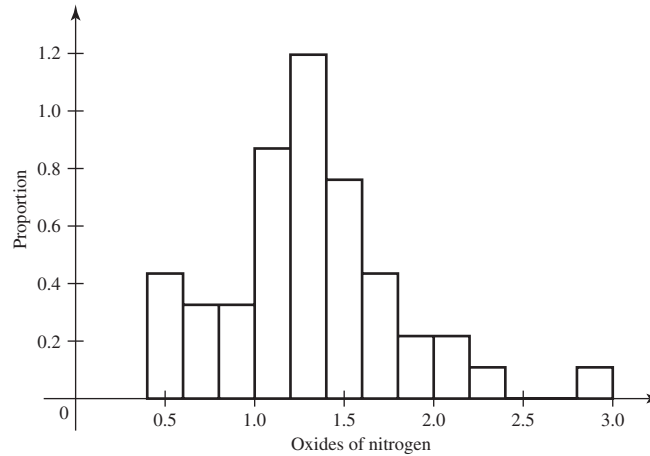
Importance of the Normal Distributions

Example 5.6.1

Automobile Emissions. Automobile engines emit a number of undesirable pollutants when they burn gasoline. Lorenzen (1980) studied the amounts of various pollutants emitted by 46 automobile engines. One class of pollutants consists of the oxides of nitrogen. Figure 5.1 shows a histogram of the 46 amounts of oxides of nitrogen (in grams per mile) that are reported by Lorenzen (1980). The bars in the histogram have areas that equal the proportions of the sample of 46 measurements that lie between the points on the horizontal axis where the sides of the bars stand. For example, the fourth bar (which runs from 1.0 to 1.2 on the horizontal axis) has area $0.870 \times 0.2 = 0.174$, which equals $8/46$ because there are eight observations between 1.0 and 1.2. When we want to make statements about probabilities related to emissions, we will need a distribution with which to model emissions. The family of normal distributions introduced in this section will prove to be valuable in examples such as this. ◀

The family of normal distributions, which will be defined and discussed in this section, is by far the single most important collection of probability distributions

Figure 5.1 Histogram of emissions of oxides of nitrogen for Example 5.6.1 in grams per mile over a common driving regimen.



in statistics. There are three main reasons for this preeminent position of these distributions.

The first reason is directly related to the mathematical properties of the normal distributions. We shall demonstrate in this section and in several later sections of this book that if a random sample is taken from a normal distribution, then the distributions of various important functions of the observations in the sample can be derived explicitly and will themselves have simple forms. Therefore, it is a mathematical convenience to be able to assume that the distribution from which a random sample is drawn is a normal distribution.

The second reason is that many scientists have observed that the random variables studied in various physical experiments often have distributions that are approximately normal. For example, a normal distribution will usually be a close approximation to the distribution of the heights or weights of individuals in a homogeneous population of people, corn stalks, or mice, or to the distribution of the tensile strength of pieces of steel produced by a certain process. Sometimes, a simple transformation of the observed random variables has a normal distribution.

The third reason for the preeminence of the normal distributions is the central limit theorem, which will be stated and proved in Sec. 6.3. If a large random sample is taken from some distribution, then even though this distribution is not itself approximately normal, a consequence of the central limit theorem is that many important functions of the observations in the sample will have distributions which are approximately normal. In particular, for a large random sample from any distribution that has a finite variance, the distribution of the average of the random sample will be approximately normal. We shall return to this topic in the next chapter.

Properties of Normal Distributions

Definition 5.6.1

Definition and p.d.f. A random variable X has the *normal distribution with mean μ and variance σ^2* ($-\infty < \mu < \infty$ and $\sigma > 0$) if X has a continuous distribution with the following p.d.f.:

$$f(x|\mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{for } -\infty < x < \infty. \quad (5.6.1)$$

We should first verify that the function defined in Eq. (5.6.1) is a p.d.f. Shortly thereafter, we shall verify that the mean and variance of the distribution with p.d.f. (5.6.1) are indeed μ and σ^2 , respectively.

Theorem 5.6.1

The function defined in Eq. (5.6.1) is a p.d.f.

Proof Clearly, the function is nonnegative. We must also show that

$$\int_{-\infty}^{\infty} f(x|\mu, \sigma^2) dx = 1. \quad (5.6.2)$$

If we let $y = (x - \mu)/\sigma$, then

$$\int_{-\infty}^{\infty} f(x|\mu, \sigma^2) dx = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}y^2\right) dy.$$

We shall now let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy. \quad (5.6.3)$$

Then we must show that $I = (2\pi)^{1/2}$.

From Eq. (5.6.3), it follows that

$$\begin{aligned} I^2 &= I \cdot I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}z^2\right) dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(y^2 + z^2)\right] dy dz. \end{aligned}$$

We shall now change the variables in this integral from y and z to the polar coordinates r and θ by letting $y = r \cos \theta$ and $z = r \sin \theta$. Then, since $y^2 + z^2 = r^2$,

$$I^2 = \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{1}{2}r^2\right) r dr d\theta = 2\pi, \quad (5.6.4)$$

where the inner integral in (5.6.4) is performed by substituting $v = r^2/2$ with $dv = r dr$, so the inner integral is

$$\int_0^{\infty} \exp(-v) dv = 1,$$

and the outer integral is 2π . Therefore, $I = (2\pi)^{1/2}$ and Eq. (5.6.2) has been established. ■

Example 5.6.2

Automobile Emissions. Consider the automobile engines described in Example 5.6.1. Figure 5.2 shows the histogram from Fig. 5.1 together with the normal p.d.f. having mean and variance chosen to match the observed data. Although the p.d.f. does not exactly match the shape of the histogram, it does correspond remarkably well. ◀

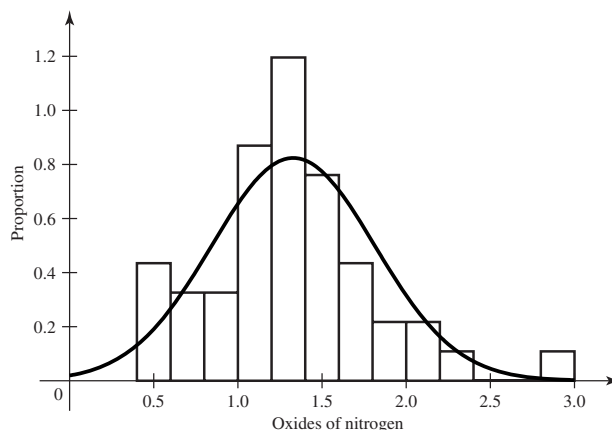
We could verify directly, using integration by parts, that the mean and variance of the distribution with p.d.f. given by Eq. (5.6.1) are, respectively, μ and σ^2 . (See Exercise 26.) However, we need the moment generating function anyway, and then we can just take two derivatives of the m.g.f. to find the first two moments.

Theorem 5.6.2

Moment Generating Function. The m.g.f. of the distribution with p.d.f. given by Eq. (5.6.1) is

$$\psi(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \quad \text{for } -\infty < t < \infty. \quad (5.6.5)$$

Figure 5.2 Histogram of emissions of oxides of nitrogen for Example 5.6.2 together with a matching normal p.d.f.



Proof By the definition of an m.g.f.,

$$\psi(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[tx - \frac{(x - \mu)^2}{2\sigma^2}\right] dx.$$

By completing the square inside the brackets (see Exercise 24), we obtain the relation

$$tx - \frac{(x - \mu)^2}{2\sigma^2} = \mu t + \frac{1}{2}\sigma^2 t^2 - \frac{[x - (\mu + \sigma^2 t)]^2}{2\sigma^2}.$$

Therefore,

$$\psi(t) = C \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

where

$$C = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{[x - (\mu + \sigma^2 t)]^2}{2\sigma^2}\right\} dx.$$

If we now replace μ with $\mu + \sigma^2 t$ in Eq. (5.6.1), it follows from Eq. (5.6.2) that $C = 1$. Hence, the m.g.f. of the normal distribution is given by Eq. (5.6.5). ■

We are now ready to verify the mean and variance.

Theorem 5.6.3 **Mean and Variance.** The mean and variance of the distribution with p.d.f. given by Eq. (5.6.1) are μ and σ^2 , respectively.

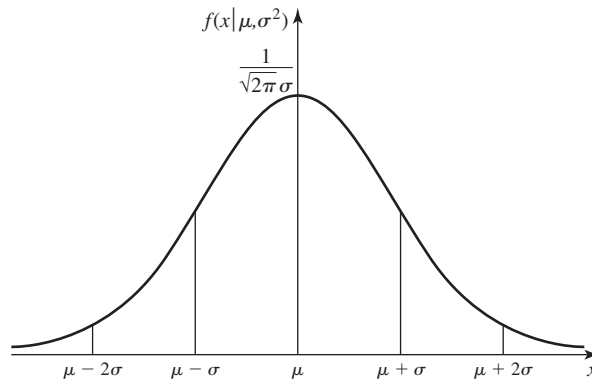
Proof The first two derivatives of the m.g.f. in Eq. (5.6.5) are

$$\begin{aligned}\psi'(t) &= (\mu + \sigma^2 t) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \\ \psi''(t) &= ([\mu + \sigma^2 t]^2 + \sigma^2) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)\end{aligned}$$

Plugging $t = 0$ into each of these derivatives yields

$$E(X) = \psi'(0) = \mu \quad \text{and} \quad \text{Var}(X) = \psi''(0) - [\psi'(0)]^2 = \sigma^2. \quad \blacksquare$$

Since the m.g.f. $\psi(t)$ is finite for all values of t , all the moments $E(X^k)$ ($k = 1, 2, \dots$) will also be finite.

Figure 5.3 The p.d.f. of a normal distribution.**Example 5.6.3**

Stock Price Changes. A popular model for the change in the price of a stock over a period of time of length u is to say that the price after time u is $S_u = S_0 e^{Z_u}$, where Z_u has the normal distribution with mean μu and variance $\sigma^2 u$. In this formula, S_0 is the present price of the stock, and σ is called the *volatility* of the stock price. The expected value of S_u can be computed from the m.g.f. ψ of Z_u :

$$E(S_u) = S_0 E(e^{Z_u}) = S_0 \psi(1) = S_0 e^{\mu u + \sigma^2 u/2}.$$

The Shapes of Normal Distributions It can be seen from Eq. (5.6.1) that the p.d.f. $f(x|\mu, \sigma^2)$ of the normal distribution with mean μ and variance σ^2 is symmetric with respect to the point $x = \mu$. Therefore, μ is both the mean and the median of the distribution. Furthermore, μ is also the mode of the distribution. In other words, the p.d.f. $f(x|\mu, \sigma^2)$ attains its maximum value at the point $x = \mu$. Finally, by differentiating $f(x|\mu, \sigma^2)$ twice, it can be found that there are points of inflection at $x = \mu + \sigma$ and at $x = \mu - \sigma$.

The p.d.f. $f(x|\mu, \sigma^2)$ is sketched in Fig. 5.3. It is seen that the curve is “bell-shaped.” However, it is not necessarily true that every arbitrary bell-shaped p.d.f. can be approximated by the p.d.f. of a normal distribution. For example, the p.d.f. of a Cauchy distribution, as sketched in Fig. 4.3, is a symmetric bell-shaped curve which apparently resembles the p.d.f. sketched in Fig. 5.3. However, since no moment of the Cauchy distribution—not even the mean—exists, the tails of the Cauchy p.d.f. must be quite different from the tails of the normal p.d.f.

Linear Transformations We shall now show that if a random variable X has a normal distribution, then every linear function of X will also have a normal distribution.

Theorem 5.6.4

If X has the normal distribution with mean μ and variance σ^2 and if $Y = aX + b$, where a and b are given constants and $a \neq 0$, then Y has the normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$.

Proof The m.g.f. ψ of X is given by Eq. (5.6.5). If ψ_Y denotes the m.g.f. of Y , then

$$\psi_Y(t) = e^{bt} \psi(at) = \exp \left[(a\mu + b)t + \frac{1}{2} a^2 \sigma^2 t^2 \right] \quad \text{for } -\infty < t < \infty.$$

By comparing this expression for ψ_Y with the m.g.f. of a normal distribution given in Eq. (5.6.5), we see that ψ_Y is the m.g.f. of the normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. Hence, Y must have this normal distribution. ■

The Standard Normal Distribution

Definition 5.6.2 **Standard Normal Distribution.** The normal distribution with mean 0 and variance 1 is called the *standard normal distribution*. The p.d.f. of the standard normal distribution is usually denoted by the symbol ϕ , and the c.d.f. is denoted by the symbol Φ . Thus,

$$\phi(x) = f(x|0, 1) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}x^2\right) \quad \text{for } -\infty < x < \infty \quad (5.6.6)$$

and

$$\Phi(x) = \int_{-\infty}^x \phi(u) du \quad \text{for } -\infty < x < \infty, \quad (5.6.7)$$

where the symbol u is used in Eq. (5.6.7) as a dummy variable of integration.

The c.d.f. $\Phi(x)$ cannot be expressed in closed form in terms of elementary functions. Therefore, probabilities for the standard normal distribution or any other normal distribution can be found only by numerical approximations or by using a table of values of $\Phi(x)$ such as the one given at the end of this book. In that table, the values of $\Phi(x)$ are given only for $x \geq 0$. Most computer packages that do statistical analysis contain functions that compute the c.d.f. and the quantile function of the standard normal distribution. Knowing the values of $\Phi(x)$ for $x \geq 0$ and $\Phi^{-1}(p)$ for $0.5 < p < 1$ is sufficient for calculating the c.d.f. and the quantile function of any normal distribution at any value, as the next two results show.

Theorem 5.6.5 **Consequences of Symmetry.** For all x and all $0 < p < 1$,

$$\Phi(-x) = 1 - \Phi(x) \quad \text{and} \quad \Phi^{-1}(p) = -\Phi^{-1}(1 - p). \quad (5.6.8)$$

Proof Since the p.d.f. of the standard normal distribution is symmetric with respect to the point $x = 0$, it follows that $\Pr(X \leq x) = \Pr(X \geq -x)$ for every number x ($-\infty < x < \infty$). Since $\Pr(X \leq x) = \Phi(x)$ and $\Pr(X \geq -x) = 1 - \Phi(-x)$, we have the first equation in Eq. (5.6.8). The second equation follows by letting $x = \Phi^{-1}(p)$ in the first equation and then applying the function Φ^{-1} to both sides of the equation. ■

Theorem 5.6.6 **Converting Normal Distributions to Standard.** Let X have the normal distribution with mean μ and variance σ^2 . Let F be the c.d.f. of X . Then $Z = (X - \mu)/\sigma$ has the standard normal distribution, and, for all x and all $0 < p < 1$,

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad (5.6.9)$$

$$F^{-1}(p) = \mu + \sigma \Phi^{-1}(p). \quad (5.6.10)$$

Proof It follows immediately from Theorem 5.6.4 that $Z = (X - \mu)/\sigma$ has the standard normal distribution. Therefore,

$$F(x) = \Pr(X \leq x) = \Pr\left(Z \leq \frac{x - \mu}{\sigma}\right),$$

which establishes Eq. (5.6.9). For Eq. (5.6.10), let $p = F(x)$ in Eq. (5.6.9) and then solve for x in the resulting equation. ■

**Example
5.6.4**

Determining Probabilities for a Normal Distribution. Suppose that X has the normal distribution with mean 5 and standard deviation 2. We shall determine the value of $\Pr(1 < X < 8)$.

If we let $Z = (X - 5)/2$, then Z will have the standard normal distribution and

$$\Pr(1 < X < 8) = \Pr\left(\frac{1-5}{2} < \frac{X-5}{2} < \frac{8-5}{2}\right) = \Pr(-2 < Z < 1.5).$$

Furthermore,

$$\begin{aligned}\Pr(-2 < Z < 1.5) &= \Pr(Z < 1.5) - \Pr(Z \leq -2) \\ &= \Phi(1.5) - \Phi(-2) \\ &= \Phi(1.5) - [1 - \Phi(2)].\end{aligned}$$

From the table at the end of this book, it is found that $\Phi(1.5) = 0.9332$ and $\Phi(2) = 0.9773$. Therefore,

$$\Pr(1 < X < 8) = 0.9105. \quad \blacktriangleleft$$

**Example
5.6.5**

Quantiles of Normal Distributions. Suppose that the engineers who collected the automobile emissions data in Example 5.6.1 are interested in finding out whether most engines are serious polluters. For example, they could compute the 0.05 quantile of the distribution of emissions and declare that 95 percent of the engines of the type tested exceed this quantile. Let X be the average grams of oxides of nitrogen per mile for a typical engine. Then the engineers modeled X as having a normal distribution. The normal distribution plotted in Fig. 5.2 has mean 1.329 and standard deviation 0.4844. The c.d.f. of X would then be $F(x) = \Phi([x - 1.329]/0.4844)$, and the quantile function would be $F^{-1}(p) = 1.329 + 0.4844\Phi^{-1}(p)$, where Φ^{-1} is the quantile function of the standard normal distribution, which can be evaluated using a computer or from tables. To find $\Phi^{-1}(p)$ from the table of Φ , find the closest value to p in the $\Phi(x)$ column and read the inverse from the x column. Since the table only has values of $p > 0.5$, we use Eq. (5.6.8) to conclude that $\Phi^{-1}(0.05) = -\Phi^{-1}(0.95)$. So, look up 0.95 in $\Phi(x)$ column (halfway between 0.9495 and 0.9505) to find $x = 1.645$ (halfway between 1.64 and 1.65) and conclude that $\Phi^{-1}(0.05) = -1.645$. The 0.05 quantile of X is then $1.329 + 0.4844 \times (-1.645) = 0.5322$. \blacktriangleleft

Comparisons of Normal Distributions

The p.d.f.'s of three normal distributions are sketched in Fig. 5.4 for a fixed value of μ and three different values of σ ($\sigma = 1/2, 1$, and 2). It can be seen from this figure that the p.d.f. of a normal distribution with a small value of σ has a high peak and is very concentrated around the mean μ , whereas the p.d.f. of a normal distribution with a larger value of σ is relatively flat and is spread out more widely over the real line.

An important fact is that every normal distribution contains the same total amount of probability within one standard deviation of its mean, the same amount within two standard deviations of its mean, and the same amount within any other fixed number of standard deviations of its mean. In general, if X has the normal distribution with mean μ and variance σ^2 , and if Z has the standard normal distribution, then for $k > 0$,

$$p_k = \Pr(|X - \mu| \leq k\sigma) = \Pr(|Z| \leq k).$$

In Table 5.2, the values of this probability p_k are given for various values of k . These probabilities can be computed from a table of Φ or using computer programs.

Figure 5.4 The normal p.d.f. for $\mu = 0$ and $\sigma = 1/2, 1, 2$.

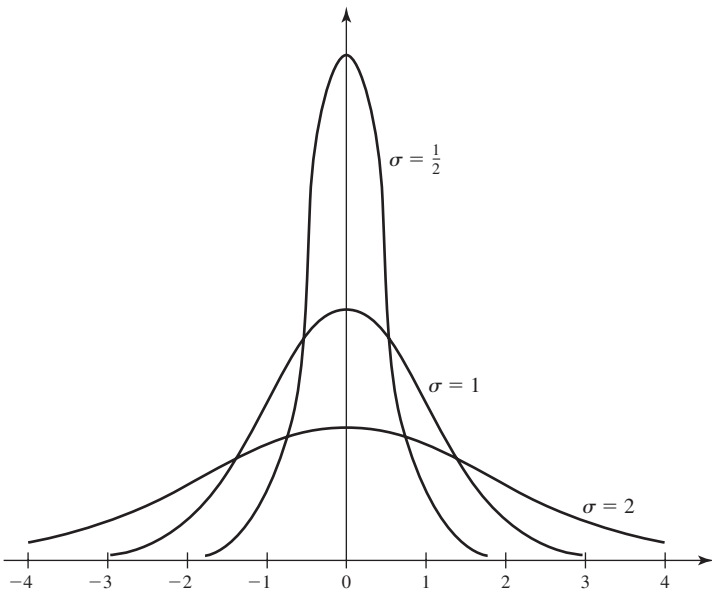


Table 5.2 Probabilities that normal random variables are within k standard deviations of their means	
k	p_k
1	0.6826
2	0.9544
3	0.9974
4	0.99994
5	$1 - 6 \times 10^{-7}$
10	$1 - 2 \times 10^{-23}$

Although the p.d.f. of a normal distribution is positive over the entire real line, it can be seen from this table that the total amount of probability outside an interval of four standard deviations on each side of the mean is only 0.00006.

Linear Combinations of Normally Distributed Variables

In the next theorem and corollary, we shall prove the following important result: Every linear combination of random variables that are independent and normally distributed will also have a normal distribution.

Theorem 5.6.7 If the random variables X_1, \dots, X_k are independent and if X_i has the normal distribution with mean μ_i and variance σ_i^2 ($i = 1, \dots, k$), then the sum $X_1 + \dots + X_k$ has the normal distribution with mean $\mu_1 + \dots + \mu_k$ and variance $\sigma_1^2 + \dots + \sigma_k^2$.

Proof Let $\psi_i(t)$ denote the m.g.f. of X_i for $i = 1, \dots, k$, and let $\psi(t)$ denote the m.g.f. of $X_1 + \dots + X_k$. Since the variables X_1, \dots, X_k are independent, then

$$\begin{aligned}\psi(t) &= \prod_{i=1}^k \psi_i(t) = \prod_{i=1}^k \exp\left(\mu_i t + \frac{1}{2}\sigma_i^2 t^2\right) \\ &= \exp\left[\left(\sum_{i=1}^k \mu_i\right)t + \frac{1}{2}\left(\sum_{i=1}^k \sigma_i^2\right)t^2\right] \quad \text{for } -\infty < t < \infty.\end{aligned}$$

From Eq. (5.6.5), the m.g.f. $\psi(t)$ can be identified as the m.g.f. of the normal distribution for which the mean is $\sum_{i=1}^k \mu_i$ and the variance is $\sum_{i=1}^k \sigma_i^2$. Hence, the distribution of $X_1 + \dots + X_k$ must be as stated in the theorem. ■

The following corollary is now obtained by combining Theorems 5.6.4 and 5.6.7.

Corollary
5.6.1

If the random variables X_1, \dots, X_k are independent, if X_i has the normal distribution with mean μ_i and variance σ_i^2 ($i = 1, \dots, k$), and if a_1, \dots, a_k and b are constants for which at least one of the values a_1, \dots, a_k is different from 0, then the variable $a_1 X_1 + \dots + a_k X_k + b$ has the normal distribution with mean $a_1 \mu_1 + \dots + a_k \mu_k + b$ and variance $a_1^2 \sigma_1^2 + \dots + a_k^2 \sigma_k^2$. ■

Example
5.6.6

Heights of Men and Women. Suppose that the heights, in inches, of the women in a certain population follow the normal distribution with mean 65 and standard deviation 1, and that the heights of the men follow the normal distribution with mean 68 and standard deviation 3. Suppose also that one woman is selected at random and, independently, one man is selected at random. We shall determine the probability that the woman will be taller than the man.

Let W denote the height of the selected woman, and let M denote the height of the selected man. Then the difference $W - M$ has the normal distribution with mean $65 - 68 = -3$ and variance $1^2 + 3^2 = 10$. Therefore, if we let

$$Z = \frac{1}{10^{1/2}}(W - M + 3),$$

then Z has the standard normal distribution. It follows that

$$\begin{aligned}\Pr(W > M) &= \Pr(W - M > 0) \\ &= \Pr\left(Z > \frac{3}{10^{1/2}}\right) = \Pr(Z > 0.949) \\ &= 1 - \Phi(0.949) = 0.171.\end{aligned}$$

Thus, the probability that the woman will be taller than the man is 0.171. ◀

Averages of random samples of normal random variables figure prominently in many statistical calculations. To fix notation, we start with a general definition.

Definition
5.6.3

Sample Mean. Let X_1, \dots, X_n be random variables. The average of these n random variables, $\frac{1}{n} \sum_{i=1}^n X_i$, is called their *sample mean* and is commonly denoted \bar{X}_n .

The following simple corollary to Corollary 5.6.1 gives the distribution of the sample mean of a random sample of normal random variables.

**Corollary
5.6.2**

Suppose that the random variables X_1, \dots, X_n form a random sample from the normal distribution with mean μ and variance σ^2 , and let \bar{X}_n denote their sample mean. Then \bar{X}_n has the normal distribution with mean μ and variance σ^2/n .

Proof Since $\bar{X}_n = \sum_{i=1}^n (1/n)X_i$, it follows from Corollary 5.6.1 that the distribution of \bar{X}_n is normal with mean $\sum_{i=1}^n (1/n)\mu = \mu$ and variance $\sum_{i=1}^n (1/n)^2\sigma^2 = \sigma^2/n$. ■

**Example
5.6.7**

Determining a Sample Size. Suppose that a random sample of size n is to be taken from the normal distribution with mean μ and variance 9. (The heights of men in Example 5.6.6 have such a distribution with $\mu = 68$.) We shall determine the minimum value of n for which

$$\Pr(|\bar{X}_n - \mu| \leq 1) \geq 0.95.$$

It is known from Corollary 5.6.2 that the sample mean \bar{X}_n will have the normal distribution for which the mean is μ and the standard deviation is $3/n^{1/2}$. Therefore, if we let

$$Z = \frac{n^{1/2}}{3}(\bar{X}_n - \mu),$$

then Z will have the standard normal distribution. In this example, n must be chosen so that

$$\Pr(|\bar{X}_n - \mu| \leq 1) = \Pr\left(|Z| \leq \frac{n^{1/2}}{3}\right) \geq 0.95. \quad (5.6.11)$$

For each positive number x , it will be true that $\Pr(|Z| \leq x) \geq 0.95$ if and only if $1 - \Phi(x) = \Pr(Z > x) \leq 0.025$. From the table of the standard normal distribution at the end of this book, it is found that $1 - \Phi(x) \leq 0.025$ if and only if $x \geq 1.96$. Therefore, the inequality in relation (5.6.11) will be satisfied if and only if

$$\frac{n^{1/2}}{3} \geq 1.96.$$

Since the smallest permissible value of n is 34.6, the sample size must be at least 35 in order that the specified relation will be satisfied. ◀

**Example
5.6.8**

Interval for Mean. Consider a population with a normal distribution such as the heights of men in Example 5.6.6. Suppose that we are not willing to specify the precise distribution as we did in that example, but rather only that the standard deviation is 3, leaving the mean μ unspecified. If we sample a number of men from this population, we could try to use their sampled heights to give us some idea what μ equals. A popular form of statistical inference that will be discussed in Sec. 8.5 finds an interval that has a specified probability of containing μ . To be specific, suppose that we observe a random sample of size n from the normal distribution with mean μ and standard deviation 3. Then, \bar{X}_n has the normal distribution with mean μ and standard deviation $3/n^{1/2}$ as in Example 5.6.7. Similarly, we can define

$$Z = \frac{n^{1/2}}{3}(\bar{X}_n - \mu),$$

which then has the standard normal distribution. Hence,

$$0.95 = \Pr(|Z| < 1.96) = \Pr\left(|\bar{X}_n - \mu| < 1.96 \frac{3}{n^{1/2}}\right). \quad (5.6.12)$$

It is easy to verify that

$$|\bar{X}_n - \mu| < 1.96 \frac{3}{n^{1/2}} \text{ if and only if } \bar{X}_n - 1.96 \frac{3}{n^{1/2}} < \mu < \bar{X}_n + 1.96 \frac{3}{n^{1/2}}. \quad (5.6.13)$$

The two inequalities in Eq. (5.6.13) hold if and only if the interval

$$\left(\bar{X}_n - 1.96 \frac{3}{n^{1/2}}, \bar{X}_n + 1.96 \frac{3}{n^{1/2}} \right) \quad (5.6.14)$$

contains the value of μ . It follows from Eq. (5.6.12) that the probability is 0.95 that the interval in (5.6.14) contains μ . Now, suppose that the sample size is $n = 36$. Then the half-width of the interval (5.6.14) is then $3/36^{1/2} = 0.98$. We will not know the endpoints of the interval until after we observe \bar{X}_n . However, we know now that the interval $(\bar{X}_n - 0.98, \bar{X}_n + 0.98)$ has probability 0.95 of containing μ . ◀

The Lognormal Distributions

It is very common to use normal distributions to model logarithms of random variables. For this reason, a name is given to the distribution of the original random variables before transforming.

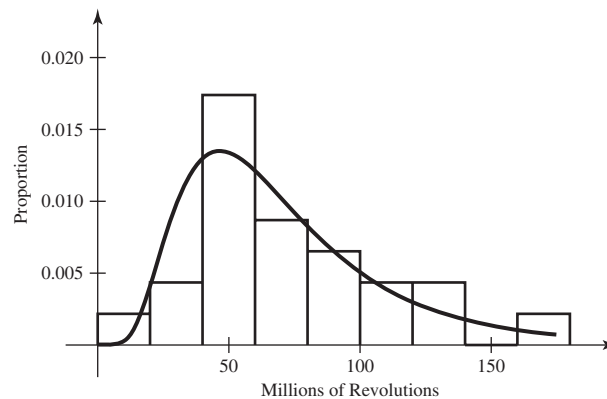
Definition 5.6.4

Lognormal Distribution. If $\log(X)$ has the normal distribution with mean μ and variance σ^2 , we say that X has the *lognormal distribution* with parameters μ and σ^2 .

Example 5.6.9

Failure Times of Ball Bearings. Products that are subject to wear and tear are generally tested for endurance in order to estimate their useful lifetimes. Lawless (1982, example 5.2.2) describes data taken from Lieblein and Zelen (1956), which are measurements of the numbers of millions of revolutions before failure for 23 ball bearings. The lognormal distribution is one popular model for times until failure. Figure 5.5 shows a histogram of the 23 lifetimes together with a lognormal p.d.f. with parameters chosen to match the observed data. The bars of the histogram in Fig. 5.5 have areas that equal the proportions of the sample that lie between the points on the horizontal axis where the sides of the bars stand. Suppose that the engineers are interested in knowing how long to wait until there is a 90 percent chance that a ball

Figure 5.5 Histogram of lifetimes of ball bearings and fitted lognormal p.d.f. for Example 5.6.9.



bearing will have failed. Then they want the 0.9 quantile of the distribution of life-times. Let X be the time to failure of a ball bearing. The lognormal distribution of X plotted in Fig. 5.5 has parameters 4.15 and 0.5334². The c.d.f. of X would then be $F(x) = \Phi([\log(x) - 4.15]/0.5334)$, and the quantile function would be

$$F^{-1}(p) = e^{4.15 + 0.5334\Phi^{-1}(p)},$$

where Φ^{-1} is the quantile function of the standard normal distribution. With $p = 0.9$, we get $\Phi^{-1}(0.9) = 1.28$ and $F^{-1}(0.9) = 125.6$. ◀

The moments of a lognormal random variable are easy to compute based on the m.g.f. of a normal distribution. If $Y = \log(X)$ has the normal distribution with mean μ and variance σ^2 , then the m.g.f. of Y is $\psi(t) = \exp(\mu t + 0.5\sigma^2 t^2)$. However, the definition of ψ is $\psi(t) = E(e^{tY})$. Since $Y = \log(X)$, we have

$$\psi(t) = E(e^{tY}) = E(e^{t\log(X)}) = E(X^t).$$

It follows that $E(X^t) = \psi(t)$ for all real t . In particular, the mean and variance of X are

$$E(X) = \psi(1) = \exp(\mu + 0.5\sigma^2), \quad (5.6.15)$$

$$\text{Var}(X) = \psi(2) - \psi(1)^2 = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1].$$

Example 5.6.10

Stock and Option Prices. Consider a stock like the one in Example 5.6.3 whose current price is S_0 . Suppose that the price at u time units in the future is $S_u = S_0 e^{Z_u}$, where Z_u has the normal distribution with mean μu and variance $\sigma^2 u$. Note that $S_0 e^{Z_u} = e^{Z_u + \log(S_0)}$ and $Z_u + \log(S_0)$ has the normal distribution with mean $\mu u + \log(S_0)$ and variance $\sigma^2 u$. So S_u has the lognormal distribution with parameters $\mu u + \log(S_0)$ and $\sigma^2 u$.

Black and Scholes (1973) developed a pricing scheme for options on stocks whose prices follow a lognormal distribution. For the remainder of this example, we shall consider a single time u and write the stock price as $S_u = S_0 e^{\mu u + \sigma u^{1/2} Z}$, where Z has the standard normal distribution. Suppose that we need to price the option to buy one share of the above stock for the price q at a particular time u in the future. As in Example 4.1.14 on page 214, we shall use risk-neutral pricing. That is, we force the present value of $E(S_u)$ to equal S_0 . If u is measured in years and the risk-free interest rate is r per year, then the present value of $E(S_u)$ is $e^{-ru} E(S_u)$. (This assumes that compounding of interest is done continuously instead of just once as it was in Example 4.1.14. The effect of continuous compounding is examined in Exercise 25.) But $E(S_u) = S_0 e^{\mu u + \sigma^2 u/2}$. Setting S_0 equal to $e^{-ru} S_0 e^{\mu u + \sigma^2 u/2}$ yields $\mu = r - \sigma^2/2$ when doing risk-neutral pricing.

Now we can determine a price for the specified option. The value of the option at time u will be $h(S_u)$, where

$$h(s) = \begin{cases} s - q & \text{if } s > q, \\ 0 & \text{otherwise.} \end{cases}$$

Set $\mu = r - \sigma^2/2$, and it is easy to see that $h(S_u) > 0$ if and only if

$$Z > \frac{\log\left(\frac{q}{S_0}\right) - (r - \sigma^2/2)u}{\sigma u^{1/2}}. \quad (5.6.16)$$

We shall refer to the constant on the right-hand side of Eq. (5.6.16) as c . The risk-neutral price of the option is the present value of $E(h(S_u))$, which equals

$$e^{-ru} E[h(S_u)] = e^{-ru} \int_c^\infty \left[S_0 e^{[r-\sigma^2/2]u + \sigma u^{1/2}z} - q \right] \frac{1}{(2\pi)^{1/2}} e^{-z^2/2} dz. \quad (5.6.17)$$

To compute the integral in Eq. (5.6.17), split the integrand into two parts at the $-q$. The second integral is then just a constant times the integral of a normal p.d.f., namely,

$$-e^{-ru} q \int_c^\infty \frac{1}{(2\pi)^{1/2}} e^{-z^2/2} dz = -e^{-ru} q [1 - \Phi(c)].$$

The first integral in Eq. (5.6.17), is

$$e^{-\sigma^2 u/2} S_0 \int_c^\infty \frac{1}{(2\pi)^{1/2}} e^{-z^2/2 + \sigma u^{1/2}z} dz.$$

This can be converted into the integral of a normal p.d.f. times a constant by completing the square (see Exercise 24). The result of completing the square is

$$e^{-\sigma^2 u/2} S_0 \int_c^\infty \frac{1}{(2\pi)^{1/2}} e^{-(z - \sigma u^{1/2})^2/2 + \sigma^2 u/2} dz = S_0 [1 - \Phi(c - \sigma u^{1/2})].$$

Finally, combine the two integrals into the option price, using the fact that $1 - \Phi(x) = \Phi(-x)$:

$$S_0 \Phi(\sigma u^{1/2} - c) - q e^{-ru} \Phi(-c). \quad (5.6.18)$$

This is the famous *Black-Scholes formula* for pricing options. As a simple example, suppose that $q = S_0$, $r = 0.06$ (6 percent interest), $u = 1$ (one year wait), and $\sigma = 0.1$. Then (5.6.18) says that the option price should be $0.0746S_0$. If the distribution of S_u is different from the form used here, simulation techniques (see Chapter 12) can be used to help price options. ◀

The p.d.f.'s of the lognormal distributions will be found in Exercise 17 of this section. The c.d.f. of each lognormal distribution is easily constructed from the standard normal c.d.f. Φ . Let X have the lognormal distribution with parameters μ and σ^2 . Then

$$\Pr(X \leq x) = \Pr(\log(X) \leq \log(x)) = \Phi\left(\frac{\log(x) - \mu}{\sigma}\right).$$

The results from earlier in this section about linear combinations of normal random variables translate into results about products of powers of lognormal random variables. Results about sums of independent normal random variables translate into results about products of independent lognormal random variables.

Summary

We introduced the family of normal distributions. The parameters of each normal distribution are its mean and variance. A linear combination of independent normal random variables has the normal distribution with mean equal to the linear combination of the means and variance determined by Corollary 4.3.1. In particular, if X has the normal distribution with mean μ and variance σ^2 , then $(X - \mu)/\sigma$ has the standard normal distribution (mean 0 and variance 1). Probabilities and quantiles for normal distributions can be obtained from tables or computer programs for standard normal probabilities and quantiles. For example, if X has the normal distribution with mean μ and variance σ^2 , then the c.d.f. of X is $F(x) = \Phi([x - \mu]/\sigma)$ and the quantile function of X is $F^{-1}(p) = \mu + \Phi^{-1}(p)\sigma$, where Φ is the standard normal c.d.f.

Exercises

1. Find the 0.5, 0.25, 0.75, 0.1, and 0.9 quantiles of the standard normal distribution.

2. Suppose that X has the normal distribution for which the mean is 1 and the variance is 4. Find the value of each of the following probabilities:

- a. $\Pr(X \leq 3)$ b. $\Pr(X > 1.5)$
- c. $\Pr(X = 1)$ d. $\Pr(2 < X < 5)$
- e. $\Pr(X \geq 0)$ f. $\Pr(-1 < X < 0.5)$
- g. $\Pr(|X| \leq 2)$ h. $\Pr(1 \leq -2X + 3 \leq 8)$

3. If the temperature in degrees Fahrenheit at a certain location is normally distributed with a mean of 68 degrees and a standard deviation of 4 degrees, what is the distribution of the temperature in degrees Celsius at the same location?

4. Find the 0.25 and 0.75 quantiles of the Fahrenheit temperature at the location mentioned in Exercise 3.

5. Let X_1 , X_2 , and X_3 be independent lifetimes of memory chips. Suppose that each X_i has the normal distribution with mean 300 hours and standard deviation 10 hours. Compute the probability that at least one of the three chips lasts at least 290 hours.

6. If the m.g.f. of a random variable X is $\psi(t) = e^{t^2}$ for $-\infty < t < \infty$, what is the distribution of X ?

7. Suppose that the measured voltage in a certain electric circuit has the normal distribution with mean 120 and standard deviation 2. If three independent measurements of the voltage are made, what is the probability that all three measurements will lie between 116 and 118?

8. Evaluate the integral $\int_0^\infty e^{-3x^2} dx$.

9. A straight rod is formed by connecting three sections A , B , and C , each of which is manufactured on a different machine. The length of section A , in inches, has the normal distribution with mean 20 and variance 0.04. The length of section B , in inches, has the normal distribution with mean 14 and variance 0.01. The length of section C , in inches, has the normal distribution with mean 26 and variance 0.04. As indicated in Fig. 5.6, the three sections are joined so that there is an overlap of 2 inches at each connection. Suppose that the rod can be used in the construction of an airplane wing if its total length in inches is between 55.7 and 56.3. What is the probability that the rod can be used?

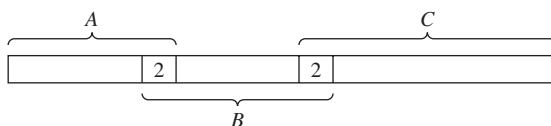


Figure 5.6 Sections of the rod in Exercise 9.

10. If a random sample of 25 observations is taken from the normal distribution with mean μ and standard deviation 2, what is the probability that the sample mean will lie within one unit of μ ?

11. Suppose that a random sample of size n is to be taken from the normal distribution with mean μ and standard deviation 2. Determine the smallest value of n such that

$$\Pr(|\bar{X}_n - \mu| < 0.1) \geq 0.9.$$

12.

a. Sketch the c.d.f. Φ of the standard normal distribution from the values given in the table at the end of this book.

b. From the sketch given in part (a) of this exercise, sketch the c.d.f. of the normal distribution for which the mean is -2 and the standard deviation is 3.

13. Suppose that the diameters of the bolts in a large box follow a normal distribution with a mean of 2 centimeters and a standard deviation of 0.03 centimeter. Also, suppose that the diameters of the holes in the nuts in another large box follow the normal distribution with a mean of 2.02 centimeters and a standard deviation of 0.04 centimeter. A bolt and a nut will fit together if the diameter of the hole in the nut is greater than the diameter of the bolt and the difference between these diameters is not greater than 0.05 centimeter. If a bolt and a nut are selected at random, what is the probability that they will fit together?

14. Suppose that on a certain examination in advanced mathematics, students from university A achieve scores that are normally distributed with a mean of 625 and a variance of 100, and students from university B achieve scores which are normally distributed with a mean of 600 and a variance of 150. If two students from university A and three students from university B take this examination, what is the probability that the average of the scores of the two students from university A will be greater than the average of the scores of the three students from university B ? *Hint:* Determine the distribution of the difference between the two averages.

15. Suppose that 10 percent of the people in a certain population have the eye disease glaucoma. For persons who have glaucoma, measurements of eye pressure X will be normally distributed with a mean of 25 and a variance of 1. For persons who do not have glaucoma, the pressure X will be normally distributed with a mean of 20 and a variance of 1. Suppose that a person is selected at random from the population and her eye pressure X is measured.

a. Determine the conditional probability that the person has glaucoma given that $X = x$.

b. For what values of x is the conditional probability in part (a) greater than $1/2$?

- 16.** Suppose that the joint p.d.f. of two random variables X and Y is

$$f(x, y) = \frac{1}{2\pi} e^{-(1/2)(x^2 + y^2)} \quad \text{for } -\infty < x < \infty \\ \text{and } -\infty < y < \infty.$$

Find $\Pr(-\sqrt{2} < X + Y < 2\sqrt{2})$.

- 17.** Consider a random variable X having the lognormal distribution with parameters μ and σ^2 . Determine the p.d.f. of X .

- 18.** Suppose that the random variables X and Y are independent and that each has the standard normal distribution. Show that the quotient X/Y has the Cauchy distribution.

- 19.** Suppose that the measurement X of pressure made by a device in a particular system has the normal distribution with mean μ and variance 1, where μ is the true pressure. Suppose that the true pressure μ is unknown but has the uniform distribution on the interval $[5, 15]$. If $X = 8$ is observed, find the conditional p.d.f. of μ given $X = 8$.

- 20.** Let X have the lognormal distribution with parameters 3 and 1.44. Find the probability that $X \leq 6.05$.

- 21.** Let X and Y be independent random variables such that $\log(X)$ has the normal distribution with mean 1.6 and variance 4.5 and $\log(Y)$ has the normal distribution with mean 3 and variance 6. Find the distribution of the product XY .

- 22.** Suppose that X has the lognormal distribution with parameters μ and σ^2 . Find the distribution of $1/X$.

- 23.** Suppose that X has the lognormal distribution with parameters 4.1 and 8. Find the distribution of $3X^{1/2}$.

- 24.** The method of *completing the square* is used several times in this text. It is a useful method for combining several quadratic and linear polynomials into a perfect square plus a constant. Prove the following identity, which is one general form of completing the square:

$$\begin{aligned} & \sum_{i=1}^n a_i (x - b_i)^2 + cx \\ &= \left(\sum_{i=1}^n a_i \right) \left(x - \frac{\sum_{i=1}^n a_i b_i - c/2}{\sum_{i=1}^n a_i} \right)^2 \\ &+ \sum_{i=1}^n a_i \left(b_i - \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n a_i} \right)^2 \\ &+ \left(\sum_{i=1}^n a_i \right)^{-1} \left[c \sum_{i=1}^n a_i b_i - c^2/4 \right] \end{aligned}$$

if $\sum_{i=1}^n a_i \neq 0$.

- 25.** In Example 5.6.10, we considered the effect of continuous compounding of interest. Suppose that S_0 dollars earn a rate of r per year compounded continuously for u years. Prove that the principal plus interest at the end of this time equals $S_0 e^{ru}$. *Hint:* Suppose that interest is compounded n times at intervals of u/n years each. At the end of each of the n intervals, the principal gets multiplied by $1 + ru/n$. Take the limit of the result as $n \rightarrow \infty$.

- 26.** Let X have the normal distribution whose p.d.f. is given by (5.6.6). Instead of using the m.g.f., derive the variance of X using integration by parts.

5.7 The Gamma Distributions

The family of gamma distributions is a popular model for random variables that are known to be positive. The family of exponential distributions is a subfamily of the gamma distributions. The times between successive occurrences in a Poisson process have an exponential distribution. The gamma function, related to the gamma distributions, is an extension of factorials from integers to all positive numbers.

The Gamma Function

Example 5.7.1

Mean and Variance of Lifetime of a Light Bulb. Suppose that we model the lifetime of a light bulb as a continuous random variable with the following p.d.f.:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If we wish to compute the mean and variance of such a lifetime, we need to compute the following integrals:

$$\int_0^{\infty} x e^{-x} dx, \quad \text{and} \quad \int_0^{\infty} x^2 e^{-x} dx. \quad (5.7.1)$$

These integrals are special cases of an important function that we examine next. ◀

Definition 5.7.1 The Gamma Function. For each positive number α , let the value $\Gamma(\alpha)$ be defined by the following integral:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx. \quad (5.7.2)$$

The function Γ defined by Eq. (5.7.2) for $\alpha > 0$ is called the *gamma function*.

As an example,

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1. \quad (5.7.3)$$

The following result, together with Eq. (5.7.3), shows that $\Gamma(\alpha)$ is finite for every value of $\alpha > 0$.

Theorem 5.7.1 If $\alpha > 1$, then

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1). \quad (5.7.4)$$

Proof We shall apply the method of integration by parts to the integral in Eq. (5.7.2). If we let $u = x^{\alpha-1}$ and $dv = e^{-x} dx$, then $du = (\alpha - 1)x^{\alpha-2} dx$ and $v = -e^{-x}$. Therefore,

$$\begin{aligned} \Gamma(\alpha) &= \int_0^{\infty} u dv = [uv]_0^{\infty} - \int_0^{\infty} v du \\ &= [-x^{\alpha-1}e^{-x}]_{x=0}^{\infty} + (\alpha - 1) \int_0^{\infty} x^{\alpha-2}e^{-x} dx \\ &= 0 + (\alpha - 1)\Gamma(\alpha - 1). \quad \blacksquare \end{aligned}$$

For integer values of α , we have a simple expression for the gamma function.

Theorem 5.7.2 For every positive integer n ,

$$\Gamma(n) = (n - 1)!. \quad (5.7.5)$$

Proof It follows from Theorem 5.7.1 that for every integer $n \geq 2$,

$$\begin{aligned} \Gamma(n) &= (n - 1)\Gamma(n - 1) = (n - 1)(n - 2)\Gamma(n - 2) \\ &= (n - 1)(n - 2) \cdots 1 \cdot \Gamma(1) \\ &= (n - 1)!\Gamma(1). \end{aligned}$$

Since $\Gamma(1) = 1 = 0!$ by Eq. (5.7.3), the proof is complete. ◻

Example 5.7.2

Mean and Variance of Lifetime of a Light Bulb. The two integrals in (5.7.1) are, respectively, $\Gamma(2) = 1! = 1$ and $\Gamma(3) = 2! = 2$. It follows that the mean of each lifetime is 1, and the variance is $2 - 1^2 = 1$. ◀

In many statistical applications, $\Gamma(\alpha)$ must be evaluated when α is either a positive integer or of the form $\alpha = n + (1/2)$ for some positive integer n . It follows from

Eq. (5.7.4) that for each positive integer n ,

$$\Gamma\left(n + \frac{1}{2}\right) = \left(n - \frac{1}{2}\right) \left(n - \frac{3}{2}\right) \cdots \left(\frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right). \quad (5.7.6)$$

Hence, it will be possible to determine the value of $\Gamma\left(n + \frac{1}{2}\right)$ if we can evaluate $\Gamma\left(\frac{1}{2}\right)$.

From Eq. (5.7.2),

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} x^{-1/2} e^{-x} dx.$$

If we let $x = (1/2)y^2$ in this integral, then $dx = y dy$ and

$$\Gamma\left(\frac{1}{2}\right) = 2^{1/2} \int_0^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy. \quad (5.7.7)$$

Because the integral of the p.d.f. of the standard normal distribution is equal to 1, it follows that

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy = (2\pi)^{1/2}. \quad (5.7.8)$$

Because the integrand in (5.7.8) is symmetric around $y = 0$,

$$\int_0^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy = \frac{1}{2}(2\pi)^{1/2} = \left(\frac{\pi}{2}\right)^{1/2}.$$

It now follows from Eq. (5.7.7) that

$$\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}. \quad (5.7.9)$$

For example, it is found from Eqs. (5.7.6) and (5.7.9) that

$$\Gamma\left(\frac{7}{2}\right) = \left(\frac{5}{2}\right) \left(\frac{3}{2}\right) \left(\frac{1}{2}\right) \pi^{1/2} = \frac{15}{8} \pi^{1/2}.$$

We present two final useful results before we introduce the gamma distributions.

**Theorem
5.7.3**

For each $\alpha > 0$ and each $\beta > 0$,

$$\int_0^{\infty} x^{\alpha-1} \exp(\beta x) dx = \frac{\Gamma(\alpha)}{\beta^{\alpha}}. \quad (5.7.10)$$

Proof Make the change of variables $y = \beta x$ so that $x = y/\beta$ and $dx = dy/\beta$. The result now follows easily from Eq. (5.7.2). ■

There is a version of Stirling's formula (Theorem 1.7.5) for the gamma function, which we state without proof.

**Theorem
5.7.4**

Stirling's Formula. $\lim_{x \rightarrow \infty} \frac{(2\pi)^{1/2} x^{x-1/2} e^{-x}}{\Gamma(x)} = 1.$ ■

**Example
5.7.3**

Service Times in a Queue. For $i = 1, \dots, n$, suppose that customer i in a queue must wait time X_i for service once reaching the head of the queue. Let Z be the rate at which the average customer is served. A typical probability model for this situation

is to say that, conditional on $Z = z$, X_1, \dots, X_n are i.i.d. with a distribution having the conditional p.d.f. $g_1(x_i|z) = z \exp(-zx_i)$ for $x_i > 0$. Suppose that Z is also unknown and has the p.d.f. $f_2(z) = 2 \exp(-2z)$ for $z > 0$. The joint p.d.f. of X_1, \dots, X_n, Z is then

$$\begin{aligned} f(x_1, \dots, x_n, z) &= \prod_{i=1}^n g_1(x_i|z) f_2(z) \\ &= 2z^n \exp(-z[2 + x_1 + \dots + x_n]), \end{aligned} \quad (5.7.11)$$

if $z, x_1, \dots, x_n > 0$ and 0 otherwise. In order to calculate the marginal joint distribution of X_1, \dots, X_n , we must integrate z out of the joint p.d.f. above. We can apply Theorem 5.7.3 with $\alpha = n + 1$ and $\beta = 2 + x_1 + \dots + x_n$ together with Theorem 5.7.2 to integrate the function in Eq. (5.7.11). The result is

$$\int_0^\infty f(x_1, \dots, x_n, z) dz = \frac{2(n!)}{(2 + \sum_{i=1}^n x_i)^{n+1}}, \quad (5.7.12)$$

for all $x_i > 0$ and 0 otherwise. This is the same joint p.d.f. that was used in Example 3.7.5 on page 154. ◀

The Gamma Distributions

Example 5.7.4

Service Times in a Queue. In Example 5.7.3, suppose that we observe the service times of n customers and want to find the conditional distribution of the rate Z . We can easily find the conditional p.d.f. $g_2(z|x_1, \dots, x_n)$ of Z given $X_1 = x_1, \dots, X_n = x_n$ by dividing the joint p.d.f. of X_1, \dots, X_n, Z in Eq. (5.7.11) by the p.d.f. of X_1, \dots, X_n in Eq. (5.7.12). The calculation is simplified by defining $y = 2 + \sum_{i=1}^n x_i$. We then obtain

$$g_2(z|x_1, \dots, x_n) = \begin{cases} \frac{y^{n+1}}{n!} e^{-yz}, & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \quad \blacktriangleleft$$

Distributions with p.d.f.'s like the one at the end of Example 5.7.4 are members of a commonly used family, which we now define.

Definition 5.7.2

Gamma Distributions. Let α and β be positive numbers. A random variable X has the *gamma distribution with parameters α and β* if X has a continuous distribution for which the p.d.f. is

$$f(x|\alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases} \quad (5.7.13)$$

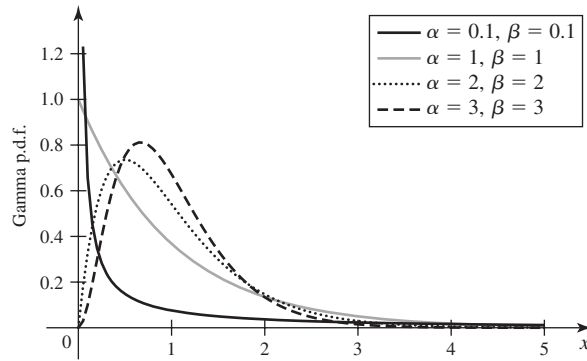
That the integral of the p.d.f. in Eq. (5.7.13) is 1 follows easily from Theorem 5.7.3.

Example 5.7.5

Service Times in a Queue. In Example 5.7.4, we can easily recognize the conditional p.d.f. as the p.d.f. of the gamma distribution with parameters $\alpha = n + 1$ and $\beta = y$. ◀

If X has a gamma distribution, then the moments of X are easily found from Eqs. (5.7.13) and (5.7.10).

Figure 5.7 Graphs of the p.d.f.'s of several different gamma distributions with common mean of 1.



Theorem 5.7.5 Moments. Let X have the gamma distribution with parameters α and β . For $k = 1, 2, \dots$,

$$E(X^k) = \frac{\Gamma(\alpha + k)}{\beta^k \Gamma(\alpha)} = \frac{\alpha(\alpha + 1) \cdots (\alpha + k - 1)}{\beta^k}.$$

In particular, $E(X) = \frac{\alpha}{\beta}$, and $\text{Var}(X) = \frac{\alpha}{\beta^2}$.

Proof For $k = 1, 2, \dots$,

$$\begin{aligned} E(X^k) &= \int_0^\infty x^k f(x|\alpha, \beta) dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+k-1} e^{-\beta x} dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha + k)}{\beta^{\alpha+k}} = \frac{\Gamma(\alpha + k)}{\beta^k \Gamma(\alpha)}. \end{aligned} \quad (5.7.14)$$

The expression for $E(X)$ follows immediately from (5.7.14). The variance can be computed as

$$\text{Var}(X) = \frac{\alpha(\alpha + 1)}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}. \quad \blacksquare$$

Figure 5.7 shows several gamma distribution p.d.f.'s that all have mean equal to 1 but different values of α and β .

Example 5.7.6 Service Times in a Queue. In Example 5.7.5, the conditional mean service rate given the observations $X_1 = x_1, \dots, X_n = x_n$ is

$$E(Z|x_1, \dots, x_n) = \frac{n + 1}{2 + \sum_{i=1}^n x_i}.$$

For large n , the conditional mean is approximately 1 over the sample average of the service times. This makes sense since 1 over the average service time is what we generally mean by service rate. \blacktriangleleft

The m.g.f. ψ of X can be obtained similarly.

Theorem 5.7.6 Moment Generating Function. Let X have the gamma distribution with parameters α and β . The m.g.f. of X is

$$\psi(t) = \left(\frac{\beta}{\beta - t}\right)^\alpha \quad \text{for } t < \beta. \quad (5.7.15)$$

Proof The m.g.f. is

$$\psi(t) = \int_0^\infty e^{tx} f(x|\alpha, \beta) dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x} dx.$$

This integral will be finite for every value of t such that $t < \beta$. Therefore, it follows from Eq. (5.7.10) that, for $t < \beta$,

$$\psi(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{(\beta - t)^\alpha} = \left(\frac{\beta}{\beta - t} \right)^\alpha. \quad \blacksquare$$

We can now show that the sum of independent random variables that have gamma distributions with a common value of the parameter β will also have a gamma distribution.

Theorem 5.7.7

If the random variables X_1, \dots, X_k are independent, and if X_i has the gamma distribution with parameters α_i and β ($i = 1, \dots, k$), then the sum $X_1 + \dots + X_k$ has the gamma distribution with parameters $\alpha_1 + \dots + \alpha_k$ and β .

Proof If ψ_i denotes the m.g.f. of X_i , then it follows from Eq. (5.7.15) that for $i = 1, \dots, k$,

$$\psi_i(t) = \left(\frac{\beta}{\beta - t} \right)^{\alpha_i} \quad \text{for } t < \beta.$$

If ψ denotes the m.g.f. of the sum $X_1 + \dots + X_k$, then by Theorem 4.4.4,

$$\psi(t) = \prod_{i=1}^k \psi_i(t) = \left(\frac{\beta}{\beta - t} \right)^{\alpha_1 + \dots + \alpha_k} \quad \text{for } t < \beta.$$

The m.g.f. ψ can now be recognized as the m.g.f. of the gamma distribution with parameters $\alpha_1 + \dots + \alpha_k$ and β . Hence, the sum $X_1 + \dots + X_k$ must have this gamma distribution. \blacksquare

The Exponential Distributions

A special case of gamma distributions provide a common model for phenomena such as waiting times. For instance, in Example 5.7.3, the conditional distribution of each service time X_i given Z (the rate of service) is a member of the following family of distributions.

Definition 5.7.3

Exponential Distributions. Let $\beta > 0$. A random variable X has the *exponential distribution with parameter β* if X has a continuous distribution with the p.d.f.

$$f(x|\beta) = \begin{cases} \beta e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases} \quad (5.7.16)$$

A comparison of the p.d.f.'s for gamma and exponential distributions makes the following result obvious.

Theorem 5.7.8

The exponential distribution with parameter β is the same as the gamma distribution with parameters $\alpha = 1$ and β . If X has the exponential distribution with parameter β , then

$$E(X) = \frac{1}{\beta} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\beta^2}, \quad (5.7.17)$$

and the m.g.f. of X is

$$\psi(t) = \frac{\beta}{\beta - t} \quad \text{for } t < \beta. \quad \blacksquare$$

Exponential distributions have a memoryless property similar to that stated in Theorem 5.5.5 for geometric distributions.

Theorem 5.7.9 **Memoryless Property of Exponential Distributions.** Let X have the exponential distribution with parameter β , and let $t > 0$. Then for every number $h > 0$,

$$\Pr(X \geq t + h | X \geq t) = \Pr(X \geq h). \quad (5.7.18)$$

Proof For each $t > 0$,

$$\Pr(X \geq t) = \int_t^{\infty} \beta e^{-\beta x} dx = e^{-\beta t}. \quad (5.7.19)$$

Hence, for each $t > 0$ and each $h > 0$,

$$\begin{aligned} \Pr(X \geq t + h | X \geq t) &= \frac{\Pr(X \geq t + h)}{\Pr(X \geq t)} \\ &= \frac{e^{-\beta(t+h)}}{e^{-\beta t}} = e^{-\beta h} = \Pr(X \geq h). \end{aligned} \quad (5.7.20) \quad \blacksquare$$

You can prove (see Exercise 23) that the exponential distributions are the only continuous distributions with the memoryless property.

To illustrate the memoryless property, we shall suppose that X represents the number of minutes that elapse before some event occurs. According to Eq. (5.7.20), if the event has not occurred during the first t minutes, then the probability that the event will not occur during the next h minutes is simply $e^{-\beta h}$. This is the same as the probability that the event would not occur during an interval of h minutes starting from time 0. In other words, regardless of the length of time that has elapsed without the occurrence of the event, the probability that the event will occur during the next h minutes always has the same value.

This memoryless property will not strictly be satisfied in all practical problems. For example, suppose that X is the length of time for which a light bulb will burn before it fails. The length of time for which the bulb can be expected to continue to burn in the future will depend on the length of time for which it has been burning in the past. Nevertheless, the exponential distribution has been used effectively as an approximate distribution for such variables as the lengths of the lives of various products.

Life Tests

Example 5.7.7

Light Bulbs. Suppose that n light bulbs are burning simultaneously in a test to determine the lengths of their lives. We shall assume that the n bulbs burn independently of one another and that the lifetime of each bulb has the exponential distribution with parameter β . In other words, if X_i denotes the lifetime of bulb i , for $i = 1, \dots, n$, then it is assumed that the random variables X_1, \dots, X_n are i.i.d. and that each has the exponential distribution with parameter β . What is the distribution of the length of time Y_1 until the first failure of one of the n bulbs? What is the distribution of the length of time Y_2 after the first failure until a second bulb fails? ◀

The random variable Y_1 in Example 5.7.7 is the minimum of a random sample of n exponential random variables. The distribution of Y_1 is easy to find.

**Theorem
5.7.10**

Suppose that the variables X_1, \dots, X_n form a random sample from the exponential distribution with parameter β . Then the distribution of $Y_1 = \min\{X_1, \dots, X_n\}$ will be the exponential distribution with parameter $n\beta$.

Proof For every number $t > 0$,

$$\begin{aligned}\Pr(Y_1 > t) &= \Pr(X_1 > t, \dots, X_n > t) \\ &= \Pr(X_1 > t) \cdots \Pr(X_n > t) \\ &= e^{-\beta t} \cdots e^{-\beta t} = e^{-n\beta t}.\end{aligned}$$

By comparing this result with Eq. (5.7.19), we see that the distribution of Y_1 must be the exponential distribution with parameter $n\beta$. ■

The memoryless property of the exponential distributions allows us to answer the second question at the end of Example 5.7.7, as well as similar questions about later failures. After one bulb has failed, $n - 1$ bulbs are still burning. Furthermore, regardless of the time at which the first bulb failed or which bulb failed first, it follows from the memoryless property of the exponential distribution that the distribution of the remaining lifetime of each of the other $n - 1$ bulbs is still the exponential distribution with parameter β . In other words, the situation is the same as it would be if we were starting the test over again from time $t = 0$ with $n - 1$ new bulbs. Therefore, Y_2 will be equal to the smallest of $n - 1$ i.i.d. random variables, each of which has the exponential distribution with parameter β . It follows from Theorem 5.7.10 that Y_2 will have the exponential distribution with parameter $(n - 1)\beta$. The next result deals with the remaining waiting times between failures.

**Theorem
5.7.11**

Suppose that the variables X_1, \dots, X_n form a random sample from the exponential distribution with parameter β . Let $Z_1 \leq Z_2 \leq \dots \leq Z_n$ be the random variables X_1, \dots, X_n sorted from smallest to largest. For each $k = 2, \dots, n$, let $Y_k = Z_k - Z_{k-1}$. Then the distribution of Y_k is the exponential distribution with parameter $(n + 1 - k)\beta$.

Proof At the time Z_{k-1} , exactly $k - 1$ of the lifetimes have ended and there are $n + 1 - k$ lifetimes that have not yet ended. For each of the remaining lifetimes, the conditional distribution of what remains of that lifetime given that it has lasted at least Z_{k-1} is still exponential with parameter β by the memoryless property. So, $Y_k = Z_k - Z_{k-1}$ has the same distribution as the minimum lifetime from a random sample of size $n + 1 - k$ from the exponential distribution with parameter β . According to Theorem 5.7.10, that distribution is exponential with parameter $(n + 1 - k)\beta$. ■

Relation to the Poisson Process

**Example
5.7.8**

Radioactive Particles. Suppose that radioactive particles strike a target according to a Poisson process with rate β , as defined in Definition 5.4.2. Let Z_k be the time until the k th particle strikes the target for $k = 1, 2, \dots$. What is the distribution of Z_1 ? What is the distribution of $Y_k = Z_k - Z_{k-1}$ for $k \geq 2$? ◀

Although the random variables defined at the end of Example 5.7.8 look similar to those in Theorem 5.7.11, there are major differences. In Theorem 5.7.11, we were

observing a fixed number n of lifetimes that all started simultaneously. The n lifetimes are all labeled in advance, and each could be observed independently of the others. In Example 5.7.8, there is no fixed number of particles being contemplated, and we have no well-defined notion of when each particle “starts” toward the target. In fact, we cannot even tell which particle is which until after they are observed. We merely start observing at an arbitrary time and record each time a particle hits. Depending on how long we observe the process, we could see an arbitrary number of particles hit the target in Example 5.7.8, but we could never see more than n failures in the setup of Theorem 5.7.11, no matter how long we observe. Theorem 5.7.12 gives the distributions for the times between arrivals in Example 5.7.8, and one can see how the distributions differ from those in Theorem 5.7.11.

Theorem 5.7.12 **Times between Arrivals in a Poisson Process.** Suppose that arrivals occur according to a Poisson process with rate β . Let Z_k be the time until the k th arrival for $k = 1, 2, \dots$. Define $Y_1 = Z_1$ and $Y_k = Z_k - Z_{k-1}$ for $k \geq 2$. Then Y_1, Y_2, \dots are i.i.d. and they each have the exponential distribution with parameter β .

Proof Let $t > 0$, and define X to be the number of arrivals from time 0 until time t . It is easy to see that $Y_1 \leq t$ if and only if $X \geq 1$. That is, the first particle strikes the target by time t if and only if at least one particle strikes the target by time t . We already know that X has the Poisson distribution with mean βt , where β is the rate of the process. So, for $t > 0$,

$$\Pr(Y_1 \leq t) = \Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - e^{-\beta t}.$$

Comparing this to Eq. (5.7.19), we see that $1 - e^{-\beta t}$ is the c.d.f. of the exponential distribution with parameter β .

What happens in a Poisson process after time t is independent of what happens up to time t . Hence, the conditional distribution given $Y_1 = t$ of the gap from time t until the next arrival at Z_2 is the same as the distribution of the time from time 0 until the first arrival. That is, the distribution of $Y_2 = Z_2 - Z_1$ given $Y_1 = t$ (i.e., $Z_1 = t$) is the exponential distribution with parameter β no matter what t is. Hence, Y_2 is independent of Y_1 and they have the same distribution. The same argument can be applied to find the distributions for Y_3, Y_4, \dots ■

An exponential distribution is often used in a practical problem to represent the distribution of the time that elapses before the occurrence of some event. For example, this distribution has been used to represent such periods of time as the period for which a machine or an electronic component will operate without breaking down, the period required to take care of a customer at some service facility, and the period between the arrivals of two successive customers at a facility.

If the events being considered occur in accordance with a Poisson process, then both the waiting time until an event occurs and the period of time between any two successive events will have exponential distributions. This fact provides theoretical support for the use of the exponential distribution in many types of problems.

We can combine Theorem 5.7.12 with Theorem 5.7.7 to obtain the following.

Corollary 5.7.1 **Time until k th Arrival.** In the situation of Theorem 5.7.12, the distribution of Z_k is the gamma distribution with parameters k and β . ■

Summary

The gamma function is defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ and has the property that $\Gamma(n) = (n-1)!$ for $n = 1, 2, \dots$. If X_1, \dots, X_n are independent random variables with gamma distributions all having the same second parameter β , then $\sum_{i=1}^n X_i$ has the gamma distribution with first parameter equal to the sum of the first parameters of X_1, \dots, X_n and second parameter equal to β . The exponential distribution with parameter β is the same as the gamma distribution with parameters 1 and β . Hence, the sum of a random sample of n exponential random variables with parameter β has the gamma distribution with parameters n and β . For a Poisson process with rate β , the times between successive occurrences have the exponential distribution with parameter β , and they are independent. The waiting time until the k th occurrence has the gamma distribution with parameters k and β .

Exercises

1. Suppose that X has the gamma distribution with parameters α and β , and c is a positive constant. Show that cX has the gamma distribution with parameters α and β/c .

2. Compute the quantile function of the exponential distribution with parameter β .

3. Sketch the p.d.f. of the gamma distribution for each of the following pairs of values of the parameters α and β : **(a)** $\alpha = 1/2$ and $\beta = 1$, **(b)** $\alpha = 1$ and $\beta = 1$, **(c)** $\alpha = 2$ and $\beta = 1$.

4. Determine the mode of the gamma distribution with parameters α and β .

5. Sketch the p.d.f. of the exponential distribution for each of the following values of the parameter β : **(a)** $\beta = 1/2$, **(b)** $\beta = 1$, and **(c)** $\beta = 2$.

6. Suppose that X_1, \dots, X_n form a random sample of size n from the exponential distribution with parameter β . Determine the distribution of the sample mean \bar{X}_n .

7. Let X_1, X_2, X_3 be a random sample from the exponential distribution with parameter β . Find the probability that at least one of the random variables is greater than t , where $t > 0$.

8. Suppose that the random variables X_1, \dots, X_k are independent and X_i has the exponential distribution with parameter β_i ($i = 1, \dots, k$). Let $Y = \min\{X_1, \dots, X_k\}$. Show that Y has the exponential distribution with parameter $\beta_1 + \dots + \beta_k$.

9. Suppose that a certain system contains three components that function independently of each other and are connected in series, as defined in Exercise 5 of Sec. 3.7, so that the system fails as soon as one of the components fails. Suppose that the length of life of the first compo-

nent, measured in hours, has the exponential distribution with parameter $\beta = 0.001$, the length of life of the second component has the exponential distribution with parameter $\beta = 0.003$, and the length of life of the third component has the exponential distribution with parameter $\beta = 0.006$. Determine the probability that the system will not fail before 100 hours.

10. Suppose that an electronic system contains n similar components that function independently of each other and that are connected in series so that the system fails as soon as one of the components fails. Suppose also that the length of life of each component, measured in hours, has the exponential distribution with mean μ . Determine the mean and the variance of the length of time until the system fails.

11. Suppose that n items are being tested simultaneously, the items are independent, and the length of life of each item has the exponential distribution with parameter β . Determine the expected length of time until three items have failed. *Hint:* The required value is $E(Y_1 + Y_2 + Y_3)$ in the notation of Theorem 5.7.11.

12. Consider again the electronic system described in Exercise 10, but suppose now that the system will continue to operate until two components have failed. Determine the mean and the variance of the length of time until the system fails.

13. Suppose that a certain examination is to be taken by five students independently of one another, and the number of minutes required by any particular student to complete the examination has the exponential distribution for which the mean is 80. Suppose that the examination begins at 9:00 A.M. Determine the probability that at least one of the students will complete the examination before 9:40 A.M.

14. Suppose again that the examination considered in Exercise 13 is taken by five students, and the first student to complete the examination finishes at 9:25 A.M. Determine the probability that at least one other student will complete the examination before 10:00 A.M.

15. Suppose again that the examination considered in Exercise 13 is taken by five students. Determine the probability that no two students will complete the examination within 10 minutes of each other.

16. It is said that a random variable X has the *Pareto distribution with parameters* x_0 and α ($x_0 > 0$ and $\alpha > 0$) if X has a continuous distribution for which the p.d.f. $f(x|x_0, \alpha)$ is as follows:

$$f(x|x_0, \alpha) = \begin{cases} \frac{\alpha x_0^\alpha}{x^{\alpha+1}} & \text{for } x \geq x_0, \\ 0 & \text{for } x < x_0. \end{cases}$$

Show that if X has this Pareto distribution, then the random variable $\log(X/x_0)$ has the exponential distribution with parameter α .

17. Suppose that a random variable X has the normal distribution with mean μ and variance σ^2 . Determine the value of $E[(X - \mu)^{2n}]$ for $n = 1, 2, \dots$.

18. Consider a random variable X for which $\Pr(X > 0) = 1$, the p.d.f. is f , and the c.d.f. is F . Consider also the function h defined as follows:

$$h(x) = \frac{f(x)}{1 - F(x)} \quad \text{for } x > 0.$$

The function h is called the *failure rate* or the *hazard function* of X . Show that if X has an exponential distribution, then the failure rate $h(x)$ is constant for $x > 0$.

19. It is said that a random variable has the *Weibull distribution with parameters* a and b ($a > 0$ and $b > 0$) if X has a continuous distribution for which the p.d.f. $f(x|a, b)$ is as follows:

$$f(x|a, b) = \begin{cases} \frac{b}{a^b} x^{b-1} e^{-(x/a)^b} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases}$$

Show that if X has this Weibull distribution, then the random variable X^b has the exponential distribution with parameter $\beta = a^{-b}$.

20. It is said that a random variable X has an *increasing failure rate* if the failure rate $h(x)$ defined in Exercise 18 is an increasing function of x for $x > 0$, and it is said that X has a *decreasing failure rate* if $h(x)$ is a decreasing function of x for $x > 0$. Suppose that X has the Weibull distribution with parameters a and b , as defined in Exercise 19. Show

that X has an increasing failure rate if $b > 1$, and X has a decreasing failure rate if $b < 1$.

21. Let X have the gamma distribution with parameters $\alpha > 2$ and $\beta > 0$.

- a. Prove that the mean of $1/X$ is $\beta/(\alpha - 1)$.
- b. Prove that the variance of $1/X$ is $\beta^2/[(\alpha - 1)^2(\alpha - 2)]$.

22. Consider the Poisson process of radioactive particle hits in Example 5.7.8. Suppose that the rate β of the Poisson process is unknown and has the gamma distribution with parameters α and γ . Let X be the number of particles that strike the target during t time units. Prove that the conditional distribution of β given $X = x$ is a gamma distribution, and find the parameters of that gamma distribution.

23. Let F be a continuous c.d.f. satisfying $F(0) = 0$, and suppose that the distribution with c.d.f. F has the memoryless property (5.7.18). Define $\ell(x) = \log[1 - F(x)]$ for $x > 0$.

- a. Show that for all $t, h > 0$,

$$1 - F(h) = \frac{1 - F(t + h)}{1 - F(t)}.$$

- b. Prove that $\ell(t + h) = \ell(t) + \ell(h)$ for all $t, h > 0$.
- c. Prove that for all $t > 0$ and all positive integers k and m , $\ell(kt/m) = (k/m)\ell(t)$.
- d. Prove that for all $t, c > 0$, $\ell(ct) = c\ell(t)$.
- e. Prove that $g(t) = \ell(t)/t$ is constant for $t > 0$.
- f. Prove that F must be the c.d.f. of an exponential distribution.

24. Review the derivation of the Black-Scholes formula (5.6.18). For this exercise, assume that our stock price at time u in the future is $S_0 e^{\mu u + W_u}$, where W_u has the gamma distribution with parameters αu and β with $\beta > 1$. Let r be the risk-free interest rate.

- a. Prove that $e^{-ru} E(S_u) = S_0$ if and only if $\mu = r - \alpha \log(\beta/[\beta - 1])$.
- b. Assume that $\mu = r - \alpha \log(\beta/[\beta - 1])$. Let R be 1 minus the c.d.f. of the gamma distribution with parameters αu and 1. Prove that the risk-neutral price for the option to buy one share of the stock for the price q at time u is $S_0 R(c[\beta - 1]) - q e^{-ru} R(c\beta)$, where

$$c = \log\left(\frac{q}{S_0}\right) + \alpha u \log\left(\frac{\beta}{\beta - 1}\right) - ru.$$

- c. Find the price for the option being considered when $u = 1$, $q = S_0$, $r = 0.06$, $\alpha = 1$, and $\beta = 10$.

5.8 The Beta Distributions

The family of beta distributions is a popular model for random variables that are known to take values in the interval $[0, 1]$. One common example of such a random variable is the unknown proportion of successes in a sequence of Bernoulli trials.

The Beta Function

Example 5.8.1

Defective Parts. A machine produces parts that are either defective or not, as in Example 3.6.9 on page 148. Let P denote the proportion of defectives among all parts that might be produced by this machine. Suppose that we observe n such parts, and let X be the number of defectives among the n parts observed. If we assume that the parts are conditionally independent given P , then we have the same situation as in Example 3.6.9, where we computed the conditional p.d.f. of P given $X = x$ as

$$g_2(p|x) = \frac{p^x(1-p)^{n-x}}{\int_0^1 q^x(1-q)^{n-x}dq}, \quad \text{for } 0 < p < 1. \quad (5.8.1)$$

We are now in a position to calculate the integral in the denominator of Eq. (5.8.1). The distribution with the resulting p.d.f. is a member a useful family that we shall study in this section. ◀

Definition 5.8.1

The Beta Function. For each positive α and β , define

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

The function B is called the *beta function*.

We can show that the beta function B is finite for all $\alpha, \beta > 0$. The proof of the following result relies on the methods from the end of Sec. 3.9 and is given at the end of this section.

Theorem 5.8.1

For all $\alpha, \beta > 0$,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}. \quad (5.8.2)$$

Example 5.8.2

Defective Parts. It follows from Theorem 5.8.1 that the integral in the denominator of Eq. (5.8.1) is

$$\int_0^1 q^x(1-q)^{n-x}dq = \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{x!(n-x)!}{(n+1)!}.$$

The conditional p.d.f. of P given $X = x$ is then

$$g_2(p|x) = \frac{(n+1)!}{x!(n-x)!} p^x(1-p)^{n-x}, \quad \text{for } 0 < p < 1. \quad \blacktriangleleft$$

Definition of the Beta Distributions

The distribution in Example 5.8.2 is a special case of the following.

Definition 5.8.2 Beta Distributions. Let $\alpha, \beta > 0$ and let X be a random variable with p.d.f.

$$f(x|\alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (5.8.3)$$

Then X has the *beta distribution with parameters α and β* .

The conditional distribution of P given $X = x$ in Example 5.8.2 is the beta distribution with parameters $x + 1$ and $n - x + 1$. It can also be seen from Eq. (5.8.3) that the beta distribution with parameters $\alpha = 1$ and $\beta = 1$ is simply the uniform distribution on the interval $[0, 1]$.

Example 5.8.3

Castaneda v. Partida. In Example 5.2.6 on page 278, 220 grand jurors were chosen from a population that is 79.1 percent Mexican American, but only 100 grand jurors were Mexican American. The expected value of a binomial random variable X with parameters 220 and 0.791 is $E(X) = 220 \times 0.791 = 174.02$. This is much larger than the observed value of $X = 100$. Of course, such a discrepancy could occur by chance. After all, there is positive probability of $X = x$ for all $x = 0, \dots, 220$. Let P stand for the proportion of Mexican Americans among all grand jurors that would be chosen under the current system being used. The court assumed that X had the binomial distribution with parameters $n = 220$ and p , conditional on $P = p$. We should then be interested in whether P is substantially less than the value 0.791, which represents impartial juror choice. For example, suppose that we define discrimination to mean that $P \leq 0.8 \times 0.791 = 0.6328$. We would like to compute the conditional probability of $P \leq 0.6328$ given $X = 100$.

Suppose that the distribution of P prior to observing X was the beta distribution with parameters α and β . Then the p.d.f. of P was

$$f_2(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad \text{for } 0 < p < 1.$$

The conditional p.f. of X given $P = p$ is the binomial p.f.

$$g_1(x|p) = \binom{220}{x} p^x (1-p)^{220-x}, \quad \text{for } x = 0, \dots, 220.$$

We can now apply Bayes' theorem for random variables (3.6.13) to obtain the conditional p.d.f. of P given $X = 100$:

$$\begin{aligned} g_2(p|100) &= \frac{\binom{220}{100} p^{100} (1-p)^{120} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}}{f_1(100)} \\ &= \frac{\binom{220}{100} \Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta) f_1(100)} p^{\alpha+100-1} (1-p)^{\beta+120-1}, \end{aligned} \quad (5.8.4)$$

for $0 < p < 1$, where $f_1(100)$ is the marginal p.f. of X at 100. As a function of p the far right side of Eq. (5.8.4) is a constant times $p^{\alpha+100-1} (1-p)^{\beta+120-1}$ for $0 < p < 1$. As such, it is clearly the p.d.f. of a beta distribution. The parameters

of that beta distribution are $\alpha + 100$ and $\beta + 120$. Hence, the constant must be $1/B(100 + \alpha, 120 + \beta)$. That is,

$$g_2(p|100) = \frac{\Gamma(\alpha + \beta + 220)}{\Gamma(\alpha + 100)\Gamma(\beta + 120)} p^{\alpha+100-1} (1-p)^{\beta+120-1}, \quad \text{for } 0 < p < 1. \quad (5.8.5)$$

After choosing values of α and β , we could compute $\Pr(P \leq 0.6328|X = 100)$ and decide how likely it is that there was discrimination. We will see how to choose α and β after we learn how to compute the expected value of a beta random variable. ◀

Note: Conditional Distribution of P after Observing X with Binomial Distribution. The calculation of the conditional distribution of P given $X = 100$ in Example 5.8.3 is a special case of a useful general result. In fact, the proof of the following result is essentially given in Example 5.8.3, and will not be repeated.

Theorem 5.8.2 Suppose that P has the beta distribution with parameters α and β , and the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p . Then the conditional distribution of P given $X = x$ is the beta distribution with parameters $\alpha + x$ and $\beta + n - x$. ■

Moments of Beta Distributions

Theorem 5.8.3 Moments. Suppose that X has the beta distribution with parameters α and β . Then for each positive integer k ,

$$E(X^k) = \frac{\alpha(\alpha + 1) \cdots (\alpha + k - 1)}{(\alpha + \beta)(\alpha + \beta + 1) \cdots (\alpha + \beta + k - 1)}. \quad (5.8.6)$$

In particular,

$$E(X) = \frac{\alpha}{\alpha + \beta},$$

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Proof For $k = 1, 2, \dots$,

$$\begin{aligned} E(X^k) &= \int_0^1 x^k f(x|\alpha, \beta) dx \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1} (1-x)^{\beta-1} dx. \end{aligned}$$

Therefore, by Eq. (5.8.2),

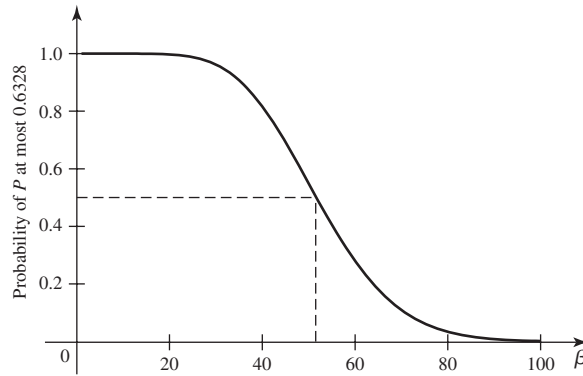
$$E(X^k) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha + k)\Gamma(\beta)}{\Gamma(\alpha + k + \beta)},$$

which simplifies to Eq. (5.8.6). The special case of the mean is simple, while the variance follows easily from

$$E(X^2) = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}. \quad \blacksquare$$

There are too many beta distributions to provide tables in the back of the book. Any good statistical package will be able to calculate the c.d.f.'s of many beta

Figure 5.8 Probability of discrimination as a function of β .



distributions, and some packages will also be able to calculate the quantile functions. The next example illustrates the importance of being able to calculate means and c.d.f.'s of beta distributions.

Example 5.8.4

Castaneda v. Partida. Continuing Example 5.8.3, we are now prepared to see why, for every reasonable choice one makes for α and β , the probability of discrimination in Castaneda v. Partida is quite large. To avoid bias either for or against the defendant, we shall suppose that, before learning X , the probability that a Mexican American juror would be selected on each draw from the pool was 0.791. Let $Y = 1$ if a Mexican American juror is selected on a single draw, and let $Y = 0$ if not. Then Y has the Bernoulli distribution with parameter p given $P = p$ and $E(Y|p) = p$. So the law of total probability for expectations, Theorem 4.7.1, says that

$$\Pr(Y = 1) = E(Y) = E[E(Y|P)] = E(P).$$

This means that we should choose α and β so that $E(P) = 0.791$. Because $E(P) = \alpha/(\alpha + \beta)$, this means that $\alpha = 3.785\beta$. The conditional distribution of P given $X = 100$ is the beta distribution with parameters $\alpha + 100$ and $\beta + 120$. For each value of $\beta > 0$, we can compute $\Pr(P \leq 0.6328|X = 100)$ using $\alpha = 3.785\beta$. Then, for each β we can check whether or not that probability is small. A plot of $\Pr(P \leq 0.6328|X = 100)$ for various values of β is given in Fig. 5.8. From the figure, we see that $\Pr(P \leq 0.6328|X = 100) < 0.5$ only for $\beta \geq 51.5$. This makes $\alpha \geq 194.9$. We claim that the beta distribution with parameters 194.9 and 51.5 as well as all others that make $\Pr(P \leq 0.6328|X = 100) < 0.5$ are unreasonable because they are incredibly prejudiced about the possibility of discrimination. For example, suppose that someone actually believed, before observing $X = 100$, that the distribution of P was the beta distribution with parameters 194.9 and 51.5. For this beta distribution, the probability that there is discrimination would be $\Pr(P \leq 0.6328) = 3.28 \times 10^{-8}$, which is essentially 0. All of the other priors with $\beta \geq 51.5$ and $\alpha = 3.785\beta$ have even smaller probabilities of $\{P \leq 0.6328\}$. Arguing from the other direction, we have the following: Anyone who believed, before observing $X = 100$, that $E(P) = 0.791$ and the probability of discrimination was greater than 3.28×10^{-8} , would believe that the probability of discrimination is at least 0.5 after learning $X = 100$. This is then fairly convincing evidence that there was discrimination in this case. ◀

Example 5.8.5

A Clinical Trial. Consider the clinical trial described in Example 2.1.4. Let P be the proportion of all patients in a large group receiving imipramine who have no relapse (called success). A popular model for P is that P has the beta distribution with

parameters α and β . Choosing α and β can be done based on expert opinion about the chance of success and on the effect that data should have on the distribution of P after observing the data. For example, suppose that the doctors running the clinical trial think that the probability of success should be around $1/3$. Let $X_i = 1$ if the i th patient is a success and $X_i = 0$ if not. We are supposing that $E(X_i|p) = \Pr(X_i = 1|p) = p$, so the law of total probability for expectations (Theorem 4.7.1) says that

$$\Pr(X_i = 1) = E(X_i) = E[E(X_i|P)] = E(P) = \frac{\alpha}{\alpha + \beta}.$$

If we want $\Pr(X_i = 1) = 1/3$, we need $\alpha/(\alpha + \beta) = 1/3$, so $\beta = 2\alpha$. Of course, the doctors will revise the probability of success after observing patients from the study. The doctors can choose α and β based on how that revision will occur.

Assume that the random variables X_1, X_2, \dots (the indicators of success) are conditionally independent given $P = p$. Let $X = X_1 + \dots + X_n$ be the number of patients out of the first n who are successes. The conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p , and the marginal distribution of P is the beta distribution with parameters α and β . Theorem 5.8.2 tells us that the conditional distribution of P given $X = x$ is the beta distribution with parameters $\alpha + x$ and $\beta + n - x$. Suppose that a sequence of 20 patients, all of whom are successes, would raise the doctors' probability of success from $1/3$ up to 0.9 . Then

$$0.9 = E(P|X = 20) = \frac{\alpha + 20}{\alpha + \beta + 20}.$$

This equation implies that $\alpha + 20 = 9\beta$. Combining this with $\beta = 2\alpha$, we get $\alpha = 1.18$ and $\beta = 2.35$.

Finally, we can ask, what will be the distribution of P after observing some patients in the study? Suppose that 40 patients are actually observed, and 22 of them recover (as in Table 2.1). Then the conditional distribution of P given this observation is the beta distribution with parameters $1.18 + 22 = 23.18$ and $2.35 + 18 = 20.35$. It follows that

$$E(P|X = 22) = \frac{23.18}{23.18 + 20.35} = 0.5325.$$

Notice how much closer this is to the proportion of successes (0.55) than was $E(P) = 1/3$. ◀



Proof of Theorem 5.8.1.

Theorem 5.8.1, i.e., Eq. (5.8.2), is part of the following useful result. The proof uses Theorem 3.9.5 (multivariate transformation of random variables). If you did not study Theorem 3.9.5, you will not be able to follow the proof of Theorem 5.8.4.

Theorem 5.8.4

Let U and V be independent random variables with U having the gamma distribution with parameters α and 1 and V having the gamma distribution with parameters β and 1 . Then

- $X = U/(U + V)$ and $Y = U + V$ are independent,
- X has the beta distribution with parameters α and β , and
- Y has the gamma distribution with parameters $\alpha + \beta$ and 1 .

Also, Eq. (5.8.2) holds.

Proof Because U and V are independent, the joint p.d.f. of U and V is the product of their marginal p.d.f.'s, which are

$$f_1(u) = \frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}, \text{ for } u > 0,$$

$$f_2(v) = \frac{v^{\beta-1}e^{-v}}{\Gamma(\beta)}, \text{ for } v > 0.$$

So, the joint p.d.f. is

$$f(u, v) = \frac{u^{\alpha-1}v^{\beta-1}e^{-(u+v)}}{\Gamma(\alpha)\Gamma(\beta)},$$

for $u > 0$ and $v > 0$.

The transformation from (u, v) to (x, y) is

$$x = r_1(u, v) = \frac{u}{u+v} \quad \text{and} \quad y = r_2(u, v) = u+v,$$

and the inverse is

$$u = s_1(x, y) = xy \quad \text{and} \quad v = s_2(x, y) = (1-x)y.$$

The Jacobian is the determinant of the matrix

$$J = \begin{bmatrix} y & x \\ -y & 1-x \end{bmatrix},$$

which equals y . According to Theorem 3.9.5, the joint p.d.f. of (X, Y) is then

$$\begin{aligned} g(x, y) &= f(s_1(x, y), s_2(x, y))y \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}y^{\alpha+\beta-1}e^{-y}}{\Gamma(\alpha)\Gamma(\beta)}, \end{aligned} \quad (5.8.7)$$

for $0 < x < 1$ and $y > 0$. Notice that this joint p.d.f. factors into separate functions of x and y , and hence X and Y are independent. The marginal distribution of Y is available from Theorem 5.7.7. The marginal p.d.f. of X is obtained by integrating y out of (5.8.7):

$$\begin{aligned} g_1(x) &= \int_0^\infty \frac{x^{\alpha-1}(1-x)^{\beta-1}y^{\alpha+\beta-1}e^{-y}}{\Gamma(\alpha)\Gamma(\beta)} dy \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty y^{\alpha+\beta-1}e^{-y} dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \end{aligned} \quad (5.8.8)$$

where the last equation follows from (5.7.2). Because the far right side of (5.8.8) is a p.d.f., it integrates to 1, which proves Eq. (5.8.2). Also, one can recognize the far right side of (5.8.8) as the p.d.f. of the beta distribution with parameters α and β . ■



Summary

The family of beta distributions is a popular model for random variables that lie in the interval $(0, 1)$, such as unknown proportions of success for sequences of Bernoulli trials. The mean of the beta distribution with parameters α and β is $\alpha/(\alpha + \beta)$. If X

has the binomial distribution with parameters n and p conditional on $P = p$, and if P has the beta distribution with parameters α and β , then, conditional on $X = x$, the distribution of P is the beta distribution with parameters $\alpha + x$ and $\beta + n - x$.

Exercises

1. Compute the quantile function of the beta distribution with parameters $\alpha > 0$ and $\beta = 1$.
2. Determine the mode of the beta distribution with parameters α and β , assuming that $\alpha > 1$ and $\beta > 1$.
3. Sketch the p.d.f. of the beta distribution for each of the following pairs of values of the parameters:
 - a. $\alpha = 1/2$ and $\beta = 1/2$
 - b. $\alpha = 1/2$ and $\beta = 1$
 - c. $\alpha = 1/2$ and $\beta = 2$
 - d. $\alpha = 1$ and $\beta = 1$
 - e. $\alpha = 1$ and $\beta = 2$
 - f. $\alpha = 2$ and $\beta = 2$
 - g. $\alpha = 25$ and $\beta = 100$
 - h. $\alpha = 100$ and $\beta = 25$
4. Suppose that X has the beta distribution with parameters α and β . Show that $1 - X$ has the beta distribution with parameters β and α .
5. Suppose that X has the beta distribution with parameters α and β , and let r and s be given positive integers. Determine the value of $E[X^r(1 - X)^s]$.
6. Suppose that X and Y are independent random variables, X has the gamma distribution with parameters α_1 and β , and Y has the gamma distribution with parameters α_2 and β . Let $U = X/(X + Y)$ and $V = X + Y$. Show that
 - (a) U has the beta distribution with parameters α_1 and α_2 ,
 - and (b) U and V are independent. *Hint:* Look at the steps in the proof of Theorem 5.8.1.
7. Suppose that X_1 and X_2 form a random sample of two observed values from the exponential distribution with parameter β . Show that $X_1/(X_1 + X_2)$ has the uniform distribution on the interval $[0, 1]$.
8. Suppose that the proportion X of defective items in a large lot is unknown and that X has the beta distribution with parameters α and β .
 - a. If one item is selected at random from the lot, what is the probability that it will be defective?
 - b. If two items are selected at random from the lot, what is the probability that both will be defective?
9. A manufacturer believes that an unknown proportion P of parts produced will be defective. She models P as having a beta distribution. The manufacturer thinks that P should be around 0.05, but if the first 10 observed products were all defective, the mean of P would rise from 0.05 to 0.9. Find the beta distribution that has these properties.
10. A marketer is interested in how many customers are likely to buy a particular product in a particular store. Let P be the proportion of all customers in the store who will buy the product. Let the distribution of P be uniform on the interval $[0, 1]$ before observing any data. The marketer then observes 25 customers and only six buy the product. If the customers were conditionally independent given P , find the conditional distribution of P given the observed customers.

5.9 The Multinomial Distributions

Many times we observe data that can assume three or more possible values. The family of multinomial distributions is an extension of the family of binomial distributions to handle these cases. The multinomial distributions are multivariate distributions.

Definition and Derivation of Multinomial Distributions

Example 5.9.1

Blood Types. In Example 1.8.4 on page 34, we discussed human blood types, of which there are four: O, A, B, and AB. If a number of people are chosen at random, we might be interested in the probability of obtaining certain numbers of each blood type. Such calculations are used in the courts during paternity suits. ◀

In general, suppose that a population contains items of k different types ($k \geq 2$) and that the proportion of the items in the population that are of type i is p_i

($i = 1, \dots, k$). It is assumed that $p_i > 0$ for $i = 1, \dots, k$, and $\sum_{i=1}^k p_i = 1$. Let $\mathbf{p} = (p_1, \dots, p_k)$ denote the vector of these probabilities.

Next, suppose that n items are selected at random from the population, with replacement, and let X_i denote the number of selected items that are of type i ($i = 1, \dots, k$). Because the n items are selected from the population at random with replacement, the selections will be independent of each other. Hence, the probability that the first item will be of type i_1 , the second item of type i_2 , and so on, is simply $p_{i_1} p_{i_2} \dots p_{i_n}$. Therefore, the probability that the sequence of n outcomes will consist of exactly x_1 items of type 1, x_2 items of type 2, and so on, selected in a *particular prespecified order*, is $p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$. It follows that the probability of obtaining exactly x_i items of type i ($i = 1, \dots, k$) is equal to the probability $p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$ multiplied by the total number of different ways in which the order of the n items can be specified.

From the discussion that led to the definition of multinomial coefficients (Definition 1.9.1), it follows that the total number of different ways in which n items can be arranged when there are x_i items of type i ($i = 1, \dots, k$) is given by the multinomial coefficient

$$\binom{n}{x_1, \dots, x_k} = \frac{n!}{x_1! x_2! \dots x_k!}.$$

In the notation of multivariate distributions, let $\mathbf{X} = (X_1, \dots, X_k)$ denote the random vector of counts, and let $\mathbf{x} = (x_1, \dots, x_k)$ denote a possible value for that random vector. Finally, let $f(\mathbf{x}|n, \mathbf{p})$ denote the joint p.f. of \mathbf{X} . Then

$$\begin{aligned} f(\mathbf{x}|n, \mathbf{p}) &= \Pr(\mathbf{X} = \mathbf{x}) = \Pr(X_1 = x_1, \dots, X_k = x_k) \\ &= \begin{cases} \binom{n}{x_1, \dots, x_k} p_1^{x_1} \dots p_k^{x_k} & \text{if } x_1 + \dots + x_k = n, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (5.9.1)$$

Definition 5.9.1 Multinomial Distributions. A discrete random vector $\mathbf{X} = (X_1, \dots, X_k)$ whose p.f. is given by Eq. (5.9.1) has the *multinomial distribution with parameters n and $\mathbf{p} = (p_1, \dots, p_k)$* .

Example 5.9.2 Attendance at a Baseball Game. Suppose that 23 percent of the people attending a certain baseball game live within 10 miles of the stadium, 59 percent live between 10 and 50 miles from the stadium, and 18 percent live more than 50 miles from the stadium. Suppose also that 20 people are selected at random from the crowd attending the game. We shall determine the probability that seven of the people selected live within 10 miles of the stadium, eight of them live between 10 and 50 miles from the stadium, and five of them live more than 50 miles from the stadium.

We shall assume that the crowd attending the game is so large that it is irrelevant whether the 20 people are selected with or without replacement. We can therefore assume that they were selected with replacement. It then follows from Eq. (5.9.1) that the required probability is

$$\frac{20!}{7! 8! 5!} (0.23)^7 (0.59)^8 (0.18)^5 = 0.0094. \quad \blacktriangleleft$$

Example 5.9.3 Blood Types. Berry and Geisser (1986) estimate the probabilities of the four blood types in Table 5.3 based on a sample of 6004 white Californians that was analyzed by Grunbaum et al. (1978). Suppose that we will select two people at random from this population and observe their blood types. What is the probability that they will both have the same blood type? The event that the two people have the same blood type is the union of four disjoint events, each of which is the event that the two people

Table 5.3 Estimated probabilities of blood types for white Californians

A	B	AB	O
0.360	0.123	0.038	0.479

both have one of the four different blood types. Each of these events has probability $\binom{2}{2,0,0,0}$ times the square of one of the four probabilities. The probability that we want is the sum of the probabilities of the four events:

$$\binom{2}{2,0,0,0}(0.360^2 + 0.123^2 + 0.038^2 + 0.479^2) = 0.376. \quad \blacktriangleleft$$

Relation between the Multinomial and Binomial Distributions

When the population being sampled contains only two different types of items, that is, when $k = 2$, each multinomial distribution reduces to essentially a binomial distribution. The precise form of this relationship is as follows.

Theorem 5.9.1 Suppose that the random vector $\mathbf{X} = (X_1, X_2)$ has the multinomial distribution with parameters n and $\mathbf{p} = (p_1, p_2)$. Then X_1 has the binomial distribution with parameters n and p_1 , and $X_2 = n - X_1$.

Proof It is clear from the definition of multinomial distributions that $\mathbf{X}_2 = n - X_1$ and $p_2 = 1 - p_1$. Therefore, the random vector \mathbf{X} is actually determined by the single random variable X_1 . From the derivation of the multinomial distribution, we see that X_1 is the number of items of type 1 that are selected if n items are selected from a population consisting of two types of items. If we call items of type 1 “success,” then X_1 is the number of successes in n Bernoulli trials with probability of success on each trial equal to p_1 . It follows that X_1 has the binomial distribution with parameters n and p_1 . ■

The proof of Theorem 5.9.1 extends easily to the following result.

Corollary 5.9.1 Suppose that the random vector $\mathbf{X} = (X_1, \dots, X_k)$ has the multinomial distribution with parameters n and $\mathbf{p} = (p_1, \dots, p_k)$. The marginal distribution of each variable X_i ($i = 1, \dots, k$) is the binomial distribution with parameters n and p_i .

Proof Choose one i from $1, \dots, k$, and define success to be the selection of an item of type i . Then X_i is the number of successes in n Bernoulli trials with probability of success on each trial equal to p_i . ■

A further generalization of Corollary 5.9.1 is that the marginal distribution of the sum of some of the coordinates of a multinomial vector has a binomial distribution. The proof is left to Exercise 1 in this section.

Corollary 5.9.2 Suppose that the random vector $\mathbf{X} = (X_1, \dots, X_k)$ has the multinomial distribution with parameters n and $\mathbf{p} = (p_1, \dots, p_k)$ with $k > 2$. Let $\ell < k$, and let i_1, \dots, i_ℓ be distinct elements of the set $\{1, \dots, k\}$. The distribution of $Y = X_{i_1} + \dots + X_{i_\ell}$ is the binomial distribution with parameters n and $p_{i_1} + \dots + p_{i_\ell}$. ■

As a final note, the relationship between Bernoulli and binomial distributions extends to multinomial distributions. The Bernoulli distribution with parameter p is the same as the binomial distribution with parameters 1 and p . However, there is no separate name for a multinomial distribution with first parameter $n = 1$. A random vector with such a distribution will consist of a single 1 in one of its coordinates and $k - 1$ zeros in the other coordinates. The probability is p_i that the i th coordinate is the 1. A k -dimensional vector seems an unwieldy way to represent a random object that can take only k different values. A more common representation would be as a single discrete random variable X that takes one of the k values $1, \dots, k$ with probabilities p_1, \dots, p_k , respectively. The univariate distribution just described has no famous name associated with it; however, we have just shown that it is closely related to the multinomial distribution with parameters 1 and (p_1, \dots, p_k) .

Means, Variances, and Covariances

The means, variances, and covariances of the coordinates of a multinomial random vector are given by the next result.

Theorem **5.9.2**

Means, Variances, and Covariances. Let the random vector \mathbf{X} have the multinomial distribution with parameters n and \mathbf{p} . The means and variances of the coordinates of \mathbf{X} are

$$E(X_i) = np_i \quad \text{and} \quad \text{Var}(X_i) = np_i(1 - p_i) \quad \text{for } i = 1, \dots, k. \quad (5.9.2)$$

Also, the covariances between the coordinates are

$$\text{Cov}(X_i, X_j) = -np_i p_j. \quad (5.9.3)$$

Proof Corollary 5.9.1 says that the marginal distribution of each component X_i is the binomial distribution with parameters n and p_i . Eq. 5.9.2 follows directly from this fact.

Corollary 5.9.2 says that $X_i + X_j$ has the binomial distribution with parameters n and $p_i + p_j$. Hence,

$$\text{Var}(X_i + X_j) = n(p_i + p_j)(1 - p_i - p_j). \quad (5.9.4)$$

According to Theorem 4.6.6, it is also true that

$$\begin{aligned} \text{Var}(X_i + X_j) &= \text{Var}(X_i) + \text{Var}(X_j) + 2 \text{Cov}(X_i, X_j) \\ &= np_i(1 - p_i) + np_j(1 - p_j) + 2 \text{Cov}(X_i, X_j). \end{aligned} \quad (5.9.5)$$

Equate the right sides of (5.9.4) and (5.9.5), and solve for $\text{Cov}(X_i, X_j)$. The result is (5.9.3). ■

Note: Negative Covariance Is Natural for Multinomial Distributions. The negative covariance between different coordinates of a multinomial vector is natural since there are only n selections to be distributed among the k coordinates of the vector. If one of the coordinates is large, at least some of the others have to be small because the sum of the coordinates is fixed at n .

Summary

Multinomial distributions extend binomial distributions to counts of more than two possible outcomes. The i th coordinate of a vector having the multinomial distribution

with parameters n and $\mathbf{p} = (p_1, \dots, p_k)$ has the binomial distribution with parameters n and p_i for $i = 1, \dots, k$. Hence, the means and variances of the coordinates of a multinomial vector are the same as those of a binomial random variable. The covariance between the i th and j th coordinates is $-np_i p_j$.

Exercises

1. Prove Corollary 5.9.2.

2. Suppose that F is a continuous c.d.f. on the real line, and let α_1 and α_2 be numbers such that $F(\alpha_1) = 0.3$ and $F(\alpha_2) = 0.8$. If 25 observations are selected at random from the distribution for which the c.d.f. is F , what is the probability that six of the observed values will be less than α_1 , 10 of the observed values will be between α_1 and α_2 , and nine of the observed values will be greater than α_2 ?

3. If five balanced dice are rolled, what is the probability that the number 1 and the number 4 will appear the same number of times?

4. Suppose that a die is loaded so that each of the numbers 1, 2, 3, 4, 5, and 6 has a different probability of appearing when the die is rolled. For $i = 1, \dots, 6$, let p_i denote the probability that the number i will be obtained, and suppose that $p_1 = 0.11$, $p_2 = 0.30$, $p_3 = 0.22$, $p_4 = 0.05$, $p_5 = 0.25$, and $p_6 = 0.07$. Suppose also that the die is to be rolled 40 times. Let X_1 denote the number of rolls for which an even number appears, and let X_2 denote the number of rolls for which either the number 1 or the number 3 appears. Find the value of $\Pr(X_1 = 20 \text{ and } X_2 = 15)$.

5. Suppose that 16 percent of the students in a certain high school are freshmen, 14 percent are sophomores, 38 percent are juniors, and 32 percent are seniors. If 15 students are selected at random from the school, what is the probability that at least eight will be either freshmen or sophomores?

6. In Exercise 5, let X_3 denote the number of juniors in the random sample of 15 students, and let X_4 denote the number of seniors in the sample. Find the value of $E(X_3 - X_4)$ and the value of $\text{Var}(X_3 - X_4)$.

7. Suppose that the random variables X_1, \dots, X_k are independent and that X_i has the Poisson distribution with mean λ_i ($i = 1, \dots, k$). Show that for each fixed positive integer n , the conditional distribution of the random vector $\mathbf{X} = (X_1, \dots, X_k)$, given that $\sum_{i=1}^k X_i = n$, is the multinomial distribution with parameters n and $\mathbf{p} = (p_1, \dots, p_k)$, where

$$p_i = \frac{\lambda_i}{\sum_{j=1}^k \lambda_j} \quad \text{for } i = 1, \dots, k.$$

8. Suppose that the parts produced by a machine can have three different levels of functionality: working, impaired, defective. Let p_1 , p_2 , and $p_3 = 1 - p_1 - p_2$ be the probabilities that a part is working, impaired, and defective, respectively. Suppose that the vector $\mathbf{p} = (p_1, p_2)$ is unknown but has a joint distribution with p.d.f.

$$f(p_1, p_2) = \begin{cases} 12p_1^2 & \text{for } 0 < p_1, p_2 < 1 \\ & \text{and } p_1 + p_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we observe 10 parts that are conditionally independent given \mathbf{p} , and among those 10 parts, eight are working and two are impaired. Find the conditional p.d.f. of \mathbf{p} given the observed parts. *Hint:* You might find Eq. (5.8.2) helpful.

5.10 The Bivariate Normal Distributions

The first family of multivariate continuous distributions for which we have a name is a generalization of the family of normal distributions to two coordinates. There is more structure to a bivariate normal distribution than just a pair of normal marginal distributions.

Definition and Derivation of Bivariate Normal Distributions

Example 5.10.1

Thyroid Hormones. Production of rocket fuel produces a chemical, perchlorate, that has found its way into drinking water supplies. Perchlorate is suspected of inhibiting thyroid function. Experiments have been performed in which laboratory rats have

been dosed with perchlorate in their drinking water. After several weeks, rats were sacrificed, and a number of thyroid hormones were measured. The levels of these hormones were then compared to the levels of the same hormones in rats that received no perchlorate in their water. Two hormones, TSH and T4, were of particular interest. Experimenters were interested in the joint distribution of TSH and T4. Although each of the hormones might be modeled with a normal distribution, a bivariate distribution is needed in order to model the two hormone levels jointly. Knowledge of thyroid activity suggests that the levels of these hormones will not be independent, because one of them is actually used by the thyroid to stimulate production of the other. ◀

If researchers are comfortable using the family of normal distributions to model each of two random variables separately, such as the hormones in Example 5.10.1, then they need a bivariate generalization of the family of normal distributions that still has normal distributions for its marginals while allowing the two random variables to be dependent. A simple way to create such a generalization is to make use of the result in Corollary 5.6.1. That result says that a linear combination of independent normal random variables has a normal distribution. If we create two different linear combinations X_1 and X_2 of the same independent normal random variables, then X_1 and X_2 will each have a normal distribution and they might be dependent. The following result formalizes this idea.

Theorem
5.10.1

Suppose that Z_1 and Z_2 are independent random variables, each of which has the standard normal distribution. Let $\mu_1, \mu_2, \sigma_1, \sigma_2$, and ρ be constants such that $-\infty < \mu_i < \infty$ ($i = 1, 2$), $\sigma_i > 0$ ($i = 1, 2$), and $-1 < \rho < 1$. Define two new random variables X_1 and X_2 as follows:

$$\begin{aligned} X_1 &= \sigma_1 Z_1 + \mu_1, \\ X_2 &= \sigma_2 \left[\rho Z_1 + (1 - \rho^2)^{1/2} Z_2 \right] + \mu_2. \end{aligned} \quad (5.10.1)$$

The joint p.d.f. of X_1 and X_2 is

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi(1 - \rho^2)^{1/2}\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right. \right. \\ &\quad \left. \left. - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}. \end{aligned} \quad (5.10.2)$$

Proof This proof relies on Theorem 3.9.5 (multivariate transformation of random variables). If you did not study Theorem 3.9.5, you won't be able to follow this proof. The joint p.d.f. $g(z_1, z_2)$ of Z_1 and Z_2 is

$$g(z_1, z_2) = \frac{1}{2\pi} \exp \left[-\frac{1}{2}(z_1^2 + z_2^2) \right], \quad (5.10.3)$$

for all z_1 and z_2 .

The inverse of the transformation (5.10.1) is $(Z_1, Z_2) = (s_1(X_1, X_2), s_2(X_1, X_2))$, where

$$\begin{aligned} s_1(x_1, x_2) &= \frac{x_1 - \mu_1}{\sigma_1}, \\ s_2(x_1, x_2) &= \frac{1}{(1 - \rho^2)^{1/2}} \left(\frac{x_2 - \mu_2}{\sigma_2} - \rho \frac{x_1 - \mu_1}{\sigma_1} \right). \end{aligned} \quad (5.10.4)$$

The Jacobian J of the transformation is

$$J = \det \begin{bmatrix} \frac{1}{\sigma_1} & 0 \\ -\rho & 1 \end{bmatrix} = \frac{1}{(1 - \rho^2)^{1/2} \sigma_1 \sigma_2}. \quad (5.10.5)$$

If one substitutes $s_i(x_1, x_2)$ for z_i ($i = 1, 2$) in Eq. (5.10.3) and then multiplies by $|J|$, one obtains Eq. (5.10.2), which is the joint p.d.f. of (X_1, X_2) according to Theorem 3.9.5. ■

Some simple properties of the distribution with p.d.f. in Eq. (5.10.2) are worth deriving before giving a name to the joint distribution.

**Theorem
5.10.2**

Suppose that X_1 and X_2 have the joint distribution whose p.d.f. is given by Eq. (5.10.2). Then there exist independent standard normal random variables Z_1 and Z_2 such that Eqs. (5.10.1) hold. Also, the mean of X_i is μ_i and the variance of X_i is σ_i^2 for $i = 1, 2$. Furthermore the correlation between X_1 and X_2 is ρ . Finally, the marginal distribution of X_i is the normal distribution with mean μ_i and variance σ_i^2 for $i = 1, 2$.

Proof Use the functions s_1 and s_2 defined in Eqs. (5.10.4) and define $Z_i = s_i(X_1, X_2)$ for $i = 1, 2$. By running the proof of Theorem 5.10.1 in reverse, we see that the joint p.d.f. of Z_1 and Z_2 is Eq. (5.10.3). Hence, Z_1 and Z_2 are independent standard normal random variables.

The values of the means and variances of X_1 and X_2 are easily obtained by applying Corollary 5.6.1 to Eq. (5.10.1). If one applies the result in Exercise 8 of Sec. 4.6, one obtains $\text{Cov}(X_1, X_2) = \sigma_1 \sigma_2 \rho$. It now follows that ρ is the correlation. The claim about the marginal distributions of X_1 and X_2 is immediate from Corollary 5.6.1. ■

We are now ready to define the family of bivariate normal distributions.

**Definition
5.10.1**

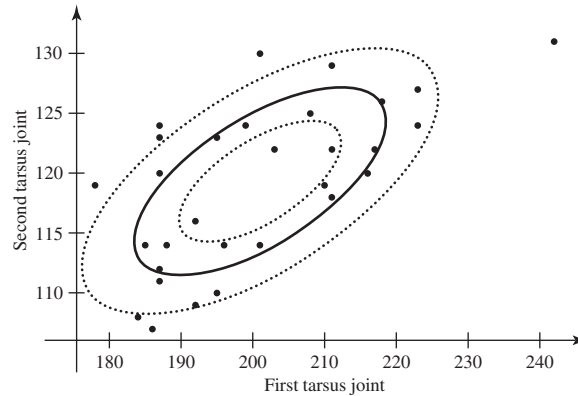
Bivariate Normal Distributions. When the joint p.d.f. of two random variables X_1 and X_2 is of the form in Eq. (5.10.2), it is said that X_1 and X_2 have the *bivariate normal distribution with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 , and correlation ρ* .

It was convenient for us to derive the bivariate normal distributions as the joint distributions of certain linear combinations of independent random variables having standard normal distributions. It should be emphasized, however, that bivariate normal distributions arise directly and naturally in many practical problems. For example, for many populations the joint distribution of two physical characteristics such as the heights and the weights of the individuals in the population will be approximately a bivariate normal distribution. For other populations, the joint distribution of the scores of the individuals in the population on two related tests will be approximately a bivariate normal distribution.

**Example
5.10.2**

Anthropometry of Flea Beetles. Lubischew (1962) reports the measurements of several physical features of a variety of species of flea beetle. The investigation was concerned with whether some combination of easily obtained measurements could be used to distinguish the different species. Figure 5.9 shows a scatterplot of measurements of the first joint in the first tarsus versus the second joint in the first tarsus for a sample of 31 from the species *Chaetocnema heikertingeri*. The plot also includes three ellipses that correspond to a fitted bivariate normal distribution. The ellipses were chosen to contain 25%, 50%, and 75% of the probability of the fitted bivariate normal

Figure 5.9 Scatterplot of flea beetle data with 25%, 50%, and 75% bivariate normal ellipses for Example 5.10.2.



distribution. The fitted distribution is the bivariate normal distribution with means 201 and 119.3, variances 222.1 and 44.2, and correlation 0.64. ◀

Properties of Bivariate Normal Distributions

For random variables with a bivariate normal distribution, we find that being independent is equivalent to being uncorrelated.

Theorem 5.10.3

Independence and Correlation. Two random variables X_1 and X_2 that have a bivariate normal distribution are independent if and only if they are uncorrelated.

Proof The “only if” direction is already known from Theorem 4.6.4. For the “if” direction, assume that X_1 and X_2 are uncorrelated. Then $\rho = 0$, and it can be seen from Eq. (5.10.2) that the joint p.d.f. $f(x_1, x_2)$ factors into the product of the marginal p.d.f. of X_1 and the marginal p.d.f. of X_2 . Hence, X_1 and X_2 are independent. ■

We have already seen in Example 4.6.4 that two random variables X_1 and X_2 with an arbitrary joint distribution can be uncorrelated without being independent. Theorem 5.10.3 says that no such examples exist in which X_1 and X_2 have a bivariate normal distribution.

When the correlation is not zero, Theorem 5.10.2 gives the marginal distributions of bivariate normal random variables. Combining the marginal and joint distributions allows us to find the conditional distributions of each X_i given the other one. The next theorem derives the conditional distributions using another technique.

Theorem 5.10.4

Conditional Distributions. Let X_1 and X_2 have the bivariate normal distribution whose p.d.f. is Eq. (5.10.2). The conditional distribution of X_2 given that $X_1 = x_1$ is the normal distribution with mean and variance given by

$$E(X_2|x_1) = \mu_2 + \rho\sigma_2 \left(\frac{x_1 - \mu_1}{\sigma_1} \right), \quad \text{Var}(X_2|x_1) = (1 - \rho^2)\sigma_2^2. \quad (5.10.6)$$

Proof We will make liberal use of Theorem 5.10.2 and its notation in this proof. Conditioning on $X_1 = x_1$ is the same as conditioning on $Z_1 = (x_1 - \mu_1)/\sigma_1$. When we want to find the conditional distribution of X_2 given $Z_1 = (x_1 - \mu_1)/\sigma_1$, we can substitute $(x_1 - \mu_1)/\sigma_1$ for Z_2 in the formula for X_2 in Eq. (5.10.1) and find the conditional distribution for the rest of the formula. That is, the conditional distribution of X_2 given

that $X_1 = x_1$ is the same as the conditional distribution of

$$(1 - \rho^2)^{1/2} \sigma_2 Z_2 + \mu_2 + \rho \sigma_2 \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \quad (5.10.7)$$

given $Z_1 = (x_1 - \mu_1)/\sigma_1$. But Z_2 is the only random variable in Eq. (5.10.7), and Z_2 is independent of Z_1 . Hence, the conditional distribution of X_2 given $X_1 = x_1$ is the marginal distribution of Eq. (5.10.7), namely, the normal distribution with mean and variance given by Eq. (5.10.6). ■

The conditional distribution of X_1 given that $X_2 = x_2$ cannot be derived so easily from Eq. (5.10.1) because of the different ways in which Z_1 and Z_2 enter Eq. (5.10.1). However, it is seen from Eq. (5.10.2) that the joint distribution of X_2 and X_1 is also bivariate normal with all of the subscripts 1 and 2 switched on all of the parameters. Hence, we can apply Theorem 5.10.4 to X_2 and X_1 to conclude that the conditional distribution of X_1 given that $X_2 = x_2$ must be the normal distribution with mean and variance

$$E(X_1|x_2) = \mu_1 + \rho \sigma_1 \left(\frac{x_2 - \mu_2}{\sigma_2} \right), \quad \text{Var}(X_1|x_2) = (1 - \rho^2) \sigma_1^2. \quad (5.10.8)$$

We have now shown that each marginal distribution and each conditional distribution of a bivariate normal distribution is a univariate normal distribution.

Some particular features of the conditional distribution of X_2 given that $X_1 = x_1$ should be noted. If $\rho \neq 0$, then $E(X_2|x_1)$ is a linear function of x_1 . If $\rho > 0$, the slope of this linear function is positive. If $\rho < 0$, the slope of the function is negative. However, the variance of the conditional distribution of X_2 given that $X_1 = x_1$ is $(1 - \rho^2) \sigma_2^2$, which does not depend on x_1 . Furthermore, this variance of the conditional distribution of X_2 is smaller than the variance σ_2^2 of the marginal distribution of X_2 .

Example 5.10.3

Predicting a Person's Weight. Let X_1 denote the height of a person selected at random from a certain population, and let X_2 denote the weight of the person. Suppose that these random variables have the bivariate normal distribution for which the p.d.f. is specified by Eq. (5.10.2) and that the person's weight X_2 must be predicted. We shall compare the smallest M.S.E. that can be attained if the person's height X_1 is known when her weight must be predicted with the smallest M.S.E. that can be attained if her height is not known.

If the person's height is not known, then the best prediction of her weight is the mean $E(X_2) = \mu_2$, and the M.S.E. of this prediction is the variance σ_2^2 . If it is known that the person's height is x_1 , then the best prediction is the mean $E(X_2|x_1)$ of the conditional distribution of X_2 given that $X_1 = x_1$, and the M.S.E. of this prediction is the variance $(1 - \rho^2) \sigma_2^2$ of that conditional distribution. Hence, when the value of X_1 is known, the M.S.E. is reduced from σ_2^2 to $(1 - \rho^2) \sigma_2^2$. ◀

Since the variance of the conditional distribution in Example 5.10.3 is $(1 - \rho^2) \sigma_2^2$, regardless of the known height x_1 of the person, it follows that the difficulty of predicting the person's weight is the same for a tall person, a short person, or a person of medium height. Furthermore, since the variance $(1 - \rho^2) \sigma_2^2$ decreases as $|\rho|$ increases, it follows that it is easier to predict a person's weight from her height when the person is selected from a population in which height and weight are highly correlated.

**Example
5.10.4**

Determining a Marginal Distribution. Suppose that a random variable X has the normal distribution with mean μ and variance σ^2 , and that for every number x , the conditional distribution of another random variable Y given that $X = x$ is the normal distribution with mean x and variance τ^2 . We shall determine the marginal distribution of Y .

We know that the marginal distribution of X is a normal distribution, and the conditional distribution of Y given that $X = x$ is a normal distribution, for which the mean is a linear function of x and the variance is constant. It follows that the joint distribution of X and Y must be a bivariate normal distribution (see Exercise 14). Hence, the marginal distribution of Y is also a normal distribution. The mean and the variance of Y must be determined.

The mean of Y is

$$E(Y) = E[E(Y|X)] = E(X) = \mu.$$

Furthermore, by Theorem 4.7.4,

$$\begin{aligned}\text{Var}(Y) &= E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)] \\ &= E(\tau^2) + \text{Var}(X) \\ &= \tau^2 + \sigma^2.\end{aligned}$$

Hence, the distribution of Y is the normal distribution with mean μ and variance $\tau^2 + \sigma^2$. ◀

Linear Combinations

**Example
5.10.5**

Heights of Husbands and Wives. Suppose that a married couple is selected at random from a certain population of married couples and that the joint distribution of the height of the wife and the height of her husband is a bivariate normal distribution. What is the probability that, in the randomly chosen couple, the wife is taller than the husband? ◀

The question asked at the end of Example 5.10.5 can be expressed in terms of the distribution of the difference between a wife's and husband's heights. This is a special case of a linear combination of a bivariate normal vector.

**Theorem
5.10.5**

Linear Combination of Bivariate Normals. Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, for which the p.d.f. is specified by Eq. (5.10.2). Let $Y = a_1X_1 + a_2X_2 + b$, where a_1 , a_2 , and b are arbitrary given constants. Then Y has the normal distribution with mean $a_1\mu_1 + a_2\mu_2 + b$ and variance

$$a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\rho\sigma_1\sigma_2. \quad (5.10.9)$$

Proof According to Theorem 5.10.2, both X_1 and X_2 can be represented, as in Eq. (5.10.1), as linear combinations of independent and normally distributed random variables Z_1 and Z_2 . Since Y is a linear combination of X_1 and X_2 , it follows that Y can also be represented as a linear combination of Z_1 and Z_2 . Therefore, by Corollary 5.6.1, the distribution of Y will also be a normal distribution. It only remains to compute the mean and variance of Y . The mean of Y is

$$\begin{aligned}E(Y) &= a_1E(X_1) + a_2E(X_2) + b \\ &= a_1\mu_1 + a_2\mu_2 + b.\end{aligned}$$

It also follows from Corollary 4.6.1 that

$$\text{Var}(Y) = a_1^2 \text{Var}(X_1) + a_2^2 \text{Var}(X_2) + 2a_1a_2 \text{Cov}(X_1, X_2).$$

That $\text{Var}(Y)$ is given by Eq. (5.10.9) now follows easily. ■

**Example
5.10.6**

Heights of Husbands and Wives. Consider again Example 5.10.5. Suppose that the heights of the wives have a mean of 66.8 inches and a standard deviation of 2 inches, the heights of the husbands have a mean of 70 inches and a standard deviation of 2 inches, and the correlation between these two heights is 0.68. We shall determine the probability that the wife will be taller than her husband.

If we let X denote the height of the wife, and let Y denote the height of her husband, then we must determine the value of $\Pr(X - Y > 0)$. Since X and Y have a bivariate normal distribution, it follows that the distribution of $X - Y$ will be the normal distribution, with mean

$$E(X - Y) = 66.8 - 70 = -3.2$$

and variance

$$\begin{aligned} \text{Var}(X - Y) &= \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y) \\ &= 4 + 4 - 2(0.68)(2)(2) = 2.56. \end{aligned}$$

Hence, the standard deviation of $X - Y$ is 1.6.

The random variable $Z = (X - Y + 3.2)/(1.6)$ will have the standard normal distribution. It can be found from the table given at the end of this book that

$$\begin{aligned} \Pr(X - Y > 0) &= \Pr(Z > 2) = 1 - \Phi(2) \\ &= 0.0227. \end{aligned}$$

Therefore, the probability that the wife will be taller than her husband is 0.0227. ◀

Summary

If a random vector (X, Y) has a bivariate normal distribution, then every linear combination $aX + bY + c$ has a normal distribution. In particular, the marginal distributions of X and Y are normal. Also, the conditional distribution of X given $Y = y$ is normal with the conditional mean being a linear function of y and the conditional variance being constant in y . (Similarly, for the conditional distribution of Y given $X = x$.) A more thorough treatment of the bivariate normal distributions and higher-dimensional generalizations can be found in the book by D. F. Morrison (1990).

Exercises

1. Consider again the joint distribution of heights of husbands and wives in Example 5.10.6. Find the 0.95 quantile of the conditional distribution of the height of the wife given that the height of the husband is 72 inches.
2. Suppose that two different tests A and B are to be given to a student chosen at random from a certain population. Suppose also that the mean score on test A is 85, and the

standard deviation is 10; the mean score on test B is 90, and the standard deviation is 16; the scores on the two tests have a bivariate normal distribution; and the correlation of the two scores is 0.8. If the student's score on test A is 80, what is the probability that her score on test B will be higher than 90?

3. Consider again the two tests A and B described in Exercise 2. If a student is chosen at random, what is the probability that the sum of her scores on the two tests will be greater than 200?

4. Consider again the two tests A and B described in Exercise 2. If a student is chosen at random, what is the probability that her score on test A will be higher than her score on test B ?

5. Consider again the two tests A and B described in Exercise 2. If a student is chosen at random, and her score on test B is 100, what predicted value of her score on test A has the smallest M.S.E., and what is the value of this minimum M.S.E.?

6. Suppose that the random variables X_1 and X_2 have a bivariate normal distribution, for which the joint p.d.f. is specified by Eq. (5.10.2). Determine the value of the constant b for which $\text{Var}(X_1 + bX_2)$ will be a minimum.

7. Suppose that X_1 and X_2 have a bivariate normal distribution for which $E(X_1|X_2) = 3.7 - 0.15X_2$, $E(X_2|X_1) = 0.4 - 0.6X_1$, and $\text{Var}(X_2|X_1) = 3.64$. Find the mean and the variance of X_1 , the mean and the variance of X_2 , and the correlation of X_1 and X_2 .

8. Let $f(x_1, x_2)$ denote the p.d.f. of the bivariate normal distribution specified by Eq. (5.10.2). Show that the maximum value of $f(x_1, x_2)$ is attained at the point at which $x_1 = \mu_1$ and $x_2 = \mu_2$.

9. Let $f(x_1, x_2)$ denote the p.d.f. of the bivariate normal distribution specified by Eq. (5.10.2), and let k be a constant such that

$$0 < k < \frac{1}{2\pi(1 - \rho^2)^{1/2}\sigma_1\sigma_2}.$$

Show that the points (x_1, x_2) such that $f(x_1, x_2) = k$ lie on a circle if $\rho = 0$ and $\sigma_1 = \sigma_2$, and these points lie on an ellipse otherwise.

10. Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, and two other random variables Y_1 and Y_2 are defined as follows:

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + b_1, \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + b_2, \end{aligned}$$

where

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0.$$

Show that Y_1 and Y_2 also have a bivariate normal distribution.

11. Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, and $\text{Var}(X_1) = \text{Var}(X_2)$. Show that the sum $X_1 + X_2$ and the difference $X_1 - X_2$ are independent random variables.

12. Suppose that the two measurements from flea beetles in Example 5.10.2 have the bivariate normal distribution with $\mu_1 = 201$, $\mu_2 = 118$, $\sigma_1 = 15.2$, $\sigma_2 = 6.6$, and $\rho = 0.64$. Suppose that the same two measurements from a second species also have the bivariate normal distribution with $\mu_1 = 187$, $\mu_2 = 131$, $\sigma_1 = 15.2$, $\sigma_2 = 6.6$, and $\rho = 0.64$. Let (X_1, X_2) be a pair of measurements on a flea beetle from one of these two species. Let a_1, a_2 be constants.

- a.** For each of the two species, find the mean and standard deviation of $a_1X_1 + a_2X_2$. (Note that the variances for the two species will be the same. How do you know that?)
- b.** Find a_1 and a_2 to maximize the ratio of the difference between the two means found in part (a) to the standard deviation found in part (a). There is a sense in which this linear combination $a_1X_1 + a_2X_2$ does the best job of distinguishing the two species among all possible linear combinations.

13. Suppose that the joint p.d.f. of two random variables X and Y is proportional, as a function of (x, y) , to

$$\exp\left(-[ax^2 + by^2 + cxy + ex + gy + h]\right),$$

where $a > 0$, $b > 0$, and c, e, g , and h are all constants. Assume that $ab > (c/2)^2$. Prove that X and Y have a bivariate normal distribution, and find the means, variances, and correlation.

14. Suppose that a random variable X has a normal distribution, and for every x , the conditional distribution of another random variable Y given that $X = x$ is a normal distribution with mean $ax + b$ and variance τ^2 , where a, b , and τ^2 are constants. Prove that the joint distribution of X and Y is a bivariate normal distribution.

15. Let X_1, \dots, X_n be i.i.d. random variables having the normal distribution with mean μ and variance σ^2 . Define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, the sample mean. In this problem, we shall find the conditional distribution of each X_i given \bar{X}_n .

- a.** Show that X_i and \bar{X}_n have the bivariate normal distribution with both means μ , variances σ^2 and σ^2/n , and correlation $1/\sqrt{n}$. *Hint:* Let $Y = \sum_{j \neq i} X_j$. Now show that Y and X_i are independent normals and \bar{X}_n and X_i are linear combinations of Y and X_i .
- b.** Show that the conditional distribution of X_i given $\bar{X}_n = \bar{x}_n$ is normal with mean \bar{x}_n and variance $\sigma^2(1 - 1/n)$.

5.11 Supplementary Exercises

1. Let X and P be random variables. Suppose that the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p . Suppose that the distribution of P is the beta distribution with parameters $\alpha = 1$ and $\beta = 1$. Find the marginal distribution of X .

2. Suppose that X , Y , and Z are i.i.d. random variables and each has the standard normal distribution. Evaluate $\Pr(3X + 2Y < 6Z - 7)$.

3. Suppose that X and Y are independent Poisson random variables such that $\text{Var}(X) + \text{Var}(Y) = 5$. Evaluate $\Pr(X + Y < 2)$.

4. Suppose that X has a normal distribution such that $\Pr(X < 116) = 0.20$ and $\Pr(X < 328) = 0.90$. Determine the mean and the variance of X .

5. Suppose that a random sample of four observations is drawn from the Poisson distribution with mean λ , and let \bar{X} denote the sample mean. Show that

$$\Pr\left(\bar{X} < \frac{1}{2}\right) = (4\lambda + 1)e^{-4\lambda}.$$

6. The lifetime X of an electronic component has the exponential distribution such that $\Pr(X \leq 1000) = 0.75$. What is the expected lifetime of the component?

7. Suppose that X has the normal distribution with mean μ and variance σ^2 . Express $E(X^3)$ in terms of μ and σ^2 .

8. Suppose that a random sample of 16 observations is drawn from the normal distribution with mean μ and standard deviation 12, and that independently another random sample of 25 observations is drawn from the normal distribution with the same mean μ and standard deviation 20. Let \bar{X} and \bar{Y} denote the sample means of the two samples. Evaluate $\Pr(|\bar{X} - \bar{Y}| < 5)$.

9. Suppose that men arrive at a ticket counter according to a Poisson process at the rate of 120 per hour, and women arrive according to an independent Poisson process at the rate of 60 per hour. Determine the probability that four or fewer people arrive in a one-minute period.

10. Suppose that X_1, X_2, \dots are i.i.d. random variables, each of which has m.g.f. $\psi(t)$. Let $Y = X_1 + \dots + X_N$, where the number of terms N in this sum is a random variable having the Poisson distribution with mean λ . Assume that N and X_1, X_2, \dots are independent, and $Y = 0$ if $N = 0$. Determine the m.g.f. of Y .

11. Every Sunday morning, two children, Craig and Jill, independently try to launch their model airplanes. On each Sunday, Craig has probability $1/3$ of a successful launch, and Jill has probability $1/5$ of a successful launch. Determine the expected number of Sundays required until at least one of the two children has a successful launch.

12. Suppose that a fair coin is tossed until at least one head and at least one tail have been obtained. Let X denote the number of tosses that are required. Find the p.f. of X .

13. Suppose that a pair of balanced dice are rolled 120 times, and let X denote the number of rolls on which the sum of the two numbers is 12. Use the Poisson approximation to approximate $\Pr(X = 3)$.

14. Suppose that X_1, \dots, X_n form a random sample from the uniform distribution on the interval $[0, 1]$. Let $Y_1 = \min\{X_1, \dots, X_n\}$, $Y_n = \max\{X_1, \dots, X_n\}$, and $W = Y_n - Y_1$. Show that each of the random variables Y_1 , Y_n , and W has a beta distribution.

15. Suppose that events occur in accordance with a Poisson process at the rate of five events per hour.

- Determine the distribution of the waiting time T_1 until the first event occurs.
- Determine the distribution of the total waiting time T_k until k events have occurred.
- Determine the probability that none of the first k events will occur within 20 minutes of one another.

16. Suppose that five components are functioning simultaneously, that the lifetimes of the components are i.i.d., and that each lifetime has the exponential distribution with parameter β . Let T_1 denote the time from the beginning of the process until one of the components fails; and let T_5 denote the total time until all five components have failed. Evaluate $\text{Cov}(T_1, T_5)$.

17. Suppose that X_1 and X_2 are independent random variables, and X_i has the exponential distribution with parameter β_i ($i = 1, 2$). Show that for each constant $k > 0$,

$$\Pr(X_1 > kX_2) = \frac{\beta_2}{k\beta_1 + \beta_2}.$$

18. Suppose that 15,000 people in a city with a population of 500,000 are watching a certain television program. If 200 people in the city are contacted at random, what is the approximate probability that fewer than four of them are watching the program?

19. Suppose that it is desired to estimate the proportion of persons in a large population who have a certain characteristic. A random sample of 100 persons is selected from the population without replacement, and the proportion \bar{X} of persons in the sample who have the characteristic is observed. Show that, no matter how large the population is, the standard deviation of \bar{X} is at most 0.05.

20. Suppose that X has the binomial distribution with parameters n and p , and that Y has the negative binomial distribution with parameters r and p , where r is a positive integer. Show that $\Pr(X < r) = \Pr(Y > n - r)$ by showing

that both the left side and the right side of this equation can be regarded as the probability of the same event in a sequence of Bernoulli trials with probability p of success.

21. Suppose that X has the Poisson distribution with mean λt , and that Y has the gamma distribution with parameters $\alpha = k$ and $\beta = \lambda$, where k is a positive integer. Show that $\Pr(X \geq k) = \Pr(Y \leq t)$ by showing that both the left side and the right side of this equation can be regarded as the probability of the same event in a Poisson process in which the expected number of occurrences per unit of time is λ .

22. Suppose that X is a random variable having a continuous distribution with p.d.f. $f(x)$ and c.d.f. $F(x)$, and for which $\Pr(X > 0) = 1$. Let the failure rate $h(x)$ be as defined in Exercise 18 of Sec. 5.7. Show that

$$\exp\left[-\int_0^x h(t) dt\right] = 1 - F(x).$$

23. Suppose that 40 percent of the students in a large population are freshmen, 30 percent are sophomores, 20 percent are juniors, and 10 percent are seniors. Suppose that

10 students are selected at random from the population, and let X_1, X_2, X_3, X_4 denote, respectively, the numbers of freshmen, sophomores, juniors, and seniors that are obtained.

- a. Determine $\rho(X_i, X_j)$ for each pair of values i and j ($i < j$).
- b. For what values of i and j ($i < j$) is $\rho(X_i, X_j)$ most negative?
- c. For what values of i and j ($i < j$) is $\rho(X_i, X_j)$ closest to 0?

24. Suppose that X_1 and X_2 have the bivariate normal distribution with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 , and correlation ρ . Determine the distribution of $X_1 - 3X_2$.

25. Suppose that X has the standard normal distribution, and the conditional distribution of Y given X is the normal distribution with mean $2X - 3$ and variance 12. Determine the marginal distribution of Y and the value of $\rho(X, Y)$.

26. Suppose that X_1 and X_2 have a bivariate normal distribution with $E(X_2) = 0$. Evaluate $E(X_1^2 X_2)$.