CHAPTER 1

Probability and Distributions

1.1 Introduction

Many kinds of investigations may be characterized in part by the fact that repeated experimentation, under essentially the same conditions, is more or less standard procedure. For instance, in medical research, interest may center on the effect of a drug that is to be administered; or an economist may be concerned with the prices of three specified commodities at various time intervals; or the agronomist may wish to study the effect that a chemical fertilizer has on the yield of a cereal grain. The only way in which an investigator can elicit information about any such phenomenon is to perform his experiment. Each experiment terminates with an *outcome*. But it is characteristic of these experiments that the outcome cannot be predicted with certainty prior to the performance of the experiment.

Suppose that we have such an experiment, the outcome of which cannot be predicted with certainty, but the experiment is of such a nature that a collection of every possible outcome can be described prior to its performance. If this kind of experiment can be repeated under the same conditions, it is called a *random experiment*, and the collection of every possible outcome is called the experimental space or the *sample space*.

Example 1. In the toss of a coin, let the outcome tails be denoted by T and let the outcome heads be denoted by H. If we assume that the coin may be repeatedly tossed under the same conditions, then the toss of this coin is an example of a random experiment in which the outcome is one of the two symbols T and H; that is, the sample space is the collection of these two symbols.

Example 2. In the cast of one red die and one white die, let the outcome be the ordered pair (number of spots up on the red die, number of spots up on the white die). If we assume that these two dice may be repeatedly cast under the same conditions, then the cast of this pair of dice is a random experiment and the sample space consists of the following 36 ordered pairs: $(1, 1), \ldots, (1, 6), (2, 1), \ldots, (2, 6), \ldots, (6, 6)$.

Let $\mathscr{C}$ denote a sample space, and let C represent a part of $\mathscr{C}$. If, upon the performance of the experiment, the outcome is in C, we shall say that the *event* C has occurred. Now conceive of our having made N repeated performances of the random experiment. Then we can count the number f of times (the frequency) that the event C actually occurred throughout the N performances. The ratio f/N is called the *relative frequency* of the event C in these N experiments. A relative frequency is usually quite erratic for small values of N, as you can discover by tossing a coin. But as N increases, experience indicates that we associate with the event C a number, say p, that is equal or approximately equal to that number about which the relative frequency seems to stabilize. If we do this, then the number p can be interpreted as that number which, in future performances of the experiment, the relative frequency of the event C will either equal or approximate. Thus, although we *cannot* predict the outcome of a random experiment, we *can*, for a large value of N, predict approximately the relative frequency with which the outcome will be in C. The number p associated with the event C is given various names. Sometimes it is called the *probability* that the outcome of the random experiment is in C; sometimes it is called the *probability* of the event C; and sometimes it is called the *probability measure* of C. The context usually suggests an appropriate choice of terminology.

Example 3. Let $\mathscr{C}$ denote the sample space of Example 2 and let C be the collection of every ordered pair of $\mathscr{C}$ for which the sum of the pair is equal to seven. Thus C is the collection (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), and (6, 1). Suppose that the dice are cast $N = 400$ times and let f, the frequency of a sum of seven, be $f = 60$. Then the relative frequency with which the outcome was in C is $f/N = \frac{60}{400} = 0.15$. Thus we might associate with C a number p that is close to 0.15, and p would be called the probability of the event C.

Remark. The preceding interpretation of probability is sometimes referred to as the *relative frequency approach*, and it obviously depends upon the fact that an experiment can be repeated under essentially identical conditions. However, many persons extend probability to other situations by treating it as a rational measure of belief. For example, the statement $p = \frac{2}{5}$ would mean to them that their *personal* or *subjective* probability of the event C is equal to $\frac{2}{5}$. Hence, if they are not opposed to gambling, this could be interpreted as a willingness on their part to bet on the outcome of C so that the two possible payoffs are in the ratio $p/(1 - p) = \frac{2}{5}/\frac{3}{5} = \frac{2}{3}$. Moreover, if they truly believe that $p = \frac{2}{5}$ is correct, they would be willing to accept either side of the bet: (a) win 3 units if C occurs and lose 2 if it does not occur, or (b) win 2 units if C does not occur and lose 3 if it does. However, since the mathematical properties of probability given in Section 1.3 are consistent with either of these interpretations, the subsequent mathematical development does not depend upon which approach is used.

The primary purpose of having a mathematical theory of statistics is to provide mathematical models for random experiments. Once a model for such an experiment has been provided and the theory worked out in detail, the statistician may, within this framework, make inferences (that is, draw conclusions) about the random experiment. The construction of such a model requires a theory of probability. One of the more logically satisfying theories of probability is that based on the concepts of sets and functions of sets. These concepts are introduced in Section 1.2.

1.2 Set Theory

The concept of a *set* or a *collection* of objects is usually left undefined. However, a particular set can be described so that there is no misunderstanding as to what collection of objects is under consideration. For example, the set of the first 10 positive integers is sufficiently well described to make clear that the numbers $\frac{3}{4}$ and 14 are not in the set, while the number 3 is in the set. If an object belongs to a set, it is said to be an *element* of the set. For example, if A denotes the set of real numbers x for which $0 \le x \le 1$, then $\frac{3}{4}$ is an element of the set A. The fact that $\frac{3}{4}$ is an element of the set A is indicated by writing $\frac{3}{4} \in A$. More generally, $a \in A$ means that a is an element of the set A.

The sets that concern us will frequently be *sets of numbers*. However, the language of sets of *points* proves somewhat more convenient than that of sets of numbers. Accordingly, we briefly indicate how we use this terminology. In analytic geometry considerable emphasis is placed on the fact that to each point on a line (on which an origin and a unit point have been selected) there corresponds one and only one number, say x; and that to each number x there corresponds one and only one point on the line. This one-to-one correspondence between the numbers and points on a line enables us to speak, without misunderstanding, of the "point x" instead of the "number x." Furthermore, with a plane rectangular coordinate system and with x and y numbers, to each symbol (x, y) there corresponds one and only one point in the plane; and to each point in the plane there corresponds but one such symbol. Here again, we may speak of the "point (x, y)," meaning the "ordered number pair x and y." This convenient language can be used when we have a rectangular coordinate system in a space of three or more dimensions. Thus the "point $(x_1, x_2, \ldots, x_n)$" means the numbers $x_1, x_2, \ldots, x_n$ in the order stated. Accordingly, in describing our sets, we frequently speak of a set of points (a set whose elements are points), being careful, of course, to describe the set so as to avoid any ambiguity. The notation $A = \{x : 0 \le x \le 1\}$ is read "A is the one-dimensional set of points x for which $0 \le x \le 1$." Similarly, $A = \{(x, y) : 0 \le x \le 1, 0 \le y \le 1\}$ can be read "A is the two-dimensional set of points (x, y) that are interior to, or on the boundary of, a square with opposite vertices at $(0, 0)$ and $(1, 1)$." We now give some definitions (together with illustrative examples) that lead to an elementary algebra of sets adequate for our purposes.

Definition 1. If each element of a set A_1 is also an element of set A_2, the set A_1 is called a *subset* of the set A_2. This is indicated by writing $A_1 \subset A_2$. If $A_1 \subset A_2$ and also $A_2 \subset A_1$, the two sets have the same elements, and this is indicated by writing $A_1 = A_2$.

Example 1. Let $A_1 = \{x : 0 \le x \le 1\}$ and $A_2 = \{x : -1 \le x \le 2\}$. Here the one-dimensional set A_1 is seen to be a subset of the one-dimensional set A_2; that is, $A_1 \subset A_2$. Subsequently, when the dimensionality of the set is clear, we shall not make specific reference to it.

Example 2. Let $A_1 = \{(x, y) : 0 \le x = y \le 1\}$ and $A_2 = \{(x, y) : 0 \le x \le 1, 0 \le y \le 1\}$. Since the elements of A_1 are the points on one diagonal of the square, then $A_1 \subset A_2$.

Definition 2. If a set A has no elements, A is called the *null set*. This is indicated by writing $A = \varnothing$.

Definition 3. The set of all elements that belong to at least one of the sets A_1 and A_2 is called the *union* of A_1 and A_2. The union of A_1 and A_2 is indicated by writing $A_1 \cup A_2$. The union of several sets $A_1, A_2, A_3, \ldots$ is the set of all elements that belong to at least one of the several sets. This union is denoted by $A_1 \cup A_2 \cup A_3 \cup \cdots$ or by $A_1 \cup A_2 \cup \cdots \cup A_k$ if a finite number k of sets is involved.

Example 3. Let $A_1 = \{x : x = 0, 1, \ldots, 10\}$ and $A_2 = \{x : x = 8, 9, 10, 11, \text{ or } 11 < x \le 12\}$. Then $A_1 \cup A_2 = \{x : x = 0, 1, \ldots, 8, 9, 10, 11, \text{ or } 11 < x \le 12\} = \{x : x = 0, 1, \ldots, 8, 9, 10, \text{ or } 11 \le x \le 12\}$.

Example 4. Let A_1 and A_2 be defined as in Example 1. Then $A_1 \cup A_2 = A_2$.

Example 5. Let $A_2 = \varnothing$. Then $A_1 \cup A_2 = A_1$ for every set A_1.

Example 6. For every set A, $A \cup A = A$.

Example 7. Let

$$A_k = \left\{x : \frac{1}{k+1} \le x \le 1\right\}, \qquad k = 1, 2, 3, \ldots.$$

Then $A_1 \cup A_2 \cup A_3 \cup \cdots = \{x : 0 < x \le 1\}$. Note that the number zero is not in this set, since it is not in one of the sets $A_1, A_2, A_3, \ldots$.

Definition 4. The set of all elements that belong to each of the sets A_1 and A_2 is called the *intersection* of A_1 and A_2. The intersection of A_1 and A_2 is indicated by writing $A_1 \cap A_2$. The intersection of several sets $A_1, A_2, A_3, \ldots$ is the set of all elements that belong to each of the sets $A_1, A_2, A_3, \ldots$. This intersection is denoted by $A_1 \cap A_2 \cap A_3 \cap \cdots$ or by $A_1 \cap A_2 \cap \cdots \cap A_k$ if a finite number k of sets is involved.

Example 8. Let $A_1 = \{(0, 0), (0, 1), (1, 1)\}$ and $A_2 = \{(1, 1), (1, 2), (2, 1)\}$. Then $A_1 \cap A_2 = \{(1, 1)\}$.

Example 9. Let $A_1 = \{(x, y) : 0 \le x + y \le 1\}$ and $A_2 = \{(x, y) : 1 < x + y\}$. Then A_1 and A_2 have no points in common and $A_1 \cap A_2 = \varnothing$.

Example 10. For every set A, $A \cap A = A$ and $A \cap \varnothing = \varnothing$.

Example 11. Let

$$A_k = \left\{x : 0 < x < \frac{1}{k}\right\}, \qquad k = 1, 2, 3, \ldots.$$

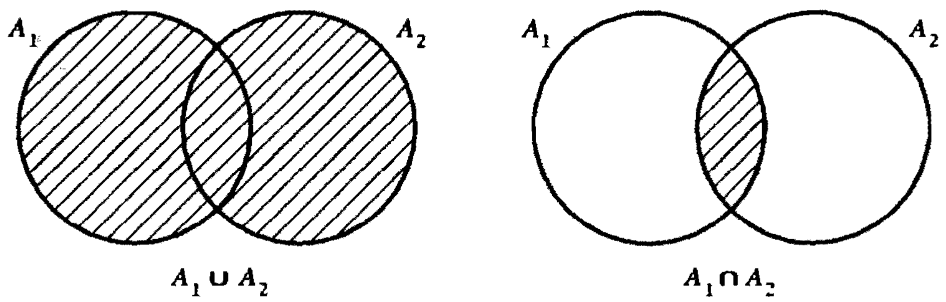

FIGURE 1.1

Then $A_1 \cap A_2 \cap A_3 \cdots$ is the null set, since there is no point that belongs to each of the sets $A_1, A_2, A_3, \ldots$.

Example 12. Let A_1 and A_2 represent the sets of points enclosed, respectively, by two intersecting circles. Then the sets $A_1 \cup A_2$ and $A_1 \cap A_2$ are represented, respectively, by the shaded regions in the *Venn diagrams* in Figure 1.1.

Example 13. Let A_1, A_2, and A_3 represent the sets of points enclosed, respectively, by three intersecting circles. Then the sets $(A_1 \cup A_2) \cap A_3$ and $(A_1 \cap A_2) \cup A_3$ are depicted in Figure 1.2.

Definition 5. In certain discussions or considerations, the totality of all elements that pertain to the discussion can be described. This set of all elements under consideration is given a special name. It is called the *space*. We shall often denote spaces by capital script letters such as $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{C}$.

Example 14. Let the number of heads, in tossing a coin four times, be denoted by x. Of necessity, the number of heads will be one of the numbers 0, 1, 2, 3, 4. Here, then, the space is the set $\mathscr{A} = \{0, 1, 2, 3, 4\}$.

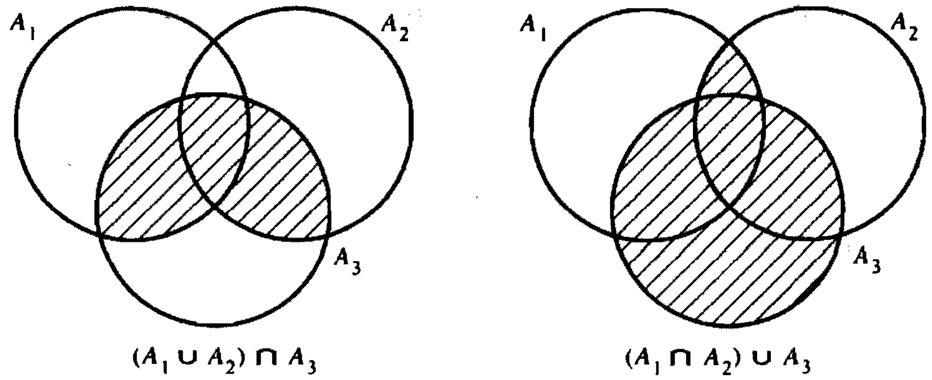

FIGURE 1.2

Example 15. Consider all nondegenerate rectangles of base x and height y. To be meaningful, both x and y must be positive. Thus the space is the set $\mathscr{A} = \{(x, y) : x > 0, y > 0\}$.

Definition 6. Let $\mathscr{A}$ denote a space and let A be a subset of the set $\mathscr{A}$. The set that consists of all elements of $\mathscr{A}$ that are not elements of A is called the *complement* of A (actually, with respect to $\mathscr{A}$). The complement of A is denoted by A^*. In particular, $\mathscr{A}^* = \varnothing$.

Example 16. Let $\mathscr{A}$ be defined as in Example 14, and let the set $A = \{0, 1\}$. The complement of A (with respect to $\mathscr{A}$) is $A^* = \{2, 3, 4\}$.

Example 17. Given $A \subset \mathscr{A}$. Then $A \cup A^* = \mathscr{A}$, $A \cap A^* = \varnothing$, $A \cup \mathscr{A} = \mathscr{A}$, $A \cap \mathscr{A} = A$, and $(A^*)^* = A$.

In the calculus, functions such as

$$f(x) = 2x, \qquad -\infty < x < \infty,$$

or

$$\begin{aligned} g(x, y) &= e^{-x-y}, \qquad 0 < x < \infty, \quad 0 < y < \infty, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

or possibly

$$\begin{aligned} h(x_1, x_2, \ldots, x_n) &= 3x_1x_2 \cdots x_n, \qquad 0 \le x_i \le 1, \quad i = 1, 2, \ldots, n, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

were of common occurrence. The value of $f(x)$ at the "point $x = 1$" is $f(1) = 2$; the value of $g(x, y)$ at the "point $(-1, 3)$" is $g(-1, 3) = 0$; the value of $h(x_1, x_2, \ldots, x_n)$ at the "point $(1, 1, \ldots, 1)$" is 3. Functions such as these are called functions of a point or, more simply, *point functions* because they are evaluated (if they have a value) at a point in a space of indicated dimension.

There is no reason why, if they prove useful, we should not have functions that can be evaluated, not necessarily at a point, but for an entire set of points. Such functions are naturally called functions of a set or, more simply, *set functions*. We shall give some examples of set functions and evaluate them for certain simple sets.

Example 18. Let A be a set in one-dimensional space and let $Q(A)$ be equal to the number of points in A which correspond to positive integers. Then $Q(A)$ is a function of the set A. Thus, if $A = \{x : 0 < x < 5\}$, then $Q(A) = 4$; if $A = \{-2, -1\}$, then $Q(A) = 0$; if $A = \{x : -\infty < x < 6\}$, then $Q(A) = 5$.

Example 19. Let A be a set in two-dimensional space and let $Q(A)$ be the area of A, if A has a finite area; otherwise, let $Q(A)$ be undefined. Thus, if

$A = \{(x, y) : x^2 + y^2 \le 1\}$, then $Q(A) = \pi$; if $A = \{(0, 0), (1, 1), (0, 1)\}$, then $Q(A) = 0$; if $A = \{(x, y) : 0 \le x, 0 \le y, x + y \le 1\}$, then $Q(A) = \frac{1}{2}$.

Example 20. Let A be a set in three-dimensional space and let $Q(A)$ be the volume of A, if A has a finite volume; otherwise, let $Q(A)$ be undefined. Thus, if $A = \{(x, y, z) : 0 \le x \le 2, 0 \le y \le 1, 0 \le z \le 3\}$, then $Q(A) = 6$; if $A = \{(x, y, z) : x^2 + y^2 + z^2 \ge 1\}$, then $Q(A)$ is undefined.

At this point we introduce the following notations. The symbol

$$\int_A f(x)\, dx$$

will mean the ordinary (Riemann) integral of $f(x)$ over a prescribed one-dimensional set A; the symbol

$$\int_A \int g(x, y)\, dx\, dy$$

will mean the Riemann integral of $g(x, y)$ over a prescribed two-dimensional set A; and so on. To be sure, unless these sets A and these functions $f(x)$ and $g(x, y)$ are chosen with care, the integrals will frequently fail to exist. Similarly, the symbol

$$\sum_A f(x)$$

will mean the sum extended over all $x \in A$; the symbol

$$\sum_A \sum g(x, y)$$

will mean the sum extended over all $(x, y) \in A$; and so on.

Example 21. Let A be a set in one-dimensional space and let $Q(A) = \sum_A f(x)$, where

$$\begin{aligned} f(x) &= (\tfrac{1}{2})^x, \qquad x = 1, 2, 3, \ldots, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

If $A = \{x : 0 \le x \le 3\}$, then

$$Q(A) = \tfrac{1}{2} + (\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 = \tfrac{7}{8}.$$

Example 22. Let $Q(A) = \sum_A f(x)$, where

$$\begin{aligned} f(x) &= p^x(1 - p)^{1-x}, \qquad x = 0, 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

If $A = \{0\}$, then

$$Q(A) = \sum_{x=0}^{0} p^x(1 - p)^{1-x} = 1 - p;$$

if $A = \{x : 1 \le x \le 2\}$, then $Q(A) = f(1) = p$.

Example 23. Let A be a one-dimensional set and let

$$Q(A) = \int_A e^{-x}\, dx.$$

Thus, if $A = \{x : 0 \le x < \infty\}$, then

$$Q(A) = \int_0^\infty e^{-x}\, dx = 1;$$

if $A = \{x : 1 \le x \le 2\}$, then

$$Q(A) = \int_1^2 e^{-x}\, dx = e^{-1} - e^{-2};$$

if $A_1 = \{x : 0 \le x \le 1\}$ and $A_2 = \{x : 1 < x \le 3\}$, then

$$\begin{aligned} Q(A_1 \cup A_2) &= \int_0^3 e^{-x}\, dx \\ &= \int_0^1 e^{-x}\, dx + \int_1^3 e^{-x}\, dx \\ &= Q(A_1) + Q(A_2); \end{aligned}$$

if $A = A_1 \cup A_2$, where $A_1 = \{x : 0 \le x \le 2\}$ and $A_2 = \{x : 1 \le x \le 3\}$, then

$$\begin{aligned} Q(A) = Q(A_1 \cup A_2) &= \int_0^3 e^{-x}\, dx \\ &= \int_0^2 e^{-x}\, dx + \int_1^3 e^{-x}\, dx - \int_1^2 e^{-x}\, dx \\ &= Q(A_1) + Q(A_2) - Q(A_1 \cap A_2). \end{aligned}$$

Example 24. Let A be a set in n-dimensional space and let

$$Q(A) = \int \cdots \int_A dx_1\, dx_2 \cdots dx_n.$$

If $A = \{(x_1, x_2, \ldots, x_n) : 0 \le x_1 \le x_2 \le \cdots \le x_n \le 1\}$, then

$$\begin{aligned} Q(A) &= \int_0^1 \int_0^{x_n} \cdots \int_0^{x_3} \int_0^{x_2} dx_1\, dx_2 \cdots dx_{n-1}\, dx_n \\ &= \frac{1}{n!}, \qquad \text{where } n! = n(n-1) \cdots 3 \cdot 2 \cdot 1. \end{aligned}$$

EXERCISES

1.1. Find the union $A_1 \cup A_2$ and the intersection $A_1 \cap A_2$ of the two sets A_1 and A_2, where:
(a) $A_1 = \{0, 1, 2\}$, $A_2 = \{2, 3, 4\}$.
(b) $A_1 = \{x : 0 < x < 2\}$, $A_2 = \{x : 1 \le x < 3\}$.
(c) $A_1 = \{(x, y) : 0 < x < 2, 0 < y < 2\}$,
$A_2 = \{(x, y) : 1 < x < 3, 1 < y < 3\}$.

1.2. Find the complement A^* of the set A with respect to the space $\mathscr{A}$ if:
(a) $\mathscr{A} = \{x : 0 < x < 1\}$, $A = \{x : \frac{5}{8} \le x < 1\}$.
(b) $\mathscr{A} = \{(x, y, z) : x^2 + y^2 + z^2 \le 1\}$, $A = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}$.
(c) $\mathscr{A} = \{(x, y) : |x| + |y| \le 2\}$, $A = \{(x, y) : x^2 + y^2 < 2\}$.

1.3. List all possible arrangements of the four letters m, a, r, and y. Let A_1 be the collection of the arrangements in which y is in the last position. Let A_2 be the collection of the arrangements in which m is in the first position. Find the union and intersection of A_1 and A_2.

1.4. By use of Venn diagrams, in which the space $\mathscr{A}$ is the set of points enclosed by a rectangle containing the circles, compare the following sets:
(a) $A_1 \cap (A_2 \cup A_3)$ and $(A_1 \cap A_2) \cup (A_1 \cap A_3)$.
(b) $A_1 \cup (A_2 \cap A_3)$ and $(A_1 \cup A_2) \cap (A_1 \cup A_3)$.
(c) $(A_1 \cup A_2)^*$ and $A_1^* \cap A_2^*$.
(d) $(A_1 \cap A_2)^*$ and $A_1^* \cup A_2^*$.

1.5. If a sequence of sets $A_1, A_2, A_3, \ldots$ is such that $A_k \subset A_{k+1}$, $k = 1, 2, 3, \ldots$, the sequence is said to be a *nondecreasing sequence*. Give an example of this kind of sequence of sets.

1.6. If a sequence of sets $A_1, A_2, A_3, \ldots$ is such that $A_k \supset A_{k+1}$, $k = 1, 2, 3, \ldots$, the sequence is said to be a *nonincreasing sequence*. Give an example of this kind of sequence of sets.

1.7. If $A_1, A_2, A_3, \ldots$ are sets such that $A_k \subset A_{k+1}$, $k = 1, 2, 3, \ldots$, $\lim_{k\to\infty} A_k$ is defined as the union $A_1 \cup A_2 \cup A_3 \cup \cdots$. Find $\lim_{k\to\infty} A_k$ if:
(a) $A_k = \{x : 1/k \le x \le 3 - 1/k\}$, $k = 1, 2, 3, \ldots$.
(b) $A_k = \{(x, y) : 1/k \le x^2 + y^2 \le 4 - 1/k\}$, $k = 1, 2, 3, \ldots$.

1.8. If $A_1, A_2, A_3, \ldots$ are sets such that $A_k \supset A_{k+1}$, $k = 1, 2, 3, \ldots$, $\lim_{k\to\infty} A_k$ is defined as the intersection $A_1 \cap A_2 \cap A_3 \cap \cdots$. Find $\lim_{k\to\infty} A_k$ if:
(a) $A_k = \{x : 2 - 1/k < x \le 2\}$, $k = 1, 2, 3, \ldots$.
(b) $A_k = \{x : 2 < x \le 2 + 1/k\}$, $k = 1, 2, 3, \ldots$.
(c) $A_k = \{(x, y) : 0 \le x^2 + y^2 \le 1/k\}$, $k = 1, 2, 3, \ldots$.

1.9. For every one-dimensional set A, let $Q(A) = \sum_A f(x)$, where $f(x) = (\frac{2}{3})(\frac{1}{3})^x$, $x = 0, 1, 2, \ldots,$ zero elsewhere. If $A_1 = \{x : x = 0, 1, 2, 3\}$ and $A_2 = \{x : x = 0, 1, 2, \ldots\}$, find $Q(A_1)$ and $Q(A_2)$.

Hint: Recall that $S_n = a + ar + \cdots + ar^{n-1} = a(1 - r^n)/(1 - r)$ and $\lim_{n\to\infty} S_n = a/(1 - r)$ provided that $|r| < 1$.

1.10. For every one-dimensional set A for which the integral exists, let $Q(A) = \int_A f(x)\,dx$, where $f(x) = 6x(1 - x)$, $0 < x < 1$, zero elsewhere; otherwise, let $Q(A)$ be undefined. If $A_1 = \{x : \frac{1}{4} < x < \frac{3}{4}\}$, $A_2 = \{\frac{1}{2}\}$, and $A_3 = \{x : 0 < x < 10\}$, find $Q(A_1)$, $Q(A_2)$, and $Q(A_3)$.

1.11. Let $Q(A) = \int_A \int (x^2 + y^2)\,dx\,dy$ for every two-dimensional set A for which the integral exists; otherwise, let $Q(A)$ be undefined. If $A_1 = \{(x, y) : -1 \le x \le 1, -1 \le y \le 1\}$, $A_2 = \{(x, y) : -1 \le x = y \le 1\}$, and $A_3 = \{(x, y) : x^2 + y^2 \le 1\}$, find $Q(A_1)$, $Q(A_2)$, and $Q(A_3)$.

Hint: In evaluating $Q(A_2)$, recall the definition of the double integral (or consider the volume under the surface $z = x^2 + y^2$ above the line segment $-1 \le x = y \le 1$ in the xy-plane). Use polar coordinates in the calculation of $Q(A_3)$.

1.12. Let $\mathscr{A}$ denote the set of points that are interior to, or on the boundary of, a square with opposite vertices at the points $(0, 0)$ and $(1, 1)$. Let $Q(A) = \int_A \int dy\,dx$.

(a) If $A \subset \mathscr{A}$ is the set $\{(x, y) : 0 < x < y < 1\}$, compute $Q(A)$.
(b) If $A \subset \mathscr{A}$ is the set $\{(x, y) : 0 < x = y < 1\}$, compute $Q(A)$.
(c) If $A \subset \mathscr{A}$ is the set $\{(x, y) : 0 < x/2 \le y \le 3x/2 < 1\}$, compute $Q(A)$.

1.13. Let $\mathscr{A}$ be the set of points interior to or on the boundary of a cube with edge of length 1. Moreover, say that the cube is in the first octant with one vertex at the point $(0, 0, 0)$ and an opposite vertex at the point $(1, 1, 1)$. Let $Q(A) = \iiint_A dx\,dy\,dz$.

(a) If $A \subset \mathscr{A}$ is the set $\{(x, y, z) : 0 < x < y < z < 1\}$, compute $Q(A)$.
(b) If A is the subset $\{(x, y, z) : 0 < x = y = z < 1\}$, compute $Q(A)$.

1.14. Let A denote the set $\{(x, y, z) : x^2 + y^2 + z^2 \le 1\}$. Evaluate $Q(A) = \iiint_A \sqrt{x^2 + y^2 + z^2}\,dx\,dy\,dz$.

Hint: Use spherical coordinates.

1.15. To join a certain club, a person must be either a statistician or a mathematician or both. Of the 25 members in this club, 19 are statisticians and 16 are mathematicians. How many persons in the club are both a statistician and a mathematician?

1.16. After a hard-fought football game, it was reported that, of the 11 starting players, 8 hurt a hip, 6 hurt an arm, 5 hurt a knee, 3 hurt both a hip and an arm, 2 hurt both a hip and a knee, 1 hurt both an arm and a knee, and no one hurt all three. Comment on the accuracy of the report.

1.3. The Probability Set Function

Let $\mathscr{C}$ denote the set of every possible outcome of a random experiment; that is, $\mathscr{C}$ is the sample space. It is our purpose to define a set function $P(C)$ such that if C is a subset of $\mathscr{C}$, then $P(C)$ is the probability that the outcome of the random experiment is an element of C. Henceforth it will be tacitly assumed that the structure of each set C is sufficiently simple to allow the computation. We have already seen that advantages accrue if we take $P(C)$ to be that number about which the relative frequency f/N of the event C tends to stabilize after a long series of experiments. This important fact suggests some of the properties that we would surely want the set function $P(C)$ to possess. For example, no relative frequency is ever negative; accordingly, we would want $P(C)$ to be a nonnegative set function. Again, the relative frequency of the whole sample space $\mathscr{C}$ is always 1. Thus we would want $P(\mathscr{C}) = 1$. Finally, if $C_1, C_2, C_3, \ldots$ are subsets of $\mathscr{C}$ such that no two of these subsets have a point in common, the relative frequency of the union of these sets is the sum of the relative frequencies of the sets, and we would want the set function $P(C)$ to reflect this additive property. We now formally define a probability set function.

Definition 7. If $P(C)$ is defined for a type of subset of the space $\mathscr{C}$, and if

(a) $P(C) \geq 0$,
(b) $P(C_1 \cup C_2 \cup C_3 \cup \cdots) = P(C_1) + P(C_2) + P(C_3) + \cdots$, where the sets C_i, $i = 1, 2, 3, \ldots$, are such that no two have a point in common (that is, where $C_i \cap C_j = \varnothing$, $i \neq j$),
(c) $P(\mathscr{C}) = 1$,

then P is called the *probability set function* of the outcome of the random experiment. For each subset C of $\mathscr{C}$, the number $P(C)$ is called the probability that the outcome of the random experiment is an element of the set C, or the probability of the event C, or the probability measure of the set C.

A probability set function tells us how the probability is distributed over various subsets C of a sample space $\mathscr{C}$. In this sense we speak of a distribution of probability.

Remark. In the definition, the phrase "a type of subset of the space $\mathscr{C}$" refers to the fact that P is a probability measure on a sigma field of subsets of $\mathscr{C}$ and would be explained more fully in a more advanced course. Nevertheless, a few observations can be made about the collection of subsets that are of the type. From condition (c) of the definition, we see that the space $\mathscr{C}$ must be in the collection. Condition (b) implies that if the sets $C_1, C_2, C_3, \ldots$ are in the collection, their union is also one of that type. Finally, we observe from the following theorems and their proofs that if the set C is in the collection, its complement must be one of those subsets. In particular, the null set, which is the complement of $\mathscr{C}$, must be in the collection.

The following theorems give us some other properties of a probability set function. In the statement of each of these theorems, $P(C)$ is taken, tacitly, to be a probability set function defined for a certain type of subset of the sample space $\mathscr{C}$.

Theorem 1. *For each $C \subset \mathscr{C}$, $P(C) = 1 - P(C^*)$.*

Proof. We have $\mathscr{C} = C \cup C^*$ and $C \cap C^* = \varnothing$. Thus, from (c) and (b) of Definition 7, it follows that

$$1 = P(C) + P(C^*),$$

which is the desired result.

Theorem 2. *The probability of the null set is zero; that is, $P(\varnothing) = 0$.*

Proof. In Theorem 1, take $C = \varnothing$ so that $C^* = \mathscr{C}$. Accordingly, we have

$$P(\varnothing) = 1 - P(\mathscr{C}) = 1 - 1 = 0,$$

and the theorem is proved.

Theorem 3. *If C_1 and C_2 are subsets of $\mathscr{C}$ such that $C_1 \subset C_2$, then $P(C_1) \le P(C_2)$.*

Proof. Now $C_2 = C_1 \cup (C_1^* \cap C_2)$ and $C_1 \cap (C_1^* \cap C_2) = \varnothing$. Hence, from (b) of Definition 7,

$$P(C_2) = P(C_1) + P(C_1^* \cap C_2).$$

However, from (a) of Definition 7, $P(C_1^* \cap C_2) \geq 0$; accordingly, $P(C_2) \geq P(C_1)$.

Theorem 4. *For each $C \subset \mathscr{C}$, $0 \leq P(C) \leq 1$.*

Proof. Since $\varnothing \subset C \subset \mathscr{C}$, we have by Theorem 3 that

$$P(\varnothing) \leq P(C) \leq P(\mathscr{C}) \quad \text{or} \quad 0 \leq P(C) \leq 1,$$

the desired result.

Theorem 5. *If C_1 and C_2 are subsets of $\mathscr{C}$, then*

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2).$$

Proof. Each of the sets $C_1 \cup C_2$ and C_2 can be represented, respectively, as a union of nonintersecting sets as follows:

$$C_1 \cup C_2 = C_1 \cup (C_1^* \cap C_2) \quad \text{and} \quad C_2 = (C_1 \cap C_2) \cup (C_1^* \cap C_2).$$

Thus, from (b) of Definition 7,

$$P(C_1 \cup C_2) = P(C_1) + P(C_1^* \cap C_2)$$

and

$$P(C_2) = P(C_1 \cap C_2) + P(C_1^* \cap C_2).$$

If the second of these equations is solved for $P(C_1^* \cap C_2)$ and this result substituted in the first equation, we obtain

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2).$$

This completes the proof.

Example 1. Let $\mathscr{C}$ denote the sample space of Example 2 of Section 1.1. Let the probability set function assign a probability of $\frac{1}{36}$ to each of the 36 points in $\mathscr{C}$. If $C_1 = \{(1, 1), (2, 1), (3, 1), (4, 1), (5, 1)\}$ and $C_2 = \{(1, 2), (2, 2), (3, 2)\}$, then $P(C_1) = \frac{5}{36}$, $P(C_2) = \frac{3}{36}$, $P(C_1 \cup C_2) = \frac{8}{36}$, and $P(C_1 \cap C_2) = 0$.

Example 2. Two coins are to be tossed and the outcome is the ordered pair (face on the first coin, face on the second coin). Thus the sample space may be represented as $\mathscr{C} = \{(\text{H}, \text{H}), (\text{H}, \text{T}), (\text{T}, \text{H}), (\text{T}, \text{T})\}$. Let the probability set function assign a probability of $\frac{1}{4}$ to each element of $\mathscr{C}$. Let $C_1 = \{(\text{H}, \text{H}), (\text{H}, \text{T})\}$ and $C_2 = \{(\text{H}, \text{H}), (\text{T}, \text{H})\}$. Then $P(C_1) = P(C_2) = \frac{1}{2}$, $P(C_1 \cap C_2) = \frac{1}{4}$, and, in accordance with Theorem 5, $P(C_1 \cup C_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$.

Let $\mathscr{C}$ denote a sample space and let $C_1, C_2, C_3, \ldots$ denote subsets of $\mathscr{C}$. If these subsets are such that no two have an element in common, they are called mutually disjoint sets and the corresponding events $C_1, C_2, C_3, \ldots$ are said to be *mutually exclusive events*. Then, for example, $P(C_1 \cup C_2 \cup C_3 \cup \cdots) = P(C_1) + P(C_2) + P(C_3) + \cdots$, in accordance with (b) of Definition 7. Moreover, if $\mathscr{C} = C_1 \cup C_2 \cup C_3 \cup \cdots$, the mutually exclusive events are further characterized as being *exhaustive* and the probability of their union is obviously equal to 1.

Let $\mathscr{C}$ be partitioned into k mutually disjoint subsets $C_1, C_2, \ldots, C_k$ in such a way that the union of these k mutually disjoint subsets is the sample space $\mathscr{C}$. Thus the events $C_1, C_2, \ldots, C_k$ are mutually exclusive and exhaustive. Suppose that the random experiment is of such a character that it is reasonable to *assume* that each of the mutually exclusive and exhaustive events C_i, $i = 1, 2, \ldots, k$, has the same probability. It is necessary, then, that $P(C_i) = 1/k$, $i = 1, 2, \ldots, k$; and we often say that the events $C_1, C_2, \ldots, C_k$ are *equally likely*. Let the event E be the union of r of these mutually exclusive events, say

$$E = C_1 \cup C_2 \cup \cdots \cup C_r, \qquad r \le k.$$

Then

$$P(E) = P(C_1) + P(C_2) + \cdots + P(C_r) = \frac{r}{k}.$$

Frequently, the integer k is called the total number of ways (for this particular partition of $\mathscr{C}$) in which the random experiment can terminate and the integer r is called the number of ways that are favorable to the event E. So, in this terminology, $P(E)$ is equal to the number of ways favorable to the event E divided by the total number of ways in which the experiment can terminate. It should be emphasized that in order to assign, *in this manner*, the probability r/k to the event E, we must assume that each of the mutually exclusive and exhaustive events $C_1, C_2, \ldots, C_k$ has the same probability $1/k$. This assumption of equally likely events then becomes a *part* of our probability model. Obviously, if this assumption is not realistic in an application, the probability of the event E cannot be computed in this way.

We next present an example that is illustrative of this model.

Example 3. Let a card be drawn at random from an ordinary deck of

52 playing cards. The sample space $\mathscr{C}$ is the union of $k = 52$ outcomes, and it is reasonable to assume that each of these outcomes has the same probability $\frac{1}{52}$. Accordingly, if E_1 is the set of outcomes that are spades, $P(E_1) = \frac{13}{52} = \frac{1}{4}$ because there are $r_1 = 13$ spades in the deck; that is, $\frac{1}{4}$ is the probability of drawing a card that is a spade. If E_2 is the set of outcomes that are kings, $P(E_2) = \frac{4}{52} = \frac{1}{13}$ because there are $r_2 = 4$ kings in the deck; that is, $\frac{1}{13}$ is the probability of drawing a card that is a king. These computations are very easy because there are no difficulties in the determination of the appropriate values of r and k. However, instead of drawing only one card, suppose that five cards are taken, at random and without replacement, from this deck. We can think of each five-card hand as being an outcome in a sample space. It is reasonable to assume that each of these outcomes has the same probability. Now if E_1 is the set of outcomes in which each card of the hand is a spade, $P(E_1)$ is equal to the number r_1 of all spade hands divided by the total number, say k, of five-card hands. It is shown in many books on algebra that

$$r_1 = \binom{13}{5} = \frac{13!}{5!\,8!} \quad \text{and} \quad k = \binom{52}{5} = \frac{52!}{5!\,47!}.$$

In general, if n is a positive integer and if x is a nonnegative integer with $x \le n$, then the binomial coefficient

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

is equal to the number of combinations of n things taken x at a time. If $x = 0$, $0! = 1$, so that $\binom{n}{0} = 1$. Thus, in the special case involving E_1,

$$P(E_1) = \frac{\binom{13}{5}}{\binom{52}{5}} = \frac{(13)(12)(11)(10)(9)}{(52)(51)(50)(49)(48)} = 0.0005,$$

approximately. Next, let E_2 be the set of outcomes in which at least one card is a spade. Then E_2^* is the set of outcomes in which no card is a spade. There are $r_2^* = \binom{39}{5}$ such outcomes. Hence

$$P(E_2^*) = \frac{\binom{39}{5}}{\binom{52}{5}} \quad \text{and} \quad P(E_2) = 1 - P(E_2^*).$$

Now suppose that E_3 is the set of outcomes in which exactly three cards are kings and exactly two cards are queens. We can select the three kings in any one of $\binom{4}{3}$ ways and the two queens in any one of $\binom{4}{2}$ ways. By a well-known counting principle, the number of outcomes in E_3 is $r_3 = \binom{4}{3}\binom{4}{2}$. Thus $P(E_3) = \binom{4}{3}\binom{4}{2}\Big/\binom{52}{5}$. Finally, let E_4 be the set of outcomes in which there are exactly two kings, two queens, and one jack. Then

$$P(E_4) = \frac{\binom{4}{2}\binom{4}{2}\binom{4}{1}}{\binom{52}{5}},$$

because the numerator of this fraction is the number of outcomes in E_4.

Example 3 and the previous discussion allow us to see one way in which we can define a probability set function, that is, a set function that satisfies the requirements of Definition 7. Suppose that our space $\mathscr{C}$ consists of k distinct points, which, for this discussion, we take to be in a one-dimensional space. If the random experiment that ends in one of those k points is such that it is reasonable to assume that these points are equally likely, we could assign $1/k$ to each point and let, for $C \subset \mathscr{C}$,

$$\begin{aligned} P(C) &= \frac{\text{number of points in } C}{k} \\ &= \sum_{x \in C} f(x), \qquad \text{where} \quad f(x) = \frac{1}{k}, \quad x \in \mathscr{C}. \end{aligned}$$

For illustration, in the cast of a die, we could take $\mathscr{C} = \{1, 2, 3, 4, 5, 6\}$ and $f(x) = \frac{1}{6}$, $x \in \mathscr{C}$, if we believe the die to be unbiased. Clearly, such a set function satisfies Definition 7.

The word *unbiased* in this illustration suggests the possibility that all six points might *not*, in all such cases, be equally likely. As a matter of fact, *loaded* dice do exist. In the case of a loaded die, some numbers occur more frequently than others in a sequence of casts of that die. For example, suppose that a die has been loaded so that the relative frequencies of the numbers in $\mathscr{C}$ *seem to stabilize* proportional to the number of spots that are on the *up* side. Thus we might assign $f(x) = x/21$, $x \in \mathscr{C}$, and the corresponding

$$P(C) = \sum_{x \in C} f(x)$$

would satisfy Definition 7. For illustration, this means that if $C = \{1, 2, 3\}$, then

$$P(C) = \sum_{x=1}^{3} f(x) = \frac{1}{21} + \frac{2}{21} + \frac{3}{21} = \frac{6}{21} = \frac{2}{7}.$$

Whether this probability set function is realistic can only be checked by performing the random experiment a large number of times.

EXERCISES

1.17. A positive integer from one to six is to be chosen by casting a die. Thus the elements c of the sample space $\mathscr{C}$ are 1, 2, 3, 4, 5, 6. Let $C_1 = \{1, 2, 3, 4\}$, $C_2 = \{3, 4, 5, 6\}$. If the probability set function P assigns a probability of $\frac{1}{6}$ to each of the elements of $\mathscr{C}$, compute $P(C_1)$, $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

1.18. A random experiment consists of drawing a card from an ordinary deck of 52 playing cards. Let the probability set function P assign a probability of $\frac{1}{52}$ to each of the 52 possible outcomes. Let C_1 denote the collection of the 13 hearts and let C_2 denote the collection of the 4 kings. Compute $P(C_1)$, $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

1.19. A coin is to be tossed as many times as necessary to turn up one head. Thus the elements c of the sample space $\mathscr{C}$ are H, TH, TTH, TTTH, and so forth. Let the probability set function P assign to these elements the respective probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, and so forth. Show that $P(\mathscr{C}) = 1$. Let $C_1 = \{c : c \text{ is H, TH, TTH, TTTH, or TTTTH}\}$. Compute $P(C_1)$. Let $C_2 = \{c : c \text{ is TTTTH or TTTTTH}\}$. Compute $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

1.20. If the sample space is $\mathscr{C} = C_1 \cup C_2$ and if $P(C_1) = 0.8$ and $P(C_2) = 0.5$, find $P(C_1 \cap C_2)$.

1.21. Let the sample space be $\mathscr{C} = \{c : 0 < c < \infty\}$. Let $C \subset \mathscr{C}$ be defined by $C = \{c : 4 < c < \infty\}$ and take $P(C) = \int_C e^{-x}\,dx$. Evaluate $P(C)$, $P(C^*)$, and $P(C \cup C^*)$.

1.22. If the sample space is $\mathscr{C} = \{c : -\infty < c < \infty\}$ and if $C \subset \mathscr{C}$ is a set for which the integral $\int_C e^{-|x|}\,dx$ exists, show that this set function is not a probability set function. What constant do we multiply the integral by to make it a probability set function?

1.23. If C_1 and C_2 are subsets of the sample space $\mathscr{C}$, show that

$$P(C_1 \cap C_2) \le P(C_1) \le P(C_1 \cup C_2) \le P(C_1) + P(C_2).$$

1.24. Let C_1, C_2, and C_3 be three mutually disjoint subsets of the sample space $\mathscr{C}$. Find $P[(C_1 \cup C_2) \cap C_3]$ and $P(C_1^* \cup C_2^*)$.

1.25. If C_1, C_2, and C_3 are subsets of $\mathscr{C}$, show that

$$P(C_1 \cup C_2 \cup C_3) = P(C_1) + P(C_2) + P(C_3) - P(C_1 \cap C_2) \\ - P(C_1 \cap C_3) - P(C_2 \cap C_3) + P(C_1 \cap C_2 \cap C_3).$$

What is the generalization of this result to four or more subsets of $\mathscr{C}$?

Hint: Write $P(C_1 \cup C_2 \cup C_3) = P[C_1 \cup (C_2 \cup C_3)]$ and use Theorem 5.

Remark. In order to solve a number of exercises, like 1.26–1.31, certain reasonable assumptions must be made.

1.26. A bowl contains 16 chips, of which 6 are red, 7 are white, and 3 are blue. If four chips are taken at random and without replacement, find the probability that: (a) each of the 4 chips is red; (b) none of the 4 chips is red; (c) there is at least 1 chip of each color.

1.27. A person has purchased 10 of 1000 tickets sold in a certain raffle. To determine the five prize winners, 5 tickets are to be drawn at random and without replacement. Compute the probability that this person will win at least one prize.

Hint: First compute the probability that the person does not win a prize.

1.28. Compute the probability of being dealt at random and without replacement a 13-card bridge hand consisting of: (a) 6 spades, 4 hearts, 2 diamonds, and 1 club; (b) 13 cards of the same suit.

1.29. Three distinct integers are chosen at random from the first 20 positive integers. Compute the probability that: (a) their sum is even; (b) their product is even.

1.30. There are 5 red chips and 3 blue chips in a bowl. The red chips are numbered 1, 2, 3, 4, 5, respectively, and the blue chips are numbered 1, 2, 3, respectively. If 2 chips are to be drawn at random and without replacement, find the probability that these chips have either the same number or the same color.

1.31. In a lot of 50 light bulbs, there are 2 bad bulbs. An inspector examines 5 bulbs, which are selected at random and without replacement.
(a) Find the probability of at least 1 defective bulb among the 5.
(b) How many bulbs should he examine so that the probability of finding at least 1 bad bulb exceeds $\frac{1}{2}$?

1.4 Conditional Probability and Independence

In some random experiments, we are interested only in those outcomes that are elements of a subset C_1 of the sample space $\mathscr{C}$. This means, for our purposes, that the sample space is effectively the subset C_1. We are now confronted with the problem of defining a probability set function with C_1 as the "new" sample space.

Let the probability set function $P(C)$ be defined on the sample space $\mathscr{C}$ and let C_1 be a subset of $\mathscr{C}$ such that $P(C_1) > 0$. We agree to consider only those outcomes of the random experiment that are elements of C_1; in essence, then, we take C_1 to be a sample space. Let C_2 be another subset of $\mathscr{C}$. How, relative to the new sample space C_1, do we want to define the probability of the event C_2? Once defined, this probability is called the *conditional probability* of the event C_2, relative to the hypothesis of the event C_1; or, more briefly, the conditional probability of C_2, given C_1. Such a conditional probability is denoted by the symbol $P(C_2|C_1)$. We now return to the question that was raised about the definition of this symbol. Since C_1 is now the sample space, the only elements of C_2 that concern us are those, if any, that are also elements of C_1, that is, the elements of $C_1 \cap C_2$. It seems desirable, then, to define the symbol $P(C_2|C_1)$ in such a way that

$$P(C_1|C_1) = 1 \qquad \text{and} \qquad P(C_2|C_1) = P(C_1 \cap C_2|C_1).$$

Moreover, from a relative frequency point of view, it would seem logically inconsistent if we did not require that the ratio of the probabilities of the events $C_1 \cap C_2$ and C_1, relative to the space C_1, be the same as the ratio of the probabilities of these events relative to the space $\mathscr{C}$; that is, we should have

$$\frac{P(C_1 \cap C_2|C_1)}{P(C_1|C_1)} = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

These three desirable conditions imply that the relation

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}$$

is a suitable *definition* of the conditional probability of the event C_2, given the event C_1, provided that $P(C_1) > 0$. Moreover, we have

1. $P(C_2|C_1) \geq 0$.
2. $P(C_2 \cup C_3 \cup \cdots |C_1) = P(C_2|C_1) + P(C_3|C_1) + \cdots$, provided that $C_2, C_3, \ldots$ are mutually disjoint sets.
3. $P(C_1|C_1) = 1$.

Properties (1) and (3) are evident; proof of property (2) is left as an exercise (1.32). But these are precisely the conditions that a probability set function must satisfy. Accordingly, $P(C_2|C_1)$ is a probability set function, defined for subsets of C_1. It may be called the conditional probability set function, relative to the hypothesis C_1; or the conditional probability set function, given C_1. It should be noted that this conditional probability set function, given C_1, is defined at this time only when $P(C_1) > 0$.

Example 1. A hand of 5 cards is to be dealt at random without replacement from an ordinary deck of 52 playing cards. The conditional probability of an all-spade hand (C_2), relative to the hypothesis that there are at least 4 spades in the hand (C_1), is, since $C_1 \cap C_2 = C_2$,

$$P(C_2|C_1) = \frac{P(C_2)}{P(C_1)} = \frac{\binom{13}{5}\Big/\binom{52}{5}}{\left[\binom{13}{4}\binom{39}{1} + \binom{13}{5}\right]\Big/\binom{52}{5}}$$

$$= \frac{\binom{13}{5}}{\binom{13}{4}\binom{39}{1} + \binom{13}{5}}.$$

From the definition of the conditional probability set function, we observe that

$$P(C_1 \cap C_2) = P(C_1)P(C_2|C_1).$$

This relation is frequently called the *multiplication rule* for probabilities. Sometimes, after considering the nature of the random experiment, it is possible to make reasonable assumptions so that both $P(C_1)$ and $P(C_2|C_1)$ can be assigned. Then $P(C_1 \cap C_2)$ can be computed under these assumptions. This will be illustrated in Examples 2 and 3.

Example 2. A bowl contains eight chips. Three of the chips are red and the remaining five are blue. Two chips are to be drawn successively, at random and without replacement. We want to compute the probability that the first draw results in a red chip (C_1) and that the second draw results in a blue chip (C_2). It is reasonable to assign the following probabilities:

$$P(C_1) = \tfrac{3}{8} \quad \text{and} \quad P(C_2|C_1) = \tfrac{5}{7}.$$

Thus, under these assignments, we have $P(C_1 \cap C_2) = (\frac{3}{8})(\frac{5}{7}) = \frac{15}{56}$.

Example 3. From an ordinary deck of playing cards, cards are to be drawn successively, at random and without replacement. The probability that the third spade appears on the sixth draw is computed as follows. Let C_1 be the event of two spades in the first five draws and let C_2 be the event of a spade on the sixth draw. Thus the probability that we wish to compute is $P(C_1 \cap C_2)$. It is reasonable to take

$$P(C_1) = \frac{\binom{13}{2}\binom{39}{3}}{\binom{52}{5}} \quad \text{and} \quad P(C_2|C_1) = \frac{11}{47}.$$

The desired probability $P(C_1 \cap C_2)$ is then the product of these two numbers.

The multiplication rule can be extended to three or more events. In the case of three events, we have, by using the multiplication rule for two events,

$$\begin{aligned} P(C_1 \cap C_2 \cap C_3) &= P[(C_1 \cap C_2) \cap C_3] \\ &= P(C_1 \cap C_2)P(C_3|C_1 \cap C_2). \end{aligned}$$

But $P(C_1 \cap C_2) = P(C_1)P(C_2|C_1)$. Hence

$$P(C_1 \cap C_2 \cap C_3) = P(C_1)P(C_2|C_1)P(C_3|C_1 \cap C_2).$$

This procedure can be used to extend the multiplication rule to four or more events. The general formula for k events can be proved by mathematical induction.

Example 4. Four cards are to be dealt successively, at random and without replacement, from an ordinary deck of playing cards. The probability of receiving a spade, a heart, a diamond, and a club, in that order, is $(\frac{13}{52})(\frac{13}{51})(\frac{13}{50})(\frac{13}{49})$. This follows from the extension of the multiplication rule. In this computation, the assumptions that are involved seem clear.

Let the space $\mathscr{C}$ be partitioned into k mutually exclusive and exhaustive events $C_1, C_2, \ldots, C_k$ such that $P(C_i) > 0$, $i = 1, 2, \ldots, k$. Here the events $C_1, C_2, \ldots, C_k$ do *not* need to be equally likely. Let C be another event such that $P(C) > 0$. Thus C occurs with one and only one of the events $C_1, C_2, \ldots, C_k$; that is,

$$\begin{aligned} C &= C \cap (C_1 \cup C_2 \cup \cdots \cup C_k) \\ &= (C \cap C_1) \cup (C \cap C_2) \cup \cdots \cup (C \cap C_k). \end{aligned}$$

Since $C \cap C_i$, $i = 1, 2, \ldots, k$, are mutually exclusive, we have

$$P(C) = P(C \cap C_1) + P(C \cap C_2) + \cdots + P(C \cap C_k).$$

However, $P(C \cap C_i) = P(C_i)P(C|C_i)$, $i = 1, 2, \ldots, k$; so

$$P(C) = P(C_1)P(C|C_1) + P(C_2)P(C|C_2) + \cdots + P(C_k)P(C|C_k)$$
$$= \sum_{i=1}^{k} P(C_i)P(C|C_i).$$

This result is sometimes called the *law of total probability*.

From the definition of conditional probability, we have, using the law of total probability, that

$$P(C_j|C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C_j)P(C|C_j)}{\sum_{i=1}^{k} P(C_i)P(C|C_i)},$$

which is the well-known *Bayes' theorem*. This permits us to calculate the conditional probability of C_j, given C, from the probabilities of $C_1, C_2, \ldots, C_k$ and the conditional probabilities of C, given C_i, $i = 1, 2, \ldots, k$.

Example 5. Say it is known that bowl C_1 contains 3 red and 7 blue chips and bowl C_2 contains 8 red and 2 blue chips. All chips are identical in size and shape. A die is cast and bowl C_1 is selected if five or six spots show on the side that is up; otherwise, bowl C_2 is selected. In a notation that is fairly obvious, it seems reasonable to assign $P(C_1) = \frac{2}{6}$ and $P(C_2) = \frac{4}{6}$. The selected bowl is handed to another person and one chip is taken at random. Say that this chip is red, an event which we denote by C. By considering the contents of the bowls, it is reasonable to assign the conditional probabilities $P(C|C_1) = \frac{3}{10}$ and $P(C|C_2) = \frac{8}{10}$. Thus the conditional probability of bowl C_1, given that a red chip is drawn, is

$$P(C_1|C) = \frac{P(C_1)P(C|C_1)}{P(C_1)P(C|C_1) + P(C_2)P(C|C_2)}$$
$$= \frac{(\frac{2}{6})(\frac{3}{10})}{(\frac{2}{6})(\frac{3}{10}) + (\frac{4}{6})(\frac{8}{10})} = \frac{3}{19}.$$

In a similar manner, we have $P(C_2|C) = \frac{16}{19}$.

In Example 5, the probabilities $P(C_1) = \frac{2}{6}$ and $P(C_2) = \frac{4}{6}$ are called *prior probabilities* of C_1 and C_2, respectively, because they are known to be due to the random mechanism used to select the bowls. After the chip is taken and observed to be red, the conditional probabilities $P(C_1|C) = \frac{3}{19}$ and $P(C_2|C) = \frac{16}{19}$ are called *posterior probabilities*. Since C_2 has a larger proportion of red chips than does C_1, it appeals to one's intuition that $P(C_2|C)$ should be larger than $P(C_2)$ and, of course, $P(C_1|C)$ should be smaller than $P(C_1)$. That is, intuitively the chances of having bowl C_2 are better once that a red chip is observed than before a chip is taken. Bayes' theorem provides a method of determining exactly what those probabilities are.

Example 6. Three plants, C_1, C_2, and C_3, produce respectively, 10, 50, and 40 percent of a company's output. Although plant C_1 is a small plant, its manager believes in high quality and only 1 percent of its products are defective. The other two, C_2 and C_3, are worse and produce items that are 3 and 4 percent defective, respectively. All products are sent to a central warehouse. One item is selected at random and observed to be defective, say event C. The conditional probability that it comes from plant C_1 is found as follows. It is natural to assign the respective prior probabilities of getting an item from the plants as $P(C_1) = 0.1$, $P(C_2) = 0.5$, and $P(C_3) = 0.4$, while the conditional probabilities of defective are $P(C|C_1) = 0.01$, $P(C|C_2) = 0.03$, and $P(C|C_3) = 0.04$. Thus the posterior probability of C_1, given a defective, is

$$P(C_1|C) = \frac{P(C_1 \cap C)}{P(C)} = \frac{(0.10)(0.01)}{(0.10)(0.01) + (0.50)(0.03) + (0.40)(0.04)},$$

which equals $\frac{1}{32}$; this is much smaller than the prior probability $P(C_1) = \frac{1}{10}$. This is as it should be because the fact that the item is defective decreases the chances that it comes from the high-quality plant C_1.

Sometimes it happens that the occurrence of event C_1 does not change the probability of event C_2; that is, when $P(C_1) > 0$,

$$P(C_2|C_1) = P(C_2).$$

In this case, we say that the events C_1 and C_2 are *independent*. Moreover, the multiplication rule becomes

$$P(C_1 \cap C_2) = P(C_1)P(C_2|C_1) = P(C_1)P(C_2).$$

This, in turn, implies, when $P(C_2) > 0$, that

$$P(C_1|C_2) = \frac{P(C_1 \cap C_2)}{P(C_2)} = \frac{P(C_1)P(C_2)}{P(C_2)} = P(C_1).$$

Remark. Events that are *independent* are sometimes called *statistically independent*, *stochastically independent*, or *independent in a probability sense*. In most instances, we use independent without a modifier if there is no possibility of misunderstanding.

It is interesting to note that C_1 and C_2 are independent if $P(C_1) = 0$ or $P(C_2) = 0$ because then $P(C_1 \cap C_2) = 0$ since $(C_1 \cap C_2) \subset C_1$ and $(C_1 \cap C_2) \subset C_2$. Thus the left- and right-hand members of

$$P(C_1 \cap C_2) = P(C_1)P(C_2)$$

are both equal to zero and are, of course, equal to each other. Also, if C_1 and C_2 are independent events, so are the three pairs: C_1 and C_2^*, C_1^* and C_2, and C_1^* and C_2^* (see Exercise 1.41).

Example 7. A red die and a white die are cast in such a way that the number of spots on the two sides that are up are independent events. If C_1 represents a four on the red die and C_2 represents a three on the white die, with an equally likely assumption for each side, we assign $P(C_1) = \frac{1}{6}$ and $P(C_2) = \frac{1}{6}$. Thus, from independence, the probability of the ordered pair (red = 4, white = 3) is

$$P[(4, 3)] = (\tfrac{1}{6})(\tfrac{1}{6}) = \tfrac{1}{36}.$$

The probability that the sum of the up spots of the two dice equals seven is

$$\begin{aligned} &P[(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)] \\ &\quad = (\tfrac{1}{6})(\tfrac{1}{6}) + (\tfrac{1}{6})(\tfrac{1}{6}) + (\tfrac{1}{6})(\tfrac{1}{6}) + (\tfrac{1}{6})(\tfrac{1}{6}) + (\tfrac{1}{6})(\tfrac{1}{6}) + (\tfrac{1}{6})(\tfrac{1}{6}) = \tfrac{6}{36}. \end{aligned}$$

In a similar manner, it is easy to show that the probabilities of the sums of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 are, respectively,

$$\tfrac{1}{36}, \tfrac{2}{36}, \tfrac{3}{36}, \tfrac{4}{36}, \tfrac{5}{36}, \tfrac{6}{36}, \tfrac{5}{36}, \tfrac{4}{36}, \tfrac{3}{36}, \tfrac{2}{36}, \tfrac{1}{36}.$$

Suppose now that we have three events, C_1, C_2, and C_3. We say that they are *mutually independent* if and only if they are *pairwise independent*:

$$P(C_1 \cap C_3) = P(C_1)P(C_3), \qquad P(C_1 \cap C_2) = P(C_1)P(C_2),$$

$$P(C_2 \cap C_3) = P(C_2)P(C_3)$$

and

$$P(C_1 \cap C_2 \cap C_3) = P(C_1)P(C_2)P(C_3).$$

More generally, the n events $C_1, C_2, \ldots, C_n$ are *mutually independent* if and only if for every collection of k of these events, $2 \le k \le n$, the following is true:

Say that $d_1, d_2, \ldots, d_k$ are k distinct integers from $1, 2, \ldots, n$; then

$$P(C_{d_1} \cap C_{d_2} \cap \cdots \cap C_{d_k}) = P(C_{d_1})P(C_{d_2}) \cdots P(C_{d_k}).$$

In particular, if $C_1, C_2, \ldots, C_n$ are mutually independent, then

$$P(C_1 \cap C_2 \cap \cdots \cap C_n) = P(C_1)P(C_2) \cdots P(C_n).$$

Also, as with two sets, many combinations of these events and their complements are independent, such as

$$C_1^* \text{ and } (C_2 \cup C_3^* \cup C_4) \qquad \text{are independent;}$$

$$C_1 \cup C_2^*, \quad C_3^*, \quad \text{and} \quad C_4 \cap C_5^* \qquad \text{are mutually independent.}$$

If there is no possibility of misunderstanding, *independent* is often used without the modifier *mutually* when considering more than two events.

We often perform a sequence of random experiments in such a way that the events associated with one of them are independent of the events associated with the others. For convenience, we refer to these events as *independent experiments*, meaning that the respective events are independent. Thus we often refer to independent flips of a coin or independent casts of a die or—more generally—independent trials of some given random experiment.

Example 8. A coin is flipped independently several times. Let the event C_i represent a head (H) on the ith toss; thus C_i^* represents a tail (T). Assume that C_i and C_i^* are equally likely; that is, $P(C_i) = P(C_i^*) = \frac{1}{2}$. Thus the probability of an ordered sequence like HHTH is, from independence,

$$P(C_1 \cap C_2 \cap C_3^* \cap C_4) = P(C_1)P(C_2)P(C_3^*)P(C_4) = (\tfrac{1}{2})^4 = \tfrac{1}{16}.$$

Similarly, the probability of observing the first head on the third flip is

$$P(C_1^* \cap C_2^* \cap C_3) = P(C_1^*)P(C_2^*)P(C_3) = (\tfrac{1}{2})^3 = \tfrac{1}{8}.$$

Also, the probability of getting at least one head on four flips is

$$\begin{aligned} P(C_1 \cup C_2 \cup C_3 \cup C_4) &= 1 - P[(C_1 \cup C_2 \cup C_3 \cup C_4)^*] \\ &= 1 - P(C_1^* \cap C_2^* \cap C_3^* \cap C_4^*) \\ &= 1 - (\tfrac{1}{2})^4 = \tfrac{15}{16}. \end{aligned}$$

See Exercise 1.43 to justify this last probability.

EXERCISES

1.32. If $P(C_1) > 0$ and if $C_2, C_3, C_4, \ldots$ are mutually disjoint sets, show that $P(C_2 \cup C_3 \cup \cdots | C_1) = P(C_2|C_1) + P(C_3|C_1) + \cdots$.

1.33. Prove that

$$P(C_1 \cap C_2 \cap C_3 \cap C_4) = P(C_1)P(C_2|C_1)P(C_3|C_1 \cap C_2)P(C_4|C_1 \cap C_2 \cap C_3).$$

1.34. A bowl contains 8 chips. Three of the chips are red and 5 are blue. Four chips are to be drawn successively at random and without replacement.
(a) Compute the probability that the colors alternate.

(b) Compute the probability that the first blue chip appears on the third draw.

1.35. A hand of 13 cards is to be dealt at random and without replacement from an ordinary deck of playing cards. Find the conditional probability that there are at least three kings in the hand relative to the hypothesis that the hand contains at least two kings.

1.36. A drawer contains eight pairs of socks. If six socks are taken at random and without replacement, compute the probability that there is at least one matching pair among these six socks.

Hint: Compute the probability that there is not a matching pair.

1.37. A bowl contains 10 chips. Four of the chips are red, 5 are white, and 1 is blue. If 3 chips are taken at random and without replacement, compute the conditional probability that there is 1 chip of each color relative to the hypothesis that there is exactly 1 red chip among the 3.

1.38. Bowl I contains 3 red chips and 7 blue chips. Bowl II contains 6 red chips and 4 blue chips. A bowl is selected at random and then 1 chip is drawn from this bowl.

(a) Compute the probability that this chip is red.

(b) Relative to the hypothesis that the chip is red, find the conditional probability that it is drawn from bowl II.

1.39. Bowl I contains 6 red chips and 4 blue chips. Five of these 10 chips are selected at random and without replacement and put in bowl II, which was originally empty. One chip is then drawn at random from bowl II. Relative to the hypothesis that this chip is blue, find the conditional probability that 2 red chips and 3 blue chips are transferred from bowl I to bowl II.

1.40. A professor of statistics has two boxes of computer disks: box C_1 contains seven Verbatim disks and three Control Data disks and box C_2 contains two Verbatim disks and eight Control Data disks. She selects a box at random with probabilities $P(C_1) = \frac{2}{3}$ and $P(C_2) = \frac{1}{3}$ because of their respective locations. A disk is then selected at random and the event C occurs if it is from Control Data. Using an equally likely assumption for each disk in the selected box, compute $P(C_1|C)$ and $P(C_2|C)$.

1.41. If C_1 and C_2 are independent events, show that the following pairs of events are also independent: (a) C_1 and C_2^*, (b) C_1^* and C_2, and (c) C_1^* and C_2^*.

Hint: In (a), write $P(C_1 \cap C_2^*) = P(C_1)P(C_2^*|C_1) = P(C_1)[1 - P(C_2|C_1)]$. From independence of C_1 and C_2, $P(C_2|C_1) = P(C_2)$.

1.42. Let C_1 and C_2 be independent events with $P(C_1) = 0.6$ and $P(C_2) = 0.3$. Compute (a) $P(C_1 \cap C_2)$; (b) $P(C_1 \cup C_2)$; (c) $P(C_1 \cup C_2^*)$.

1.43. Generalize Exercise 1.4 to obtain

$$(C_1 \cup C_2 \cup \cdots \cup C_k)^* = C_1^* \cap C_2^* \cap \cdots \cap C_k^*.$$

Say that $C_1, C_2, \ldots, C_k$ are independent events that have respective probabilities $p_1, p_2, \ldots, p_k$. Argue that the probability of at least one of $C_1, C_2, \ldots, C_k$ is equal to

$$1 - (1 - p_1)(1 - p_2) \cdots (1 - p_k).$$

1.44. Each of four persons fires one shot at a target. Let C_k denote the event that the target is hit by person k, $k = 1, 2, 3, 4$. If C_1, C_2, C_3, C_4 are independent and if $P(C_1) = P(C_2) = 0.7$, $P(C_3) = 0.9$, and $P(C_4) = 0.4$, compute the probability that (a) all of them hit the target; (b) exactly one hits the target; (c) no one hits the target; (d) at least one hits the target.

1.45. A bowl contains three red (R) balls and seven white (W) balls of exactly the same size and shape. Select balls successively at random and with replacement so that the events of white on the first trial, white on the second, and so on, can be assumed to be independent. In four trials, make certain assumptions and compute the probabilities of the following ordered sequences: (a) WWRW; (b) RWWW; (c) WWWR; and (d) WRWW. (e) Compute the probability of exactly one red ball in the four trials.

1.46. A coin is tossed two independent times, each resulting in a tail (T) or a head (H). The sample space consists of four ordered pairs: TT, TH, HT, HH. Making certain assumptions, compute the probability of each of these ordered pairs. What is the probability of at least one head?

1.5 Random Variables of the Discrete Type

The reader will perceive that a sample space $\mathscr{C}$ may be tedious to describe if the elements of $\mathscr{C}$ are not numbers. We shall now discuss how we may formulate a rule, or a set of rules, by which the elements c of $\mathscr{C}$ may be represented by numbers. We begin the discussion with a very simple example. Let the random experiment be the toss of a coin and let the sample space associated with the experiment be $\mathscr{C} = \{c : \text{where } c \text{ is T or } c \text{ is H}\}$ and T and H represent, respectively, tails and heads. Let X be a function such that $X(c) = 0$ if c is T and let $X(c) = 1$ if c is H. Thus X is a real-valued function defined on the sample space $\mathscr{C}$ which takes us from the sample space $\mathscr{C}$ to a space of real numbers $\mathscr{A} = \{0, 1\}$. We call X a random variable and, in this example, the space associated with X is $\mathscr{A} = \{0, 1\}$. We now formulate the definition of a random variable and its space.

Definition 8. Consider a random experiment with a sample space $\mathscr{C}$. A function X, which assigns to each element $c \in \mathscr{C}$ one and only one real number $X(c) = x$, is called a *random variable*. The *space* of X is the set of real numbers $\mathscr{A} = \{x : x = X(c), c \in \mathscr{C}\}$.

It may be that the set $\mathscr{C}$ has elements which are themselves real numbers. In such an instance we could write $X(c) = c$ so that $\mathscr{A} = \mathscr{C}$.

Let X be a random variable that is defined on a sample space $\mathscr{C}$, and let $\mathscr{A}$ be the space of X. Further, let A be a subset of $\mathscr{A}$. Just as we used the terminology "the event C," with $C \subset \mathscr{C}$, we shall now speak of "the event A." The probability $P(C)$ of the event C has been defined. We wish now to define the probability of the event A. This probability will be denoted by $\Pr(X \in A)$, where Pr is an abbreviation for "the probability that." With A a subset of $\mathscr{A}$, let C be that subset of $\mathscr{C}$ such that $C = \{c : c \in \mathscr{C} \text{ and } X(c) \in A\}$. Thus C has as its elements all outcomes in $\mathscr{C}$ for which the random variable X has a value that is in A. This prompts us to define, as we now do, $\Pr(X \in A)$ to be equal to $P(C)$, where $C = \{c : c \in \mathscr{C} \text{ and } X(c) \in A\}$. Thus $\Pr(X \in A)$ is an assignment of probability to a set A, which is a subset of the space $\mathscr{A}$ associated with the random variable X. This assignment is determined by the probability set function P and the random variable X and is sometimes denoted by $P_X(A)$. That is,

$$\Pr(X \in A) = P_X(A) = P(C),$$

where $C = \{c : c \in \mathscr{C} \text{ and } X(c) \in A\}$. Thus a random variable X is a function that carries the probability from a sample space $\mathscr{C}$ to a space $\mathscr{A}$ of real numbers. In this sense, with $A \subset \mathscr{A}$, the probability $P_X(A)$ is often called an *induced probability*.

Remark. In a more advanced course, it would be noted that the random variable X is a Borel measurable function. This is needed to assure that we can find the induced probabilities on the sigma field of the subsets of $\mathscr{A}$. We need this requirement throughout this book for every function that is a random variable, but no further mention of it is made.

The function $P_X(A)$ satisfies the conditions (a), (b), and (c) of the definition of a probability set function (Section 1.3). That is, $P_X(A)$ is also a probability set function. Conditions (a) and (c) are easily verified by observing, for an appropriate C, that

$$P_X(A) = P(C) \geq 0,$$

and that $\mathscr{C} = \{c : c \in \mathscr{C} \text{ and } X(c) \in \mathscr{A}\}$ requires

$$P_X(\mathscr{A}) = P(\mathscr{C}) = 1.$$

In discussing the condition (b), let us restrict our attention to the two mutually exclusive events A_1 and A_2. Here $P_X(A_1 \cup A_2) = P(C)$, where $C = \{c : c \in \mathscr{C} \text{ and } X(c) \in A_1 \cup A_2\}$. However,

$$C = \{c : c \in \mathscr{C} \text{ and } X(c) \in A_1\} \cup \{c : c \in \mathscr{C} \text{ and } X(c) \in A_2\},$$

or, for brevity, $C = C_1 \cup C_2$. But C_1 and C_2 are disjoint sets. This must be so, for if some c were common, say c_i, then $X(c_i) \in A_1$ and $X(c_i) \in A_2$. That is, the same number $X(c_i)$ belongs to both A_1 and A_2. This is a contradiction because A_1 and A_2 are disjoint sets. Accordingly,

$$P(C) = P(C_1) + P(C_2).$$

However, by definition, $P(C_1)$ is $P_X(A_1)$ and $P(C_2)$ is $P_X(A_2)$ and thus

$$P_X(A_1 \cup A_2) = P_X(A_1) + P_X(A_2).$$

This is condition (b) for two disjoint sets.

Thus each of $P_X(A)$ and $P(C)$ is a probability set function. But the reader should fully recognize that the probability set function P is defined for subsets C of $\mathscr{C}$, whereas P_X is defined for subsets A of $\mathscr{A}$, and, in general, they are not the same set function. Nevertheless, they are closely related and some authors even drop the index X and write $P(A)$ for $P_X(A)$. They think it is quite clear that $P(A)$ means the probability of A, a subset of $\mathscr{A}$, and $P(C)$ means the probability of C, a subset of $\mathscr{C}$. From this point on, we shall adopt this convention and simply write $P(A)$.

Perhaps an additional example will be helpful. Let a coin be tossed two independent times and let our interest be in the *number* of heads to be observed. Thus the sample space is $\mathscr{C} = \{c : \text{where } c \text{ is TT or TH or HT or HH}\}$. Let $X(c) = 0$ if c is TT; let $X(c) = 1$ if c is either TH or HT; and let $X(c) = 2$ if c is HH. Thus the space of the random variable X is $\mathscr{A} = \{0, 1, 2\}$. Consider the subset A of the space $\mathscr{A}$, where $A = \{1\}$. How is the probability of the event A defined? We take the subset C of $\mathscr{C}$ to have as its elements all outcomes in $\mathscr{C}$ for which the random variable X has a value that is an element of A. Because $X(c) = 1$ if c is either TH or HT, then $C = \{c : \text{where } c \text{ is TH or HT}\}$. Thus $P(A) = \Pr(X \in A) = P(C)$. Since $A = \{1\}$, then $P(A) = \Pr(X \in A)$ can be written more simply as $\Pr(X = 1)$. Let $C_1 = \{c : c \text{ is TT}\}$, $C_2 = \{c : c \text{ is TH}\}$, $C_3 = \{c : c \text{ is HT}\}$, and $C_4 = \{c : c \text{ is HH}\}$ denote subsets of $\mathscr{C}$. From independence and equally likely assumptions (see Exercise 1.46), our probability set function $P(C)$ assigns a probability of $\frac{1}{4}$ to each of the sets C_i, $i = 1, 2, 3, 4$. Then $P(C_1) = \frac{1}{4}$, $P(C_2 \cup C_3) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, and $P(C_4) = \frac{1}{4}$. Let us now point out how much simpler it is to couch these statements in a language that involves the random variable X. Because X is the number of heads to be observed in tossing a coin two times, we have

$$\Pr(X = 0) = \tfrac{1}{4}, \qquad \text{since } P(C_1) = \tfrac{1}{4};$$

$$\Pr(X = 1) = \tfrac{1}{2}, \qquad \text{since } P(C_2 \cup C_3) = \tfrac{1}{2};$$

and

$$\Pr(X = 2) = \tfrac{1}{4}, \qquad \text{since } P(C_4) = \tfrac{1}{4}.$$

This may be further condensed in the following table:

x	0	1	2
$\Pr(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

This table depicts the distribution of probability over the elements of $\mathscr{A}$, the space of the random variable X. This can be written more simply as

$$\Pr(X = x) = \binom{2}{x}\left(\frac{1}{2}\right)^2, \qquad x \in \mathscr{A}.$$

Example 1. Consider a sequence of independent flips of a coin, each resulting in a head (H) or a tail (T). Moreover, on each flip, we assume that H and T are equally likely, that is, $P(\text{H}) = P(\text{T}) = \frac{1}{2}$. The sample space $\mathscr{C}$ consists of sequences like TTHTHHT $\cdots$. Let the random variable X equal the number of flips needed to obtain the first head. For this given sequence, $X = 3$. Clearly, the space of X is $\mathscr{A} = \{1, 2, 3, 4, \ldots\}$. We see that $X = 1$ when the sequence begins with an H and thus $\Pr(X = 1) = \frac{1}{2}$. Likewise, $X = 2$ when the sequence begins with TH, which has probability $\Pr(X = 2) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ from the independence. More generally, if $X = x$, where $x = 1, 2, 3, 4, \ldots$, there must be a string of $x - 1$ tails followed by a head, that is, TT $\cdots$ TH, where there are $x - 1$ tails in TT $\cdots$ T. Thus, from independence, we have

$$\Pr(X = x) = \left(\frac{1}{2}\right)^{x-1}\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^x, \qquad x = 1, 2, 3, \ldots.$$

Let us make some observations about these three illustrations of a random variable. In each case the number of points in the space $\mathscr{A}$ was finite, as with $\{0, 1\}$ and $\{0, 1, 2\}$, or countable, as with $\{1, 2, 3, \ldots\}$. There was a function, say $f(x) = \Pr(X = x)$, that described how the probability is distributed over the space $\mathscr{A}$. In each of these illustrations, there is a simple formula (although that is not necessary in general) for that function, namely:

$$f(x) = \frac{1}{2}, \qquad x \in \{0, 1\},$$

$$f(x) = \binom{2}{x}\left(\frac{1}{2}\right)^2, \qquad x \in \{0, 1, 2\},$$

and

$$f(x) = \left(\frac{1}{2}\right)^x, \qquad x \in \{1, 2, 3, \dots\}.$$

Moreover, the sum of $f(x)$ over all $x \in \mathscr{A}$ equals 1:

$$\sum_{x=0}^{1} \left(\frac{1}{2}\right) = \frac{1}{2} + \frac{1}{2} = 1,$$

$$\sum_{x=0}^{2} \binom{2}{x}\left(\frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1,$$

$$\sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \cdots = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1.$$

Finally, if $A \subset \mathscr{A}$, we can compute the probability of $X \in A$ by the summation

$$\Pr(X \in A) = \sum_A f(x).$$

For illustrations, using the random variable of Example 1,

$$\Pr(X = 1, 2, 3) = \sum_{x=1}^{3} \left(\frac{1}{2}\right)^x = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$$

and

$$\Pr(X = 1, 3, 5, \dots) = \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \cdots$$

$$= \frac{\frac{1}{2}}{1 - \frac{1}{4}} = \frac{2}{3}.$$

We have special names for this type of random variable X and for a function $f(x)$ like that in each of these three illustrations, which we now give.

Let X denote a random variable with a one-dimensional space $\mathscr{A}$. Suppose that $\mathscr{A}$ consists of a countable number of points; that is, $\mathscr{A}$ contains a finite number of points or the points of $\mathscr{A}$ can be put into a one-to-one correspondence with the positive integers. Such a space

$\mathscr{A}$ is called a *discrete set* of points. Let a function $f(x)$ be such that $f(x) > 0$, $x \in \mathscr{A}$, and

$$\sum_{\mathscr{A}} f(x) = 1.$$

Whenever a probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed in terms of such an $f(x)$ by

$$P(A) = \Pr(X \in A) = \sum_{A} f(x),$$

then X is called a *random variable of the discrete type* and $f(x)$ is called the *probability density function* of X. Hereafter the *probability density function* is abbreviated p.d.f.

Our notation can be simplified somewhat so that we do not need to spell out the space in each instance. For illustration, let the random variable be the number of flips necessary to obtain the first head. We now extend the definition of the p.d.f. from on $\mathscr{A} = \{1, 2, 3, \ldots\}$ to all the real numbers by writing

$$\begin{aligned} f(x) &= \left(\frac{1}{2}\right)^x, \qquad x = 1, 2, 3, \ldots, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

From such a function, we see that the space $\mathscr{A}$ is clearly the set of positive integers which is a discrete set of points. Thus the corresponding random variable is one of the discrete type.

Example 2. A lot, consisting of 100 fuses, is inspected by the following procedure. Five of these fuses are chosen at random and tested; if all 5 "blow" at the correct amperage, the lot is accepted. If, in fact, there are 20 defective fuses in the lot, the probability of accepting the lot is, under appropriate assumptions,

$$\frac{\binom{80}{5}}{\binom{100}{5}} = 0.32,$$

approximately. More generally, let the random variable X be the number of defective fuses among the 5 that are inspected. The p.d.f. of X is given by

$$\begin{aligned} f(x) = \Pr(X = x) &= \frac{\binom{20}{x}\binom{80}{5-x}}{\binom{100}{5}}, \qquad x = 0, 1, 2, 3, 4, 5, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Clearly, the space of X is $\mathscr{A} = \{0, 1, 2, 3, 4, 5\}$. Thus this is an example of a random variable of the discrete type whose distribution is an illustration of a *hypergeometric distribution*.

Let the random variable X have the probability set function $P(A)$, where A is a one-dimensional set. Take x to be a real number and consider the set A which is an unbounded set from $-\infty$ to x, including the point x itself. For all such sets A we have $P(A) = \Pr(X \in A) = \Pr(X \le x)$. This probability depends on the point x; that is, this probability is a function of the point x. This point function is denoted by the symbol $F(x) = \Pr(X \le x)$. The function $F(x)$ is called the *distribution function* (sometimes, *cumulative distribution function*) of the random variable X. Since $F(x) = \Pr(X \le x)$, then, with $f(x)$ the p.d.f., we have

$$F(x) = \sum_{w \le x} f(w),$$

for the discrete type.

Example 3. Let the random variable X of the discrete type have the p.d.f. $f(x) = x/6$, $x = 1, 2, 3$, zero elsewhere. The distribution function of X is

$$\begin{aligned} F(x) &= 0, & x &< 1, \\ &= \tfrac{1}{6}, & 1 &\le x < 2, \\ &= \tfrac{3}{6}, & 2 &\le x < 3, \\ &= 1, & 3 &\le x. \end{aligned}$$

Here, as depicted in Figure 1.3, $F(x)$ is a step function that is constant in every interval not containing 1, 2, or 3, but has steps of heights $\frac{1}{6}$, $\frac{2}{6}$, and $\frac{3}{6}$, which are the probabilities at those respective points. It is also seen that $F(x)$ is everywhere continuous from the right. The p.d.f. of X is displayed as a bar

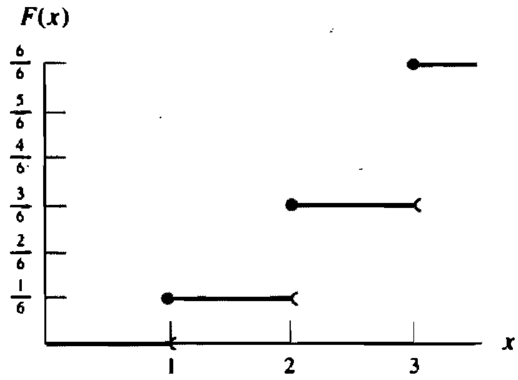

FIGURE 1.3

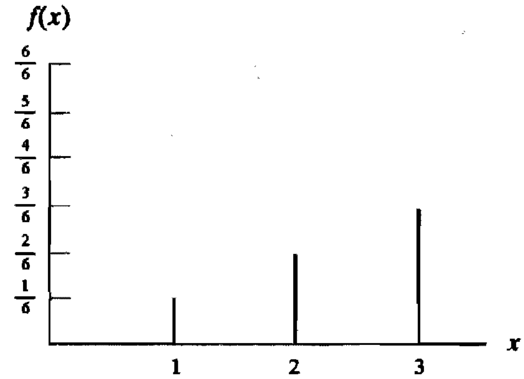

FIGURE 1.4

graph in Figure 1.4. We see that $f(x)$ represents the probability at each x while $F(x)$ cumulates all the probability of points that are less than or equal to x. Thus we can compute a probability like

$$\Pr(1.5 < X \le 4.5) = F(4.5) - F(1.5) = 1 - \tfrac{1}{6} = \tfrac{5}{6}$$

or as

$$\Pr(1.5 < X \le 4.5) = f(2) + f(3) = \tfrac{2}{6} + \tfrac{3}{6} = \tfrac{5}{6}.$$

While the properties of a *distribution function* $F(x) = \Pr(X \le x)$ are discussed in more detail in Section 1.7, we can make a few observations now since $F(x)$ is a probability.

1. $0 \le F(x) \le 1$.
2. $F(x)$ is a nondecreasing function as it cumulates probability as x increases.
3. $F(y) = 0$ for every point y that is less than the smallest value in the space of X.
4. $F(z) = 1$ for every point z that is greater than the largest value in the space of X.
5. If X is a random variable of the discrete type, then $F(x)$ is a step function and the height of the step at x in the space of X is equal to the probability $f(x) = \Pr(X = x)$.

EXERCISES

1.47. Let a card be selected from an ordinary deck of playing cards. The outcome c is one of these 52 cards. Let $X(c) = 4$ if c is an ace, let $X(c) = 3$ if c is a king, let $X(c) = 2$ if c is a queen, let $X(c) = 1$ if c is a jack, and let $X(c) = 0$ otherwise. Suppose that P assigns a probability of $\frac{1}{52}$ to each outcome c. Describe the induced probability $P_X(A)$ on the space $\mathscr{A} = \{0, 1, 2, 3, 4\}$ of the random variable X.

1.48. For each of the following, find the constant c so that $f(x)$ satisfies the condition of being a p.d.f. of one random variable X.
(a) $f(x) = c(\frac{2}{3})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.
(b) $f(x) = cx$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere.

1.49. Let $f(x) = x/15$, $x = 1, 2, 3, 4, 5$, zero elsewhere, be the p.d.f. of X. Find $\Pr(X = 1 \text{ or } 2)$, $\Pr(\frac{1}{2} < X < \frac{5}{2})$, and $\Pr(1 \le X \le 2)$.

1.50. Let $f(x)$ be the p.d.f. of a random variable X. Find the distribution function $F(x)$ of X and sketch its graph along with that of $f(x)$ if:
(a) $f(x) = 1$, $x = 0$, zero elsewhere.
(b) $f(x) = \frac{1}{3}$, $x = -1, 0, 1$, zero elsewhere.
(c) $f(x) = x/15$, $x = 1, 2, 3, 4, 5$, zero elsewhere.

1.51. Let us select five cards at random and without replacement from an ordinary deck of playing cards.
(a) Find the p.d.f. of X, the number of hearts in the five cards.
(b) Determine $\Pr(X \le 1)$.

1.52. Let X equal the number of heads in four independent flips of a coin. Using certain assumptions, determine the p.d.f. of X and compute the probability that X is equal to an odd number.

1.53. Let X have the p.d.f. $f(x) = x/5050$, $x = 1, 2, 3, \ldots, 100$, zero elsewhere.
(a) Compute $\Pr(X \le 50)$.
(b) Show that the distribution function of X is $F(x) = [x]([x] + 1)/10100$, for $1 \le x \le 100$, where $[x]$ is the greatest integer in x.

1.54. Let a bowl contain 10 chips of the same size and shape. One and only one of these chips is red. Continue to draw chips from the bowl, one at a time and at random and without replacement, until the red chip is drawn.
(a) Find the p.d.f. of X, the number of trials needed to draw the red chip.
(b) Compute $\Pr(X \le 4)$.

1.55. Cast a die a number of independent times until a six appears on the up side of the die.
(a) Find the p.d.f. $f(x)$ of X, the number of casts needed to obtain that first six.
(b) Show that $\sum_{x=1}^{\infty} f(x) = 1$.
(c) Determine $\Pr(X = 1, 3, 5, 7, \ldots)$.
(d) Find the distribution function $F(x) = \Pr(X \le x)$.

1.56. Cast a die two independent times and let X equal the absolute value of the difference of the two resulting values (the numbers on the up sides). Find the p.d.f. of X.

Hint: It is not necessary to find a formula for the p.d.f.

1.6 Random Variables of the Continuous Type

A random variable was defined in Section 1.5, and only those of the discrete type were considered there. Let us begin the discussion of random variables of the continuous type with an example.

Let a random experiment be a selection of a point that is interior to a circle of radius 1 that has center at the origin of a two-dimensional space. We call this space $\mathscr{C}$ and the area of this circle is π. The random selection is in such a way that the probability of being in a certain set C interior to $\mathscr{C}$ is proportional to the area of C; in particular, if $C \subset \mathscr{C}$,

$$P(C) = \frac{\text{area of } C}{\pi}.$$

First we observe that $P(\mathscr{C}) = 1$. In addition, if C_1 is that subset of $\mathscr{C}$ that is in the first quadrant, $P(C_1) = (\pi/4)/\pi = \frac{1}{4}$. If C_2 is the interior of a circle of radius $\frac{1}{3}$ such that $C_2 \subset \mathscr{C}$, then $P(C_2) = \pi(\frac{1}{3})^2/\pi = \frac{1}{9}$. It is interesting to note that the probability of a point, a line segment, or any curve in $\mathscr{C}$ is equal to zero because those areas would be zero. In particular, if C_3 is the boundary of the set C_2 (that is, C_3 is the actual circle of radius $\frac{1}{3}$), then $P(C_3) = 0$.

We define a random variable X, associated with this random experiment, as the distance of the selected point from the origin. The space of X is $\mathscr{A} = \{x : 0 \le x < 1\}$. Of course, for any $x \in \mathscr{A}$, $\Pr(X = x) = 0$, because $X = x$ is the event that the random point falls on a circle, symmetric with respect to the origin, of radius x and the associated area equals zero. However, it does make sense to consider the induced probability of the event $X \le x$, namely the distribution function of X. If $x \in \mathscr{A}$, then

$$F(x) = \Pr(X \le x) = \frac{\text{area of a certain circle of radius } x}{\pi}$$

$$= \frac{\pi x^2}{\pi} = x^2, \qquad 0 \le x < 1.$$

Clearly, if $x < 0$, then $F(x) = 0$; and if $x > 1$, then $F(x) = 1$. Thus we can write

$$\begin{aligned} F(x) &= 0, & x &< 0, \\ &= x^2, & 0 \le x &< 1, \\ &= 1, & 1 &\le x. \end{aligned}$$

Recall, in the discrete case, we had a function f that was associated with F through the equation

$$F(x) = \sum_{w \le x} f(w).$$

Either F or f could be used to compute probabilities like

$$\Pr(a < X \le b) = F(b) - F(a) = \sum_{w \in A} f(w),$$

where $A = \{w : a < w \le b\}$. We have observed, in this continuous case, that $\Pr(X = x) = 0$, so a summation of such probabilities is no longer appropriate. However, it is easy to find an integral that relates F to f through

$$F(x) = \int_{w \le x} f(w)\, dw.$$

Since $\mathscr{A} = \{x : 0 \le x < 1\}$, this can be written as

$$F(x) = x^2 = \int_0^x f(w)\, dw, \qquad x \in \mathscr{A}.$$

By one form of the fundamental theorem of calculus, we know that the derivative of the right-hand member of this equation is $f(x)$. Thus taking derivatives of each member of the equation, we obtain

$$2x = f(x), \qquad 0 \le x < 1$$

Of course, at $x = 0$, this is only a right-hand derivative. We observe that $f(x) \ge 0$, $x \in \mathscr{A}$, and

$$\int_0^1 2x\, dx = 1.$$

Probabilities can now be computed through

$$\Pr(X \in A) = \int_A f(w)\, dw.$$

For illustration,

$$\Pr\left(\tfrac{1}{4} < X \le \tfrac{1}{2}\right) = \int_{1/4}^{1/2} 2w\, dw = \left[w^2\right]_{1/4}^{1/2}$$

$$= F(\tfrac{1}{2}) - F(\tfrac{1}{4}) = \tfrac{1}{4} - \tfrac{1}{16} = \tfrac{3}{16}.$$

With the background of this example, we give the definition of a random variable of the continuous type.

Let X denote a random variable with a one-dimensional space $\mathscr{A}$, which consists of an interval or a union of intervals. Let a function $f(x)$ be nonnegative such that

$$\int_{\mathscr{A}} f(x)\, dx = 1.$$

Whenever a probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed in terms of such an $f(x)$ by

$$P(A) = \Pr(X \in A) = \int_A f(x)\, dx,$$

then X is said to be a *random variable of the continuous type* and $f(x)$ is called the *probability density function* (p.d.f.) of X.

Example 1. Let the random variable of the continuous type X equal the distance in feet between bad records of a used computer tape. Say that the space of X is $\mathscr{A} = \{x : 0 < x < \infty\}$. Suppose that a reasonable probability model for X is given by the p.d.f.

$$f(x) = \tfrac{1}{40} e^{-x/40}, \qquad x \in \mathscr{A}.$$

Here $f(x) \ge 0$ for $x \in \mathscr{A}$, and

$$\int_0^\infty \tfrac{1}{40} e^{-x/40}\, dx = \left[-e^{-x/40}\right]_0^\infty = 1.$$

If we are interested in the probability that the distance between bad records is greater than 40 feet, then $A = \{x : 40 < x < \infty\}$ and

$$\Pr(X \in A) = \int_{40}^\infty \tfrac{1}{40} e^{-x/40}\, dx = e^{-1}.$$

The p.d.f. and the probability of interest are depicted in Figure 1.5.

If we restrict ourselves to random variables of either the discrete type or the continuous type, we may work exclusively with the p.d.f. $f(x)$. This affords an enormous simplification; but it should be recognized that this simplification is obtained at considerable cost from a mathematical point of view. Not only shall we exclude from

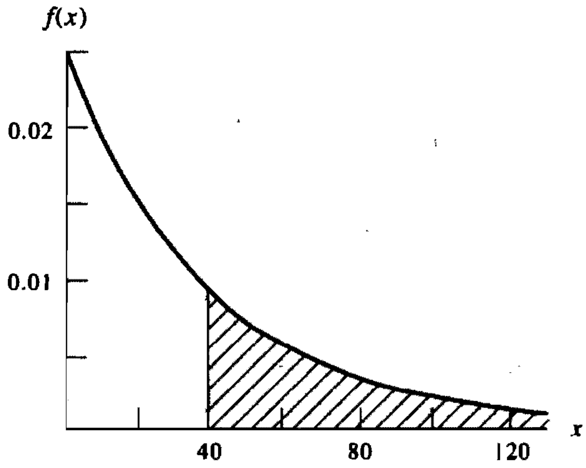

FIGURE 1.5

consideration many random variables that do not have these types of distributions, but we shall also exclude many interesting subsets of the space. In this book, however, we shall in general restrict ourselves to these simple types of random variables.

Remarks. Let X denote the number of spots that show when a die is cast. We can assume that X is a random variable with $\mathscr{A} = \{1, 2, \ldots, 6\}$ and with a p.d.f. $f(x) = \frac{1}{6}$, $x \in \mathscr{A}$. Other assumptions can be made to provide different mathematical models for this experiment. Experimental evidence can be used to help one decide which model is the more realistic. Next, let X denote the point at which a balanced pointer comes to rest. If the circumference is graduated $0 \le x < 1$, a reasonable mathematical model for this experiment is to take X to be a random variable with $\mathscr{A} = \{x : 0 \le x < 1\}$ and with a p.d.f. $f(x) = 1$, $x \in \mathscr{A}$.

Both types of probability density functions can be used as distributional models for many random variables found in real situations. For illustrations consider the following. If X is the number of automobile accidents during a given day, then $f(0), f(1), f(2), \ldots$ represent the probabilities of $0, 1, 2, \ldots$ accidents. On the other hand, if X is length of life of a female born in a certain community, the integral [area under the graph of $f(x)$ that lies above the x-axis and between the vertical lines $x = 40$ and $x = 50$]

$$\int_{40}^{50} f(x)\,dx$$

represents the probability that she dies between 40 and 50 (or the percentage of those females dying between 40 and 50). A particular $f(x)$ will be suggested later for each of these situations, but again experimental evidence must be used to decide whether we have realistic models.

Our notation can be considerably simplified when we restrict ourselves to random variables of the continuous or discrete types. Suppose that the space of a continuous type of random variable X is $\mathscr{A} = \{x : 0 < x < \infty\}$ and that the p.d.f. of X is e^{-x}, $x \in \mathscr{A}$. We shall in no manner alter the distribution of X [that is, alter any $P(A)$, $A \subset \mathscr{A}$] if we extend the definition of the p.d.f. of X by writing

$$\begin{aligned} f(x) &= e^{-x}, \qquad 0 < x < \infty, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

and then refer to $f(x)$ as the p.d.f. of X. We have

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{\infty} e^{-x}\, dx = 1.$$

Thus we may treat the entire axis of reals as though it were the space of X. Accordingly, we now replace

$$\int_{\mathscr{A}} f(x)\, dx \qquad \text{by} \qquad \int_{-\infty}^{\infty} f(x)\, dx.$$

If $f(x)$ is the p.d.f. of a continuous type of random variable X and if A is the set $\{x : a < x < b\}$, then $P(A) = \Pr(X \in A)$ can be written as

$$\Pr(a < X < b) = \int_a^b f(x)\, dx.$$

Moreover, if $A = \{a\}$, then

$$P(A) = \Pr(X \in A) = \Pr(X = a) = \int_a^a f(x)\, dx = 0,$$

since the integral $\int_a^a f(x)\, dx$ is defined in calculus to be zero. That is, if X is a random variable of the continuous type, the probability of every set consisting of a single point is zero. This fact enables us to write, say,

$$\Pr(a < X < b) = \Pr(a \le X \le b).$$

More important, this fact allows us to change the value of the p.d.f. of a continuous type of random variable X at a single point without altering the distribution of X. For instance, the p.d.f.

$$\begin{aligned} f(x) &= e^{-x}, \qquad 0 < x < \infty, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

can be written as

$$f(x) = e^{-x}, \quad 0 \le x < \infty,$$
$$= 0, \quad \text{elsewhere},$$

without changing any $P(A)$. We observe that these two functions differ only at $x = 0$ and $\Pr(X = 0) = 0$. More generally, if two probability density functions of random variables of the continuous type differ only on a set having probability zero, the two corresponding probability set functions are exactly the same. Unlike the continuous type, the p.d.f. of a discrete type of random variable may not be changed at any point, since a change in such a p.d.f. alters the distribution of probability.

Example 2. Let the random variable X of the continuous type have the p.d.f. $f(x) = 2/x^3$, $1 < x < \infty$, zero elsewhere. The distribution function of X is

$$F(x) = \int_{-\infty}^{x} 0\,dw = 0, \quad x < 1,$$
$$= \int_{1}^{x} \frac{2}{w^3}\,dw = 1 - \frac{1}{x^2}, \quad 1 \le x.$$

The graph of this distribution function is depicted in Figure 1.6. Here $F(x)$ is a continuous function for all real numbers x; in particular, $F(x)$ is everywhere continuous from the right. Moreover, the derivative of $F(x)$ with respect to x exists at all points except at $x = 1$. Thus the p.d.f. of X is defined by this derivative except at $x = 1$. Since the set $A = \{1\}$ is a set of probability measure zero [that is, $P(A) = 0$], we are free to define the p.d.f. at $x = 1$ in any manner we please. One way to do this is to write $f(x) = 2/x^3$, $1 < x < \infty$, zero elsewhere.

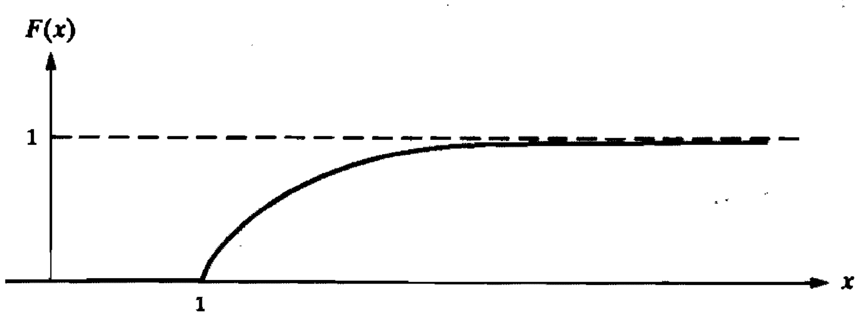

FIGURE 1.6

EXERCISES

1.57. Let a point be selected from the sample space $\mathscr{C} = \{c : 0 < c < 10\}$. Let $C \subset \mathscr{C}$ and let the probability set function be $P(C) = \int_C \frac{1}{10}\,dz$. Define the random variable X to be $X(c) = c^2$. Find the distribution function and the p.d.f. of X.

1.58. Let the probability set function $P(A)$ of the random variable X be $P(A) = \int_A f(x)\,dx$, where $f(x) = 2x/9$, $x \in \mathscr{A} = \{x : 0 < x < 3\}$. Let $A_1 = \{x : 0 < x < 1\}$, $A_2 = \{x : 2 < x < 3\}$. Compute $P(A_1) = \Pr[X \in A_1]$, $P(A_2) = \Pr(X \in A_2)$, and $P(A_1 \cup A_2) = \Pr(X \in A_1 \cup A_2)$.

1.59. Let the space of the random variable X be $\mathscr{A} = \{x : 0 < x < 1\}$. If $A_1 = \{x : 0 < x < \frac{1}{2}\}$ and $A_2 = \{x : \frac{1}{2} \le x < 1\}$, find $P(A_2)$ if $P(A_1) = \frac{1}{4}$.

1.60. Let the space of the random variable X be $\mathscr{A} = \{x : 0 < x < 10\}$ and let $P(A_1) = \frac{3}{8}$, where $A_1 = \{x : 1 < x < 5\}$. Show that $P(A_2) \le \frac{5}{8}$, where $A_2 = \{x : 5 \le x < 10\}$.

1.61. Let the subsets $A_1 = \{x : \frac{1}{4} < x < \frac{1}{2}\}$ and $A_2 = \{x : \frac{1}{2} \le x < 1\}$ of the space $\mathscr{A} = \{x : 0 < x < 1\}$ of the random variable X be such that $P(A_1) = \frac{1}{8}$ and $P(A_2) = \frac{1}{2}$. Find $P(A_1 \cup A_2)$, $P(A_1^*)$, and $P(A_1^* \cap A_2^*)$.

1.62. Given $\int_A [1/\pi(1 + x^2)]\,dx$, where $A \subset \mathscr{A} = \{x : -\infty < x < \infty\}$. Show that the integral could serve as a probability set function of a random variable X whose space is $\mathscr{A}$.

1.63. Let the probability set function of the random variable X be

$$P(A) = \int_A e^{-x}\,dx, \qquad \text{where } \mathscr{A} = \{x : 0 < x < \infty\}.$$

Let $A_k = \{x : 2 - 1/k < x \le 3\}$, $k = 1, 2, 3, \ldots$. Find $\lim_{k \to \infty} A_k$ and $P\left(\lim_{k \to \infty} A_k\right)$.
Find $P(A_k)$ and $\lim_{k \to \infty} P(A_k)$. Note that $\lim_{k \to \infty} P(A_k) = P\left(\lim_{k \to \infty} A_k\right)$.

1.64. For each of the following probability density functions of X, compute $\Pr(|X| < 1)$ and $\Pr(X^2 < 9)$.
(a) $f(x) = x^2/18$, $-3 < x < 3$, zero elsewhere.
(b) $f(x) = (x + 2)/18$, $-2 < x < 4$, zero elsewhere.

1.65. Let $f(x) = 1/x^2$, $1 < x < \infty$, zero elsewhere, be the p.d.f. of X. If $A_1 = \{x : 1 < x < 2\}$ and $A_2 = \{x : 4 < x < 5\}$, find $P(A_1 \cup A_2)$ and $P(A_1 \cap A_2)$.

1.66. A *mode* of a distribution of one random variable X is a value of x that maximizes the p.d.f. $f(x)$. For X of the continuous type, $f(x)$ must be continuous. If there is only one such x, it is called the *mode of the distribution*. Find the mode of each of the following distributions:
(a) $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.

(b) $f(x) = 12x^2(1 - x)$, $0 < x < 1$, zero elsewhere.
(c) $f(x) = (\frac{1}{2})x^2e^{-x}$, $0 < x < \infty$, zero elsewhere.

1.67. A *median* of a distribution of one random variable X of the discrete or continuous type is a value of x such that $\Pr(X < x) \le \frac{1}{2}$ and $\Pr(X \le x) \ge \frac{1}{2}$. If there is only one such x, it is called the *median of the distribution*. Find the median of each of the following distributions:

(a) $f(x) = \dfrac{4!}{x!\,(4-x)!}\left(\dfrac{1}{4}\right)^x\left(\dfrac{3}{4}\right)^{4-x}$, $x = 0, 1, 2, 3, 4$, zero elsewhere.

(b) $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = \dfrac{1}{\pi(1 + x^2)}$, $-\infty < x < \infty$.

Hint: In parts (b) and (c), $\Pr(X < x) = \Pr(X \le x)$ and thus that common value must equal $\frac{1}{2}$ if x is to be the median of the distribution.

1.68. Let $0 < p < 1$. A $(100p)$th *percentile* (*quantile* of order p) of the distribution of a random variable X is a value ξ_p such that $\Pr(X < \xi_p) \le p$ and $\Pr(X \le \xi_p) \ge p$. Find the twentieth percentile of the distribution that has p.d.f. $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

Hint: With a continuous-type random variable X, $\Pr(X < \xi_p) = \Pr(X \le \xi_p)$ and hence that common value must equal p.

1.69. Find the distribution function $F(x)$ associated with each of the following probability density functions. Sketch the graphs of $f(x)$ and $F(x)$.

(a) $f(x) = 3(1 - x)^2$, $0 < x < 1$, zero elsewhere.
(b) $f(x) = 1/x^2$, $1 < x < \infty$, zero elsewhere.
(c) $f(x) = \frac{1}{3}$, $0 < x < 1$ or $2 < x < 4$, zero elsewhere.

Also find the median and 25th percentile of each of these distributions.

1.70. Consider the distribution function $F(x) = 1 - e^{-x} - xe^{-x}$, $0 \le x < \infty$, zero elsewhere. Find the p.d.f., the mode, and the median (by numerical methods) of this distribution.

1.7 Properties of the Distribution Function

In Section 1.5 we defined the distribution function of a random variable X as $F(x) = \Pr(X \le x)$. This concept was used in Section 1.6 to find the probability distribution of a random variable of the continuous type. So, in terms of the p.d.f. $f(x)$, we know that

$$F(x) = \sum_{w \le x} f(w),$$

for the discrete type of random variable, and

$$F(x) = \int_{-\infty}^{x} f(w)\, dw,$$

for the continuous type of random variable. We speak of a distribution function $F(x)$ as being of the continuous or discrete type, depending on whether the random variable is of the continuous or discrete type.

Remark. If X is a random variable of the continuous type, the p.d.f. $f(x)$ has at most a finite number of discontinuities in every finite interval. This means (1) that the distribution function $F(x)$ is everywhere continuous and (2) that the derivative of $F(x)$ with respect to x exists and is equal to $f(x)$ at each point of continuity of $f(x)$. That is, $F'(x) = f(x)$ at each point of continuity of $f(x)$. If the random variable X is of the discrete type, most surely the p.d.f. $f(x)$ is *not* the derivative of $F(x)$ with respect to x (that is, with respect to Lebesgue measure); but $f(x)$ *is* the (Radon–Nikodym) derivative of $F(x)$ with respect to a counting measure. A derivative is often called a *density*. Accordingly, we call these derivatives *probability density functions*.

There are several properties of a distribution function $F(x)$ that can be listed as a consequence of the properties of the probability set function. Some of these are the following. In listing these properties, we shall not restrict X to be a random variable of the discrete or continuous type. We shall use the symbols $F(\infty)$ and $F(-\infty)$ to mean $\lim_{x\to\infty} F(x)$ and $\lim_{x\to-\infty} F(x)$, respectively. In like manner, the symbols $\{x : x < \infty\}$ and $\{x : x < -\infty\}$ represent, respectively, the limits of the sets $\{x : x \le b\}$ and $\{x : x \le -b\}$ as $b \to \infty$.

1. $0 \le F(x) \le 1$ because $0 \le \Pr(X \le x) \le 1$.
2. $F(x)$ is a nondecreasing function of x. For, if $x' < x''$, then

$$\{x : x \le x''\} = \{x : x \le x'\} \cup \{x : x' < x \le x''\}$$

and

$$\Pr(X \le x'') = \Pr(X \le x') + \Pr(x' < X \le x'').$$

That is,

$$F(x'') - F(x') = \Pr(x' < X \le x'') \ge 0.$$

3. $F(\infty) = 1$ and $F(-\infty) = 0$ because the set $\{x : x \le \infty\}$ is the entire one-dimensional space and the set $\{x : x \le -\infty\}$ is the null set.

From the proof of property 2, it is observed that, if $a < b$, then

$$\Pr(a < X \le b) = F(b) - F(a).$$

Suppose that we want to use $F(x)$ to compute the probability $\Pr(X = b)$. To do this, consider, with $h > 0$,

$$\lim_{h \to 0} \Pr(b - h < X \le b) = \lim_{h \to 0} [F(b) - F(b - h)].$$

Intuitively, it seems that $\lim_{h \to 0} \Pr(b - h < X \le b)$ should exist and be equal to $\Pr(X = b)$ because, as h tends to zero, the limit of the set $\{x : b - h < x \le b\}$ is the set that contains the single point $x = b$. The fact that this limit is $\Pr(X = b)$ is a theorem that we accept without proof. Accordingly, we have

$$\Pr(X = b) = F(b) - F(b-),$$

where $F(b-)$ is the left-hand limit of $F(x)$ at $x = b$. That is, the probability that $X = b$ is the height of the step that $F(x)$ has at $x = b$. Hence, if the distribution function $F(x)$ is continuous at $x = b$, then $\Pr(X = b) = 0$.

There is a fourth property of $F(x)$ that is now listed.

4. $F(x)$ is continuous from the right, that is, right-continuous.

To prove this property, consider, with $h > 0$,

$$\lim_{h \to 0} \Pr(a < X \le a + h) = \lim_{h \to 0} [F(a + h) - F(a)].$$

We accept without proof a theorem which states, with $h > 0$, that

$$\lim_{h \to 0} \Pr(a < X \le a + h) = P(\varnothing) = 0.$$

Here also, the theorem is intuitively appealing because, as h tends to zero, the limit of the set $\{x : a < x \le a + h\}$ is the null set. Accordingly, we write

$$0 = F(a+) - F(a),$$

where $F(a+)$ is the right-hand limit of $F(x)$ at $x = a$. Hence $F(x)$ is continuous from the right at every point $x = a$.

Remark. In the arguments concerning several of these properties, we appeal to the reader's intuition. However, most of these properties can be proved in formal ways using the definition of $\lim_{k \to \infty} A_k$, given in Exercises 1.7

and 1.8, and the fact that the probability set function P is countably additive; that is, P enjoys (b) of Definition 7.

The preceding discussion may be summarized in the following manner: A distribution function $F(x)$ is a nondecreasing function of x, which is everywhere continuous from the right and has $F(-\infty) = 0$, $F(\infty) = 1$. The probability $\Pr(a < X \le b)$ is equal to the difference $F(b) - F(a)$. If x is a discontinuity point of $F(x)$, then the probability $\Pr(X = x)$ is equal to the jump which the distribution function has at the point x. If x is a continuity point of $F(x)$, then $\Pr(X = x) = 0$.

Remark. The definition of the distribution function makes it clear that the probability set function P determines the distribution function F. It is true, although not so obvious, that a probability set function P can be found from a distribution function F. That is, P and F give the same information about the distribution of probability, and which function is used is a matter of convenience.

Often, probability models can be constructed that make reasonable assumptions about the probability set function and thus the distribution function. For a simple illustration, consider an experiment in which one chooses at random a point from the closed interval $[a, b]$, $a < b$, that is on the real line. Thus the sample space $\mathscr{C}$ is $[a, b]$. Let the random variable X be the identity function defined on $\mathscr{C}$. Thus the space $\mathscr{A}$ of X is $\mathscr{A} = \mathscr{C}$. Suppose that it is reasonable to *assume*, from the nature of the experiment, that if an interval A is a subset of $\mathscr{A}$, the probability of the event A is proportional to the length of A. Hence, if A is the interval $[a, x]$, $x \le b$, then

$$P(A) = \Pr(X \in A) = \Pr(a \le X \le x) = c(x - a),$$

where c is the constant of proportionality.

In the expression above, if we take $x = b$, we have

$$1 = \Pr(a \le X \le b) = c(b - a),$$

so $c = 1/(b - a)$. Thus we will have an appropriate probability model if we take the distribution function of X, $F(x) = \Pr(X \le x)$, to be

$$\begin{aligned} F(x) &= 0, \qquad x < a, \\ &= \frac{x - a}{b - a}, \qquad a \le x \le b, \\ &= 1, \qquad b < x. \end{aligned}$$

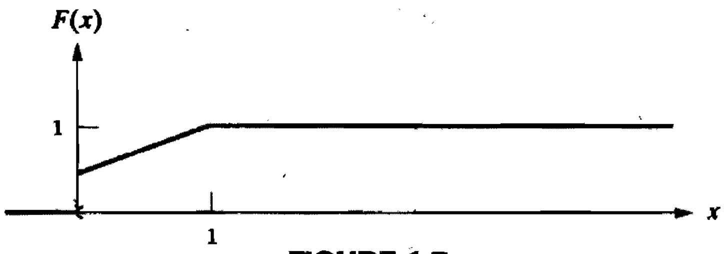

FIGURE 1.7

Accordingly, the p.d.f. of X, $f(x) = F'(x)$, may be written

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b,$$
$$= 0 \qquad \text{elsewhere.}$$

The derivative of $F(x)$ does not exist at $x = a$ nor at $x = b$; but the set $\{x : x = a, b\}$ is a set of probability measure zero, and we elect to define $f(x)$ to be equal to $1/(b-a)$ at those two points, just as a matter of convenience. We observe that this p.d.f. is a constant on $\mathscr{A}$. If the p.d.f. of one or more variables of the continuous type or of the discrete type is a constant on the space $\mathscr{A}$, we say that the probability is distributed *uniformly* over $\mathscr{A}$. Thus, in the example above, we say that X has a *uniform distribution* over the interval $[a, b]$.

We now give an illustrative example of a distribution that is neither of the discrete nor continuous type.

Example 1. Let a distribution function be given by

$$F(x) = 0, \qquad x < 0,$$
$$= \frac{x+1}{2}, \qquad 0 \le x < 1,$$
$$= 1, \qquad 1 \le x.$$

Then, for instance,

$$\Pr(-3 < X \le \tfrac{1}{2}) = F(\tfrac{1}{2}) - F(-3) = \tfrac{3}{4} - 0 = \tfrac{3}{4}$$

and

$$\Pr(X = 0) = F(0) - F(0-) = \tfrac{1}{2} - 0 = \tfrac{1}{2}.$$

The graph of $F(x)$ is shown in Figure 1.7. We see that $F(x)$ is not always continuous, nor is it a step function. Accordingly, the corresponding distribution is neither of the continuous type nor of the discrete type. It may be described as a mixture of those types.

Distributions that are mixtures of the continuous and discrete types do, in fact, occur frequently in practice. For illustration, in life testing, suppose we know that the length of life, say X, exceeds the number b, but the exact value is unknown. This is called *censoring*. For instance, this can happen when a subject in a cancer study simply disappears; the investigator knows that the subject has lived a certain number of months, but the exact length of life is unknown. Or it might happen when an investigator does not have enough time in an investigation to observe the moments of deaths of all the animals, say rats, in some study. Censoring can also occur in the insurance industry; in particular, consider a loss with a limited-pay policy in which the top amount is exceeded but it is not known by how much.

Example 2. Reinsurance companies are concerned with large losses because they might agree, for illustration, to cover losses due to wind damages that are between \$2,000,000 and \$10,000,000. Say that X equals the size of a wind loss in millions of dollars, and suppose that it has the distribution function

$$\begin{aligned} F(x) &= 0, \qquad -\infty < x < 0, \\ &= 1 - \left(\frac{10}{10+x}\right)^3, \qquad 0 \le x < \infty. \end{aligned}$$

If losses beyond \$10,000,000 are reported only as 10, then the distribution function of this censored distribution is

$$\begin{aligned} F(x) &= 0, \qquad -\infty < x < 0, \\ &= 1 - \left(\frac{10}{10+x}\right)^3, \qquad 0 \le x < 10, \\ &= 1, \qquad 10 \le x < \infty, \end{aligned}$$

which has a jump of $[10/(10+10)]^3 = \frac{1}{8}$ at $x = 10$.

We shall now point out an important fact about a function of a random variable. Let X denote a random variable with space $\mathscr{A}$. Consider the function $Y = u(X)$ of the random variable X. Since X is a function defined on a sample space $\mathscr{C}$, then $Y = u(X)$ is a composite function defined on $\mathscr{C}$. That is, $Y = u(X)$ is itself a random variable which has its own space $\mathscr{B} = \{y : y = u(x), x \in \mathscr{A}\}$ and its own probability set function. If $y \in \mathscr{B}$, the event $Y = u(X) \le y$ occurs when, and only when, the event $X \in A \subset \mathscr{A}$ occurs, where $A = \{x : u(x) \le y\}$. That is, the distribution function of Y is

$$G(y) = \Pr(Y \le y) = \Pr[u(X) \le y] = P(A).$$

The following example illustrates a method of finding the distribution function and the p.d.f. of a function of a random variable. This method is called the *distribution-function technique.*

Example 3. Let $f(x) = \frac{1}{2}$, $-1 < x < 1$, zero elsewhere, be the p.d.f. of the random variable X. Define the random variable Y by $Y = X^2$. We wish to find the p.d.f. of Y. If $y \geq 0$, the probability $\Pr(Y \leq y)$ is equivalent to

$$\Pr(X^2 \leq y) = \Pr(-\sqrt{y} \leq X \leq \sqrt{y}).$$

Accordingly, the distribution function of Y, $G(y) = \Pr(Y \leq y)$, is given by

$$\begin{aligned} G(y) &= 0, \quad y < 0, \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \tfrac{1}{2}\,dx = \sqrt{y}, \quad 0 \leq y < 1, \\ &= 1, \quad 1 \leq y. \end{aligned}$$

Since Y is a random variable of the continuous type, the p.d.f. of Y is $g(y) = G'(y)$ at all points of continuity of $g(y)$. Thus we may write

$$\begin{aligned} g(y) &= \frac{1}{2\sqrt{y}}, \quad 0 < y < 1, \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

Remarks. Many authors use f_X and f_Y to denote the respective probability density functions of the random variables X and Y. Here we use f and g because we can avoid the use of subscripts. However, at other times, we will use subscripts as in f_X and f_Y or even f_1 and f_2, depending upon the circumstances. In a given example, we do not use the same symbol, without subscripts, to represent different functions. That is, in Example 2, we do not use $f(x)$ and $f(y)$ to represent different probability density functions.

In addition, while we ordinarily use the letter x in the description of the p.d.f. of X, this is not necessary at all because it is unimportant which letter we use in describing a function. For illustration, in Example 3, we could say that the random variable Y has the p.d.f. $g(w) = 1/2\sqrt{w}, 0 < w < 1$, zero elsewhere, and it would have exactly the same meaning as Y has the p.d.f. $g(y) = 1/2\sqrt{y}$, $0 < y < 1$, zero elsewhere.

These remarks apply to other functions too, such as distribution functions. In Example 3, we could have written the distribution function of Y, where $0 \leq w < 1$, as

$$F_Y(w) = \Pr(Y \leq w) = \sqrt{w}.$$

EXERCISES

1.71. Given the distribution function

$$F(x)=0, \quad x<-1,$$
$$=\frac{x+2}{4}, \quad -1\le x<1,$$
$$=1, \quad 1\le x.$$

Sketch the graph of $F(x)$ and then compute: (a) $\Pr(-\frac{1}{2}<X\le\frac{1}{2})$; (b) $\Pr(X=0)$; (c) $\Pr(X=1)$; (d) $\Pr(2<X\le 3)$.

1.72. Let $f(x)=1$, $0<x<1$, zero elsewhere, be the p.d.f. of X. Find the distribution function and the p.d.f. of $Y=\sqrt{X}$.
Hint: $\Pr(Y\le y)=\Pr(\sqrt{X}\le y)=\Pr(X\le y^2)$, $0<y<1$.

1.73. Let $f(x)=x/6$, $x=1, 2, 3$, zero elsewhere, be the p.d.f. of X. Find the distribution function and the p.d.f. of $Y=X^2$.
Hint: Note that X is a random variable of the discrete type.

1.74. Let $f(x)=(4-x)/16$, $-2<x<2$, zero elsewhere, be the p.d.f. of X.
(a) Sketch the distribution function and the p.d.f. of X on the same set of axes.
(b) If $Y=|X|$, compute $\Pr(Y\le 1)$.
(c) If $Z=X^2$, compute $\Pr(Z\le\frac{1}{4})$.

1.75. Let X have the p.d.f. $f(x)=2x$, $0<x<1$, zero elsewhere. Find the distribution function and p.d.f. of $Y=X^2$.

1.76. Let X have the p.d.f. $f(x)=4x^3$, $0<x<1$, zero elsewhere. Find the distribution function and p.d.f. of $Y=-2\ln X^4$.

1.77. Explain why, with $h>0$, the two limits, $\lim_{h\to 0}\Pr(b-h<X\le b)$ and $\lim_{h\to 0}F(b-h)$, exist.
Hint: Note that $\Pr(b-h<X\le b)$ is bounded below by zero and $F(b-h)$ is bounded above by both $F(b)$ and 1.

1.78. Let $F(x)$ be the distribution function of the random variable X. If m is a number such that $F(m)=\frac{1}{2}$, show that m is a median of the distribution.

1.79. Let $f(x)=\frac{1}{3}$, $-1<x<2$, zero elsewhere, be the p.d.f. of X. Find the distribution function and the p.d.f. of $Y=X^2$.
Hint: Consider $\Pr(X^2\le y)$ for two cases: $0\le y<1$ and $1\le y<4$.

1.8 Expectation of a Random Variable

Let X be a random variable having a p.d.f. $f(x)$ such that we have certain absolute convergence; namely, in the discrete case,

$$\sum_x |x| f(x) \qquad \text{converges to a finite limit,}$$

or, in the continuous case,

$$\int_{-\infty}^{\infty} |x| f(x)\, dx \qquad \text{converges to a finite limit.}$$

The *expectation of a random variable* is

$$E(X) = \sum_x x f(x), \qquad \text{in the discrete case,}$$

or

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx, \qquad \text{in the continuous case.}$$

Sometimes the expectation $E(X)$ is called the *mathematical expectation of X* or the *expected value of X*.

Remark. The terminology of expectation or expected value has its origin in games of chance. This can be illustrated as follows: Four small similar chips, numbered 1, 1, 1, and 2, respectively, are placed in a bowl and are mixed. A player is blindfolded and is to draw a chip from the bowl. If she draws one of the three chips numbered 1, she will receive one dollar. If she draws the chip numbered 2, she will receive two dollars. It seems reasonable to assume that the player has a "$\frac{3}{4}$ claim" on the \$1 and a "$\frac{1}{4}$ claim" on the \$2. Her "total claim" is $(1)(\frac{3}{4}) + 2(\frac{1}{4}) = \frac{5}{4}$, that is, \$1.25. Thus the expectation of X is precisely the player's claim in this game.

Example 1. Let the random variable X of the discrete type have the p.d.f. given by the table

x	1	2	3	4
$f(x)$	$\frac{4}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{2}{10}$

Here $f(x) = 0$ if x is not equal to one of the first four positive integers. This illustrates the fact that there is no need to have a formula to describe a p.d.f. We have

$$E(X) = (1)(\tfrac{4}{10}) + 2(\tfrac{1}{10}) + 3(\tfrac{3}{10}) + 4(\tfrac{2}{10}) = \tfrac{23}{10} = 2.3.$$

Example 2. Let X have the p.d.f.

$$f(x) = 4x^3, \qquad 0 < x < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

Then

$$E(X) = \int_0^1 x(4x^3)\,dx = \int_0^1 4x^4\,dx = \left[\frac{4x^5}{5}\right]_0^1 = \frac{4}{5}.$$

Let us consider a function of a random variable X with space $\mathscr{A}$. Call this function $Y = u(X)$. For convenience, let X be of the continuous type and $y = u(x)$ be a continuous increasing function of X with an inverse function $x = w(y)$, which, of course, is also increasing. So Y is a random variable and its distribution function is

$$G(y) = \Pr(Y \le y) = \Pr[u(X) \le y] = \Pr[X \le w(y)]$$
$$= \int_{-\infty}^{w(y)} f(x)\,dx,$$

where $f(x)$ is the p.d.f. of X. By one form of the fundamental theorem of calculus,

$$g(y) = G'(y) = f[w(y)]w'(y), \qquad y \in \mathscr{B},$$
$$= 0 \qquad \text{elsewhere,}$$

where

$$\mathscr{B} = \{y : y = u(x), \quad x \in \mathscr{A}\}.$$

By definition, given absolute convergence, the expected value of Y is

$$E(Y) = \int_{-\infty}^{\infty} y g(y)\,dy.$$

Since $y = u(x)$, we might ask how $E(Y)$ compares to the integral

$$I = \int_{-\infty}^{\infty} u(x)f(x)\,dx.$$

To answer this, change the variable of integration through $y = u(x)$ or, equivalently, $x = w(y)$. Since

$$\frac{dx}{dy} = w'(y) > 0,$$

we have

$$I = \int_{-\infty}^{\infty} yf[w(y)]w'(y)\,dy = \int_{-\infty}^{\infty} yg(y)\,dy.$$

That is, in this special case,

$$E(Y) = \int_{-\infty}^{\infty} yg(y)\,dy = \int_{-\infty}^{\infty} u(x)f(x)\,dx.$$

However, this is true more generally and it also makes no difference whether X is of the discrete or continuous type and $Y = u(X)$ need not be an increasing function of X (Exercise 1.87 illustrates this).

So if $Y = u(X)$ has an expectation, we can find it from

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\,dx, \tag{1}$$

in the continuous case, and

$$E[u(X)] = \sum_x u(x)f(x), \tag{2}$$

in the discrete case. Accordingly, we say that $E[u(X)]$ is the expectation (mathematical expectation or expected value) of $u(X)$.

Remark. If the mathematical expectation of Y exists, recall that the integral (or sum)

$$\int_{-\infty}^{\infty} |y|g(y)\,dy \qquad \left[\text{or} \quad \sum_y |y|g(y)\right]$$

exists. Hence the existence of $E[u(X)]$ implies that the corresponding integral (or sum) converges absolutely.

Next, we shall point out some fairly obvious but useful facts about expectations when they exist.

1. If k is a constant, then $E(k) = k$. This follows from expression (1) [or (2)] upon setting $u = k$ and recalling that an integral (or sum) of a constant times a function is the constant times the integral (or sum) of the function. Of course, the integral (or sum) of the function f is 1.
2. If k is a constant and v is a function, then $E(kv) = kE(v)$. This follows from expression (1) [or (2)] upon setting $u = kv$ and rewriting expression (1) [or (2)] as k times the integral (or sum) of the product vf.
3. If k_1 and k_2 are constants and v_1 and v_2 are functions, then $E(k_1v_1 + k_2v_2) = k_1E(v_1) + k_2E(v_2)$. This, too, follows from ex-

pression (1) [or (2)] upon setting $u = k_1v_1 + k_2v_2$ because the integral (or sum) of $(k_1v_1 + k_2v_2)f$ is equal to the integral (or sum) of k_1v_1f plus the integral (or sum) of k_2v_2f. Repeated application of this property shows that if $k_1, k_2, \ldots, k_m$ are constants and $v_1, v_2, \ldots, v_m$ are functions, then

$$E(k_1v_1 + k_2v_2 + \cdots + k_mv_m) = k_1E(v_1) + k_2E(v_2) + \cdots + k_mE(v_m).$$

This property of expectation leads us to characterize the symbol E as a linear operator.

Example 3. Let X have the p.d.f.

$$f(x) = 2(1 - x), \quad 0 < x < 1,$$
$$= 0 \quad \text{elsewhere.}$$

Then

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^1 (x)2(1 - x)\,dx = \tfrac{1}{3},$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2f(x)\,dx = \int_0^1 (x^2)2(1 - x)\,dx = \tfrac{1}{6},$$

and, of course,

$$E(6X + 3X^2) = 6(\tfrac{1}{3}) + 3(\tfrac{1}{6}) = \tfrac{5}{2}.$$

Example 4. Let X have the p.d.f.

$$f(x) = \frac{x}{6}, \quad x = 1, 2, 3,$$
$$= 0 \quad \text{elsewhere.}$$

Then

$$E(X^3) = \sum_x x^3f(x) = \sum_{x=1}^{3} x^3\frac{x}{6}$$

$$= \tfrac{1}{6} + \tfrac{16}{6} + \tfrac{81}{6} = \tfrac{98}{6}.$$

Example 5. Let us divide, at random, a horizontal line segment of length 5 into two parts. If X is the length of the left-hand part, it is reasonable to assume that X has the p.d.f.

$$f(x) = \tfrac{1}{5}, \quad 0 < x < 5,$$
$$= 0 \quad \text{elsewhere.}$$

The expected value of the length X is $E(X) = \frac{5}{2}$ and the expected value of the length $5 - X$ is $E(5 - X) = \frac{5}{2}$. But the expected value of the product of the two lengths is equal to

$$E[X(5 - X)] = \int_0^5 x(5 - x)(\tfrac{1}{5})\,dx = \tfrac{25}{6} \neq (\tfrac{5}{2})^2.$$

That is, in general, the expected value of a product is not equal to the product of the expected values.

Example 6. A bowl contains five chips, which cannot be distinguished by a sense of touch alone. Three of the chips are marked \$1 each and the remaining two are marked \$4 each. A player is blindfolded and draws, at random and without replacement, two chips from the bowl. The player is paid an amount equal to the sum of the values of the two chips that he draws and the game is over. If it costs \$4.75 to play this game, would we care to participate for any protracted period of time? Because we are unable to distinguish the chips by sense of touch, we assume that each of the 10 pairs that can be drawn has the same probability of being drawn. Let the random variable X be the number of chips, of the two to be chosen, that are marked \$1. Then, under our assumption, X has the hypergeometric p.d.f.

$$f(x) = \frac{\binom{3}{x}\binom{2}{2-x}}{\binom{5}{2}}, \qquad x = 0, 1, 2,$$

$$= 0 \qquad \text{elsewhere.}$$

If $X = x$, the player receives $u(x) = x + 4(2 - x) = 8 - 3x$ dollars. Hence his mathematical expectation is equal to

$$E[8 - 3X] = \sum_{x=0}^{2} (8 - 3x)f(x) = \tfrac{44}{10},$$

or \$4.40.

EXERCISES

1.80. Let X have the p.d.f. $f(x) = (x + 2)/18$, $-2 < x < 4$, zero elsewhere. Find $E(X)$, $E[(X + 2)^3]$, and $E[6X - 2(X + 2)^3]$.

1.81. Suppose that $f(x) = \frac{1}{5}$, $x = 1, 2, 3, 4, 5$, zero elsewhere, is the p.d.f. of the discrete type of random variable X. Compute $E(X)$ and $E(X^2)$. Use these two results to find $E[(X + 2)^2]$ by writing $(X + 2)^2 = X^2 + 4X + 4$.

1.82. Let X be a number selected at random from a set of numbers $\{51, 52, 53, \ldots, 100\}$. Approximate $E(1/X)$.

Hint: Find reasonable upper and lower bounds by finding integrals bounding $E(1/X)$.

1.83. Let the p.d.f. $f(x)$ be positive at $x = -1, 0, 1$ and zero elsewhere.
(a) If $f(0) = \frac{1}{4}$, find $E(X^2)$.
(b) If $f(0) = \frac{1}{4}$ and if $E(X) = \frac{1}{4}$, determine $f(-1)$ and $f(1)$.

1.84. Let X have the p.d.f. $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere. Consider a random rectangle whose sides are X and $(1 - X)$. Determine the expected value of the area of the rectangle.

1.85. A bowl contains 10 chips, of which 8 are marked \$2 each and 2 are marked \$5 each. Let a person choose, at random and without replacement, 3 chips from this bowl. If the person is to receive the sum of the resulting amounts, find his expectation.

1.86. Let X be a random variable of the continuous type that has p.d.f. $f(x)$. If m is the unique median of the distribution of X and b is a real constant, show that

$$E(|X - b|) = E(|X - m|) + 2 \int_m^b (b - x)f(x)\,dx,$$

provided that the expectations exist. For what value of b is $E(|X - b|)$ a minimum?

1.87. Let $f(x) = 2x$, $0 < x < 1$, zero elsewhere, be the p.d.f. of X.
(a) Compute $E(1/X)$.
(b) Find the distribution function and the p.d.f. of $Y = 1/X$.
(c) Compute $E(Y)$ and compare this result with the answer obtained in part (a).
Hint: Here $\mathscr{A} = \{x : 0 < x < 1\}$, find $\mathscr{B}$.

1.88. Two distinct integers are chosen at random and without replacement from the first six positive integers. Compute the expected value of the absolute value of the difference of these two numbers.

1.9 Some Special Expectations

Certain expectations, if they exist, have special names and symbols to represent them. First, let X be a random variable of the discrete type having a p.d.f. $f(x)$. Then

$$E(X) = \sum_x x f(x).$$

If the discrete points of the space of positive probability density are $a_1, a_2, a_3, \ldots$, then

$$E(X) = a_1 f(a_1) + a_2 f(a_2) + a_3 f(a_3) + \cdots.$$

This sum of products is seen to be a "weighted average" of the values $a_1, a_2, a_3, \ldots$, the "weight" associated with each a_i being $f(a_i)$. This suggests that we call $E(X)$ the arithmetic mean of the values of X, or, more simply, the *mean value* of X (or the mean value of the distribution).

The mean value μ of a random variable X is defined, when it exists, to be $\mu = E(X)$, where X is a random variable of the discrete or of the continuous type.

Another special expectation is obtained by taking $u(X) = (X - \mu)^2$. If, initially, X is a random variable of the discrete type having a p.d.f. $f(x)$, then

$$\begin{aligned} E[(X-\mu)^2] &= \sum_x (x-\mu)^2 f(x) \\ &= (a_1 - \mu)^2 f(a_1) + (a_2 - \mu)^2 f(a_2) + \cdots, \end{aligned}$$

if $a_1, a_2, \ldots$ are the discrete points of the space of positive probability density. This sum of products may be interpreted as a "weighted average" of the squares of the deviations of the numbers $a_1, a_2, \ldots$ from the mean value μ of those numbers where the "weight" associated with each $(a_i - \mu)^2$ is $f(a_i)$. This mean value of the square of the deviation of X from its mean value μ is called the *variance* of X (or the variance of the distribution).

The variance of X will be denoted by σ^2, and we define σ^2, if it exists, by $\sigma^2 = E[(X - \mu)^2]$, whether X is a discrete or a continuous type of random variable. Sometimes the variance of X is written var (X).

It is worthwhile to observe that var (X) equals

$$\sigma^2 = E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2);$$

and since E is a linear operator,

$$\begin{aligned} \sigma^2 &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

This frequency affords an easier way of computing the variance of X.

It is customary to call σ (the positive square root of the variance) the *standard deviation* of X (or the standard deviation of the distribution). The number σ is sometimes interpreted as a measure of the dispersion of the points of the space relative to the mean value μ. We note that if the space contains only one point x for which $f(x) > 0$, then $\sigma = 0$.

Remark. Let the random variable X of the continuous type have the p.d.f. $f(x) = 1/2a$, $-a < x < a$, zero elsewhere, so that $\sigma = a/\sqrt{3}$ is the standard deviation of the distribution of X. Next, let the random variable Y of the continuous type have the p.d.f. $g(y) = 1/4a$, $-2a < y < 2a$, zero elsewhere, so that $\sigma = 2a/\sqrt{3}$ is the standard deviation of the distribution of Y. Here the standard deviation of Y is greater than that of X; this reflects the fact that the probability for Y is more widely distributed (relative to the mean zero) than is the probability for X.

We next define a third special mathematical expectation, called the *moment-generating function* (abbreviated m.g.f.) of a random variable X. Suppose that there is a positive number h such that for $-h < t < h$ the mathematical expectation $E(e^{tX})$ exists. Thus

$$E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}f(x)\,dx,$$

if X is a continuous type of random variable, or

$$E(e^{tX}) = \sum_{x} e^{tx}f(x),$$

if X is a discrete type of random variable. This expectation is called the moment-generating function (m.g.f.) of X (or of the distribution) and is denoted by $M(t)$. That is,

$$M(t) = E(e^{tX}).$$

It is evident that if we set $t = 0$, we have $M(0) = 1$. As will be seen by example, not every distribution has an m.g.f., but it is difficult to overemphasize the importance of an m.g.f., when it does exist. This importance stems from the fact that the m.g.f. is unique and completely determines the distribution of the random variable; thus, if two random variables have the same m.g.f., they have the same distribution. This property of an m.g.f. will be very useful in subsequent chapters. Proof of the uniqueness of the m.g.f. is based on the theory of transforms in analysis, and therefore we merely assert this uniqueness.

Although the fact that an m.g.f. (when it exists) completely determines the distribution of one random variable will not be proved, it does seem desirable to try to make the assertion plausible. This can be done if the random variable is of the discrete type. For example, let it be given that

$$M(t) = \tfrac{1}{10}e^t + \tfrac{2}{10}e^{2t} + \tfrac{3}{10}e^{3t} + \tfrac{4}{10}e^{4t}$$

is, for all real values of t, the m.g.f. of a random variable X of the discrete type. If we let $f(x)$ be the p.d.f. of X and let $a, b, c, d, \ldots$ be the discrete points in the space of X at which $f(x) > 0$, then

$$M(t) = \sum_x e^{tx} f(x),$$

or

$$\tfrac{1}{10}e^t + \tfrac{2}{10}e^{2t} + \tfrac{3}{10}e^{3t} + \tfrac{4}{10}e^{4t} = f(a)e^{at} + f(b)e^{bt} + \cdots.$$

Because this is an identity for all real values of t, it seems that the right-hand member should consist of but four terms and that each of the four should equal, respectively, one of those in the left-hand member; hence we may take $a = 1, f(a) = \tfrac{1}{10}$; $b = 2, f(b) = \tfrac{2}{10}$; $c = 3$, $f(c) = \tfrac{3}{10}$; $d = 4, f(d) = \tfrac{4}{10}$. Or, more simply, the p.d.f. of X is

$$f(x) = \frac{x}{10}, \qquad x = 1, 2, 3, 4,$$
$$= 0 \qquad \text{elsewhere.}$$

On the other hand, let X be a random variable of the continuous type and let it be given that

$$M(t) = \frac{1}{1-t}, \qquad t < 1,$$

is the m.g.f. of X. That is, we are given

$$\frac{1}{1-t} = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx, \qquad t < 1.$$

It is not at all obvious how $f(x)$ is found. However, it is easy to see that a distribution with p.d.f.

$$f(x) = e^{-x}, \qquad 0 < x < \infty,$$
$$= 0 \qquad \text{elsewhere}$$

has the m.g.f. $M(t) = (1 - t)^{-1}$, $t < 1$. Thus the random variable X

has a distribution with this p.d.f. in accordance with the assertion of the uniqueness of the m.g.f.

Since a distribution that has an m.g.f. $M(t)$ is completely determined by $M(t)$, it would not be surprising if we could obtain some properties of the distribution directly from $M(t)$. For example, the existence of $M(t)$ for $-h < t < h$ implies that derivatives of all order exist at $t = 0$. Thus, using a theorem in analysis that allows us to change the order of differentiation and integration, we have

$$\frac{dM(t)}{dt} = M'(t) = \int_{-\infty}^{\infty} xe^{tx}f(x)\,dx,$$

if X is of the continuous type, or

$$\frac{dM(t)}{dt} = M'(t) = \sum_x xe^{tx}f(x),$$

if X is of the discrete type. Upon setting $t = 0$, we have in either case

$$M'(0) = E(X) = \mu.$$

The second derivative of $M(t)$ is

$$M''(t) = \int_{-\infty}^{\infty} x^2e^{tx}f(x)\,dx \qquad \text{or} \qquad \sum_x x^2e^{tx}f(x),$$

so that $M''(0) = E(X^2)$. Accordingly, the var (X) equals

$$\sigma^2 = E(X^2) - \mu^2 = M''(0) - [M'(0)]^2.$$

For example, if $M(t) = (1 - t)^{-1}$, $t < 1$, as in the illustration above, then

$$M'(t) = (1 - t)^{-2} \qquad \text{and} \qquad M''(t) = 2(1 - t)^{-3}.$$

Hence

$$\mu = M'(0) = 1$$

and

$$\sigma^2 = M''(0) - \mu^2 = 2 - 1 = 1.$$

Of course, we could have computed μ and σ^2 from the p.d.f. by

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} x^2f(x)\,dx - \mu^2,$$

respectively. Sometimes one way is easier than the other.

In general, if m is a positive integer and if $M^{(m)}(t)$ means the mth derivative of $M(t)$, we have, by repeated differentiation with respect to t,

$$M^{(m)}(0) = E(X^m).$$

Now

$$E(X^m) = \int_{-\infty}^{\infty} x^m f(x)\,dx \qquad \text{or} \qquad \sum_x x^m f(x),$$

and integrals (or sums) of this sort are, in mechanics, called *moments*. Since $M(t)$ generates the values of $E(X^m)$, $m = 1, 2, 3, \ldots$, it is called the moment-generating function (m.g.f.). In fact, we shall sometimes call $E(X^m)$ the mth moment of the distribution, or the mth moment of X.

Example 1. Let X have the p.d.f.

$$\begin{aligned} f(x) &= \tfrac{1}{2}(x + 1), \qquad -1 < x < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Then the mean value of X is

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-1}^{1} x\frac{x+1}{2}\,dx = \frac{1}{3}$$

while the variance of X is

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \int_{-1}^{1} x^2\frac{x+1}{2}\,dx - (\tfrac{1}{3})^2 = \frac{2}{9}.$$

Example 2. If X has the p.d.f.

$$\begin{aligned} f(x) &= \frac{1}{x^2}, \qquad 1 < x < \infty, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

then the mean value of X does not exist, since

$$\begin{aligned} \int_{1}^{\infty} |x|\frac{1}{x^2}\,dx &= \lim_{b\to\infty} \int_{1}^{b} \frac{1}{x}\,dx \\ &= \lim_{b\to\infty} (\ln b - \ln 1) \end{aligned}$$

does not exist.

Example 3. It is known that the series

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$$

converges to $\pi^2/6$. Then

$$\begin{aligned} f(x) &= \frac{6}{\pi^2 x^2}, \qquad x = 1, 2, 3, \ldots, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

is the p.d.f. of a discrete type of random variable X. The m.g.f. of this distribution, if it exists, is given by

$$\begin{aligned} M(t) &= E(e^{tX}) = \sum_x e^{tx} f(x) \\ &= \sum_{x=1}^{\infty} \frac{6e^{tx}}{\pi^2 x^2}. \end{aligned}$$

The ratio test may be used to show that this series diverges if $t > 0$. Thus there does not exist a positive number h such that $M(t)$ exists for $-h < t < h$. Accordingly, the distribution having the p.d.f. $f(x)$ of this example does not have an m.g.f.

Example 4. Let X have the m.g.f. $M(t) = e^{t^2/2}$, $-\infty < t < \infty$. We can differentiate $M(t)$ any number of times to find the moments of X. However, it is instructive to consider this alternative method. The function $M(t)$ is represented by the following MacLaurin's series.

$$\begin{aligned} e^{t^2/2} &= 1 + \frac{1}{1!}\left(\frac{t^2}{2}\right) + \frac{1}{2!}\left(\frac{t^2}{2}\right)^2 + \cdots + \frac{1}{k!}\left(\frac{t^2}{2}\right)^k + \cdots \\ &= 1 + \frac{1}{2!}t^2 + \frac{(3)(1)}{4!}t^4 + \cdots + \frac{(2k-1)\cdots(3)(1)}{(2k)!}t^{2k} + \cdots. \end{aligned}$$

In general, the MacLaurin's series for $M(t)$ is

$$\begin{aligned} M(t) &= M(0) + \frac{M'(0)}{1!}t + \frac{M''(0)}{2!}t^2 + \cdots + \frac{M^{(m)}(0)}{m!}t^m + \cdots \\ &= 1 + \frac{E(X)}{1!}t + \frac{E(X^2)}{2!}t^2 + \cdots + \frac{E(X^m)}{m!}t^m + \cdots. \end{aligned}$$

Thus the coefficient of $(t^m/m!)$ in the MacLaurin's series representation of $M(t)$ is $E(X^m)$. So, for our particular $M(t)$, we have

$$E(X^{2k}) = (2k-1)(2k-3)\cdots(3)(1) = \frac{(2k)!}{2^k k!},$$

$k = 1, 2, 3, \ldots$, and $E(X^{2k-1}) = 0$, $k = 1, 2, 3, \ldots$.

Remarks. In a more advanced course, we would not work with the m.g.f. because so many distributions do not have moment-generating functions. Instead, we would let i denote the imaginary unit, t an arbitrary real, and we would define $\varphi(t) = E(e^{itX})$. This expectation exists for *every* distribution and it is called the *characteristic function* of the distribution. To see why $\varphi(t)$ exists for all real t, we note, in the continuous case, that its absolute value

$$|\varphi(t)| = \left| \int_{-\infty}^{\infty} e^{itx} f(x)\,dx \right| \le \int_{-\infty}^{\infty} |e^{itx} f(x)|\,dx.$$

However, $|f(x)| = f(x)$ since $f(x)$ is nonnegative and

$$|e^{itx}| = |\cos tx + i \sin tx| = \sqrt{\cos^2 tx + \sin^2 tx} = 1.$$

Thus

$$|\varphi(t)| \le \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Accordingly, the integral for $\varphi(t)$ exists for all real values of t. In the discrete case, a summation would replace the integral.

Every distribution has a unique characteristic function; and to each characteristic function there corresponds a unique distribution of probability. If X has a distribution with characteristic function $\varphi(t)$, then, for instance, if $E(X)$ and $E(X^2)$ exist, they are given, respectively, by $iE(X) = \varphi'(0)$ and $i^2E(X^2) = \varphi''(0)$. Readers who are familiar with complex-valued functions may write $\varphi(t) = M(it)$ and, throughout this book, may prove certain theorems in complete generality.

Those who have studied Laplace and Fourier transforms will note a similarity between these transforms and $M(t)$ and $\varphi(t)$; it is the uniqueness of these transforms that allows us to assert the uniqueness of each of the moment-generating and characteristic functions.

EXERCISES

1.89. Find the mean and variance, if they exist, of each of the following distributions.

(a) $f(x) = \dfrac{3!}{x!\,(3-x)!}\left(\dfrac{1}{2}\right)^3$, $x = 0, 1, 2, 3$, zero elsewhere.

(b) $f(x) = 6x(1-x)$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = 2/x^3$, $1 < x < \infty$, zero elsewhere.

1.90. Let $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere, be the p.d.f. of the random variable X. Find the m.g.f., the mean, and the variance of X.

1.91. For each of the following probability density functions, compute $\Pr(\mu - 2\sigma < X < \mu + 2\sigma)$.

(a) $f(x) = 6x(1 - x)$, $0 < x < 1$, zero elsewhere.
(b) $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.

1.92. If the variance of the random variable X exists, show that

$$E(X^2) \geq [E(X)]^2.$$

1.93. Let a random variable X of the continuous type have a p.d.f. $f(x)$ whose graph is symmetric with respect to $x = c$. If the mean value of X exists, show that $E(X) = c$.

Hint: Show that $E(X - c)$ equals zero by writing $E(X - c)$ as the sum of two integrals: one from $-\infty$ to c and the other from c to ∞. In the first, let $y = c - x$; and, in the second, $z = x - c$. Finally, use the symmetry condition $f(c - y) = f(c + y)$ in the first.

1.94. Let the random variable X have mean μ, standard deviation σ, and m.g.f. $M(t)$, $-h < t < h$. Show that

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

and

$$E\left\{\exp\left[t\left(\frac{X - \mu}{\sigma}\right)\right]\right\} = e^{-\mu t/\sigma} M\left(\frac{t}{\sigma}\right), \qquad -h\sigma < t < h\sigma.$$

1.95. Show that the m.g.f. of the random variable X having the p.d.f. $f(x) = \frac{1}{3}$, $-1 < x < 2$, zero elsewhere, is

$$M(t) = \frac{e^{2t} - e^{-t}}{3t}, \qquad t \neq 0,$$

$$= 1, \qquad t = 0.$$

1.96. Let X be a random variable such that $E[(X - b)^2]$ exists for all real b. Show that $E[(X - b)^2]$ is a minimum when $b = E(X)$.

1.97. Let X denote a random variable for which $E[(X - a)^2]$ exists. Give an example of a distribution of a discrete type such that this expectation is zero. Such a distribution is called a ***degenerate distribution***.

1.98. Let X be a random variable such that $K(t) = E(t^X)$ exists for all real values of t in a certain open interval that includes the point $t = 1$. Show that $K^{(m)}(1)$ is equal to the mth *factorial moment* $E[X(X - 1) \cdots (X - m + 1)]$.

1.99. Let X be a random variable. If m is a positive integer, the expectation $E[(X - b)^m]$, if it exists, is called the mth moment of the distribution about the point b. Let the first, second, and third moments of the distribution about the point 7 be 3, 11, and 15, respectively. Determine the mean μ of X, and then find the first, second, and third moments of the distribution about the point μ.

1.100. Let X be a random variable such that $R(t) = E(e^{t(X-b)})$ exists for $-h < t < h$. If m is a positive integer, show that $R^{(m)}(0)$ is equal to the mth moment of the distribution about the point b.

1.101. Let X be a random variable with mean μ and variance σ^2 such that the third moment $E[(X-\mu)^3]$ about the vertical line through μ exists. The value of the ratio $E[(X-\mu)^3]/\sigma^3$ is often used as a measure of *skewness*. Graph each of the following probability density functions and show that this measure is negative, zero, and positive for these respective distributions (which are said to be skewed to the left, not skewed, and skewed to the right, respectively).
(a) $f(x) = (x+1)/2$, $-1 < x < 1$, zero elsewhere.
(b) $f(x) = \frac{1}{2}$, $-1 < x < 1$, zero elsewhere.
(c) $f(x) = (1-x)/2$, $-1 < x < 1$, zero elsewhere.

1.102. Let X be a random variable with mean μ and variance σ^2 such that the fourth moment $E[(X-\mu)^4]$ about the vertical line through μ exists. The value of the ratio $E[(X-\mu)^4]/\sigma^4$ is often used as a measure of *kurtosis*. Graph each of the following probability density functions and show that this measure is smaller for the first distribution.
(a) $f(x) = \frac{1}{2}$, $-1 < x < 1$, zero elsewhere.
(b) $f(x) = 3(1-x^2)/4$, $-1 < x < 1$, zero elsewhere.

1.103. Let the random variable X have p.d.f.

$$\begin{aligned} f(x) &= p, \quad && x = -1, 1, \\ &= 1 - 2p, && x = 0, \\ &= 0 && \text{elsewhere,} \end{aligned}$$

where $0 < p < \frac{1}{2}$. Find the measure of kurtosis as a function of p. Determine its value when $p = \frac{1}{3}$, $p = \frac{1}{5}$, $p = \frac{1}{10}$, and $p = \frac{1}{100}$. Note that the kurtosis increases as p decreases.

1.104. Let $\psi(t) = \ln M(t)$, where $M(t)$ is the m.g.f. of a distribution. Prove that $\psi'(0) = \mu$ and $\psi''(0) = \sigma^2$.

1.105. Find the mean and the variance of the distribution that has the distribution function

$$\begin{aligned} F(x) &= 0, \quad && x < 0, \\ &= \frac{x}{8}, && 0 \le x < 2, \\ &= \frac{x^2}{16}, && 2 \le x < 4, \\ &= 1, && 4 \le x. \end{aligned}$$

1.106. Find the moments of the distribution that has m.g.f. $M(t) = (1 - t)^{-3}$, $t < 1$.

Hint: Find the MacLaurin's series for $M(t)$.

1.107. Let X be a random variable of the continuous type with p.d.f. $f(x)$, which is positive provided $0 < x < b < \infty$, and is equal to zero elsewhere. Show that

$$E(X) = \int_0^b [1 - F(x)]\, dx,$$

where $F(x)$ is the distribution function of X.

1.108. Let X be a random variable of the discrete type with p.d.f. $f(x)$ that is positive on the nonnegative integers and is equal to zero elsewhere. Show that

$$E(X) = \sum_{x=0}^{\infty} [1 - F(x)],$$

where $F(x)$ is the distribution function of X.

1.109. Let X have the p.d.f. $f(x) = 1/k$, $x = 1, 2, \ldots, k$, zero elsewhere. Show that the m.g.f. is

$$M(t) = \frac{e^t(1 - e^{kt})}{k(1 - e^t)}, \qquad t \neq 0,$$
$$= 1, \qquad t = 0.$$

1.110. Let X have the distribution function $F(x)$ that is a mixture of the continuous and discrete types, namely

$$F(x) = 0, \qquad x < 0,$$
$$= \frac{x + 1}{4}, \qquad 0 \le x < 1,$$
$$= 1, \qquad 1 \le x.$$

Find $\mu = E(X)$ and $\sigma^2 = \text{var}(X)$.

Hint: Determine that part of the p.d.f. associated with each of the discrete and continuous types, and then sum for the discrete part and integrate for the continuous part.

1.111. Consider k continuous-type distributions with the following characteristics: p.d.f. $f_i(x)$, mean μ_i, and variance σ_i^2, $i = 1, 2, \ldots, k$. If $c_i \ge 0$, $i = 1, 2, \ldots, k$, and $c_1 + c_2 + \cdots + c_k = 1$, show that the mean and the variance of the distribution having p.d.f. $c_1 f_1(x) + \cdots + c_k f_k(x)$ are $\mu = \sum_{i=1}^{k} c_i \mu_i$ and $\sigma^2 = \sum_{i=1}^{k} c_i[\sigma_i^2 + (\mu_i - \mu)^2]$, respectively.

1.10 Chebyshev's Inequality

In this section we prove a theorem that enables us to find upper (or lower) bounds for certain probabilities. These bounds, however, are not necessarily close to the exact probabilities and, accordingly, we ordinarily do not use the theorem to approximate a probability. The principal uses of the theorem and a special case of it are in theoretical discussions in other chapters.

Theorem 6. *Let $u(X)$ be a nonnegative function of the random variable X. If $E[u(X)]$ exists, then, for every positive constant c,*

$$\Pr[u(X) \geq c] \leq \frac{E[u(X)]}{c}.$$

Proof. The proof is given when the random variable X is of the continuous type; but the proof can be adapted to the discrete case if we replace integrals by sums. Let $A = \{x : u(x) \geq c\}$ and let $f(x)$ denote the p.d.f. of X. Then

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\,dx = \int_A u(x)f(x)\,dx + \int_{A^*} u(x)f(x)\,dx.$$

Since each of the integrals in the extreme right-hand member of the preceding equation is nonnegative, the left-hand member is greater than or equal to either of them. In particular,

$$E[u(X)] \geq \int_A u(x)f(x)\,dx.$$

However, if $x \in A$, then $u(x) \geq c$; accordingly, the right-hand member of the preceding inequality is not increased if we replace $u(x)$ by c. Thus

$$E[u(X)] \geq c \int_A f(x)\,dx.$$

Since

$$\int_A f(x)\,dx = \Pr(X \in A) = \Pr[u(X) \geq c],$$

it follows that

$$E[u(X)] \geq c \Pr[u(X) \geq c],$$

which is the desired result.

The preceding theorem is a generalization of an inequality that is often called *Chebyshev's inequality*. This inequality will now be established.

Theorem 7. Chebyshev's Inequality. *Let the random variable X have a distribution of probability about which we assume only that there is a finite variance σ^2. This, of course, implies that there is a mean μ. Then for every $k > 0$,*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2},$$

or, equivalently,

$$\Pr(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Proof. In Theorem 6 take $u(X) = (X - \mu)^2$ and $c = k^2\sigma^2$. Then we have

$$\Pr[(X - \mu)^2 \geq k^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2}.$$

Since the numerator of the right-hand member of the preceding inequality is σ^2, the inequality may be written

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2},$$

which is the desired result. Naturally, we would take the positive number k to be greater than 1 to have an inequality of interest.

It is seen that the number $1/k^2$ is an upper bound for the probability $\Pr(|X - \mu| \geq k\sigma)$. In the following example this upper bound and the exact value of the probability are compared in special instances.

Example 1. Let X have the p.d.f.

$$f(x) = \frac{1}{2\sqrt{3}}, \qquad -\sqrt{3} < x < \sqrt{3},$$
$$= 0 \qquad \text{elsewhere.}$$

Here $\mu = 0$ and $\sigma^2 = 1$. If $k = \frac{3}{2}$, we have the exact probability

$$\Pr(|X - \mu| \geq k\sigma) = \Pr\left(|X| \geq \frac{3}{2}\right) = 1 - \int_{-3/2}^{3/2} \frac{1}{2\sqrt{3}}\,dx = 1 - \frac{\sqrt{3}}{2}.$$

By Chebyshev's inequality, the preceding probability has the upper bound $1/k^2 = \frac{4}{9}$. Since $1 - \sqrt{3}/2 = 0.134$, approximately, the exact probability in this case is considerably less than the upper bound $\frac{4}{9}$. If we take $k = 2$, we have the exact probability $\Pr(|X - \mu| \geq 2\sigma) = \Pr(|X| \geq 2) = 0$. This again is considerably less than the upper bound $1/k^2 = \frac{1}{4}$ provided by Chebyshev's inequality.

In each of the instances in the preceding example, the probability $\Pr(|X - \mu| \geq k\sigma)$ and its upper bound $1/k^2$ differ considerably. This suggests that this inequality might be made sharper. However, if we want an inequality that holds for every $k > 0$ and holds for all random variables having finite variance, such an improvement is impossible, as is shown by the following example.

Example 2. Let the random variable X of the discrete type have probabilities $\frac{1}{8}, \frac{6}{8}, \frac{1}{8}$ at the points $x = -1, 0, 1$, respectively. Here $\mu = 0$ and $\sigma^2 = \frac{1}{4}$. If $k = 2$, then $1/k^2 = \frac{1}{4}$ and $\Pr(|X - \mu| \geq k\sigma) = \Pr(|X| \geq 1) = \frac{1}{4}$. That is, the probability $\Pr(|X - \mu| \geq k\sigma)$ here attains the upper bound $1/k^2 = \frac{1}{4}$. Hence the inequality cannot be improved without further assumptions about the distribution of X.

EXERCISES

1.112. Let X be a random variable with mean μ and let $E[(X - \mu)^{2k}]$ exist. Show, with $d > 0$, that $\Pr(|X - \mu| \geq d) \leq E[(X - \mu)^{2k}]/d^{2k}$. This is essentially Chebyshev's inequality when $k = 1$. The fact that this holds for all $k = 1, 2, 3, \ldots$, when those $(2k)$th moments exist, usually provides a much smaller upper bound for $\Pr(|X - \mu| \geq d)$ than does Chebyshev's result.

1.113. Let X be a random variable such that $\Pr(X \leq 0) = 0$ and let $\mu = E(X)$ exist. Show that $\Pr(X \geq 2\mu) \leq \frac{1}{2}$.

1.114. If X is a random variable such that $E(X) = 3$ and $E(X^2) = 13$, use Chebyshev's inequality to determine a lower bound for the probability $\Pr(-2 < X < 8)$.

1.115. Let X be a random variable with m.g.f. $M(t)$, $-h < t < h$. Prove that

$$\Pr(X \geq a) \leq e^{-at}M(t), \qquad 0 < t < h,$$

and that

$$\Pr(X \leq a) \leq e^{-at}M(t), \qquad -h < t < 0.$$

Hint: Let $u(x) = e^{tx}$ and $c = e^{ta}$ in Theorem 6. *Note.* These results imply that $\Pr(X \geq a)$ and $\Pr(X \leq a)$ are less than the respective greatest lower bounds for $e^{-at}M(t)$ when $0 < t < h$ and when $-h < t < 0$.

1.116. The m.g.f. of X exists for all real values of t and is given by

$$M(t) = \frac{e^t - e^{-t}}{2t}, \qquad t \neq 0, \quad M(0) = 1.$$

Use the results of the preceding exercise to show that $\Pr(X \geq 1) = 0$ and $\Pr(X \leq -1) = 0$. Note that here h is infinite.

ADDITIONAL EXERCISES

1.117. Players A and B play a sequence of independent games. Player A throws a die first and wins on a "six." If he fails, B throws and wins on a "five" or "six." If he fails, A throws again and wins on a "four," "five," or "six." And so on. Find the probability of each player winning the sequence.

1.118. Let X be the number of gallons of ice cream that is requested at a certain store on a hot summer day. Let us assume that the p.d.f. of X is $f(x) = 12x(1000 - x)^2/10^{12}$, $0 < x < 1000$, zero elsewhere. How many gallons of ice cream should the store have on hand each of these days, so that the probability of exhausting its supply on a particular day is 0.05?

1.119. Find the 25th percentile of the distribution having p.d.f. $f(x) = |x|/4$, $-2 < x < 2$, zero elsewhere.

1.120. Let A_1, A_2, A_3 be independent events with probabilities $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, respectively. Compute $\Pr(A_1 \cup A_2 \cup A_3)$.

1.121. From a bowl containing 5 red, 3 white, and 7 blue chips, select 4 at random and without replacement. Compute the conditional probability of 1 red, 0 white, and 3 blue chips, given that there are at least 3 blue chips in this sample of 4 chips.

1.122. Let the three independent events A, B, and C be such that $P(A) = P(B) = P(C) = \frac{1}{4}$. Find $P[(A^* \cap B^*) \cup C]$.

1.123. Person A tosses a coin and then person B rolls a die. This is repeated independently until a head or one of the numbers 1, 2, 3, 4 appears, at which time the game is stopped. Person A wins with the head and B wins with one of the numbers 1, 2, 3, 4. Compute the probability that A wins the game.

1.124. Find the mean and variance of the random variable X having distribution function

$$\begin{aligned} F(x) &= 0, & x < 0, \\ &= \frac{x}{4}, & 0 \leq x < 1, \end{aligned}$$

$$= \frac{x^2}{4}, \qquad 1 \le x < 2,$$

$$= 1, \qquad 2 \le x.$$

1.125. Let X be a random variable having distribution function

$$F(x) = 0, \qquad x < 0,$$
$$= 2x^2, \qquad 0 \le x < \tfrac{1}{2},$$
$$= 1 - 2(1-x)^2, \qquad \tfrac{1}{2} \le x < \tfrac{3}{4},$$
$$= 1, \qquad \tfrac{3}{4} \le x.$$

Find $\Pr(\frac{1}{4} < X < \frac{5}{8})$ and the variance of the distribution.
Hint: Note that there is a step in $F(x)$.

1.126. Bowl I contains 7 red and 3 white chips and bowl II has 4 red and 6 white chips. Two chips are selected at random and without replacement from I and transferred to II. Three chips are then selected at random and without replacement from II.
(a) What is the probability that all three are white?
(b) Given that three white chips are selected from II, what is the conditional probability that two white chips were transferred from I?

1.127. A bowl contains ten chips numbered 1, 2, . . . , 10, respectively. Five chips are drawn at random, one at a time, and without replacement. What is the probability that exactly two even-numbered chips are drawn and they occur on even-numbered draws?

1.128. Let $E(X^r) = \dfrac{1}{r+1}$, $r = 1, 2, 3, \ldots$. Find the series representation for the m.g.f. of X. Sum this series.

1.129. Let X have the p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Compute the probability that X is at least $\frac{3}{4}$ given that X is at least $\frac{1}{2}$.

1.130. Divide a line segment into two parts by selecting a point at random. Find the probability that the larger segment is at least three times the shorter. Assume a uniform distribution.

1.131. Three chips are selected at random and without replacement from a bowl containing 5 white, 4 black, and 7 red chips. Find the probability that these three chips are alike in color.

1.132. Factories A, B, and C produce, respectively, 20, 30, and 50% of a certain company's output. The items produced in A, B, and C are 1, 2, and 3 percent defective, respectively. We observe one item from the company's output at random and find it defective. What is the conditional probability that the item was from A?

1.133. The probabilities that the independent events A, B, and C will occur are $\frac{3}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. What is the probability that at least one of the three events will happen?

1.134. A person bets 1 dollar to b dollars that he can draw two cards from an ordinary deck without replacement and that they will be of the same suit. Find b so that the bet will be fair.

1.135. A bowl contains 6 chips: 4 are red and 2 are white. Three chips are selected at random and without replacement; then a coin is tossed a number of independent times that is equal to the number of red chips in this sample of 3. For example, if we have 2 red and 1 white, the coin is tossed twice. Given that one head results, compute the conditional probability that the sample contains 1 red and 2 white.

CHAPTER 2

Multivariate Distributions

2.1 Distributions of Two Random Variables

We begin the discussion of two random variables with the following example. A coin is to be tossed three times and our interest is in the ordered number pair (number of H's on first two tosses, number of H's on all three tosses), where H and T represent, respectively, heads and tails. Thus the sample space is $\mathscr{C} = \{c : c = c_i, i = 1, 2, \ldots, 8\}$, where c_1 is TTT, c_2 is TTH, c_3 is THT, c_4 is HTT, c_5 is THH, c_6 is HTH, c_7 is HHT, and c_8 is HHH. Let X_1 and X_2 be two functions such that $X_1(c_1) = X_1(c_2) = 0$, $X_1(c_3) = X_1(c_4) = X_1(c_5) = X_1(c_6) = 1$, $X_1(c_7) = X_1(c_8) = 2$; and $X_2(c_1) = 0$, $X_2(c_2) = X_2(c_3) = X_2(c_4) = 1$, $X_2(c_5) = X_2(c_6) = X_2(c_7) = 2$, $X_2(c_8) = 3$. Thus X_1 and X_2 are real-valued functions defined on the sample space $\mathscr{C}$, which take us from that sample space to the space of ordered number pairs

$$\mathscr{A} = \{(0, 0), (0, 1), (1, 1), (1, 2), (2, 2), (2, 3)\}.$$

Thus X_1 and X_2 are two random variables defined on the space $\mathscr{C}$, and, in this example, the space of these random variables is the two-dimensional set $\mathscr{A}$ given immediately above. We now formulate the definition of the space of two random variables.

Definition 1. Given a random experiment with a sample space $\mathscr{C}$. Consider two random variables X_1 and X_2, which assign to each element c of $\mathscr{C}$ one and only one ordered pair of numbers $X_1(c) = x_1$, $X_2(c) = x_2$. The *space* of X_1 and X_2 is the set of ordered pairs $\mathscr{A} = \{(x_1, x_2) : x_1 = X_1(c), x_2 = X_2(c), c \in \mathscr{C}\}$.

Let $\mathscr{A}$ be the space associated with the two random variables X_1 and X_2 and let A be a subset of $\mathscr{A}$. As in the case of one random variable, we shall speak of the event A. We wish to define the probability of the event A, which we denote by $\Pr[(X_1, X_2) \in A]$. Take $C = \{c : c \in \mathscr{C}$ and $[X_1(c), X_2(c)] \in A\}$, where $\mathscr{C}$ is the sample space. We then define $\Pr[(X_1, X_2) \in A] = P(C)$, where P is the probability set function defined for subsets C of $\mathscr{C}$. Here again we could denote $\Pr[(X_1, X_2) \in A]$ by the probability set function $P_{X_1,X_2}(A)$; but, with our previous convention, we simply write

$$P(A) = \Pr[(X_1, X_2) \in A].$$

Again it is important to observe that this function is a probability set function defined for subsets A of the space $\mathscr{A}$.

Let us return to the example in our discussion of two random variables. Consider the subset A of $\mathscr{A}$, where $A = \{(1, 1), (1, 2)\}$. To compute $\Pr[(X_1, X_2) \in A] = P(A)$, we must include as elements of C all outcomes in $\mathscr{C}$ for which the random variables X_1 and X_2 take values (x_1, x_2) which are elements of A. Now $X_1(c_3) = 1$, $X_2(c_3) = 1$, $X_1(c_4) = 1$, and $X_2(c_4) = 1$. Also, $X_1(c_5) = 1$, $X_2(c_5) = 2$, $X_1(c_6) = 1$, and $X_2(c_6) = 2$. Thus $P(A) = \Pr[(X_1, X_2) \in A] = P(C)$, where $C = \{c_3, c_4, c_5, \text{or } c_6\}$. Suppose that our probability set function $P(C)$ assigns a probability of $\frac{1}{8}$ to each of the eight elements of $\mathscr{C}$. This assignment seems reasonable if $P(\text{T}) = P(\text{H}) = \frac{1}{2}$ and the tosses are independent. For illustration,

$$P(\{c_1\}) = \Pr(\text{TTT}) = (\tfrac{1}{2})(\tfrac{1}{2})(\tfrac{1}{2}) = \tfrac{1}{8}.$$

Then $P(A)$, which can be written as $\Pr(X_1 = 1, X_2 = 1 \text{ or } 2)$, is equal to $\frac{4}{8} = \frac{1}{2}$. It is left for the reader to show that we can tabulate the probability, which is then assigned to each of the elements of $\mathscr{A}$, with the following result:

(x_1, x_2)	(0, 0)	(0, 1)	(1, 1)	(1, 2)	(2, 2)	(2, 3)
$\Pr[(X_1, X_2) = (x_1, x_2)]$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

This table depicts the distribution of probability over the elements of $\mathscr{A}$, the space of the random variables X_1 and X_2.

Again in statistics we are more interested in the space $\mathscr{A}$ of two random variables, say X and Y, than that of $\mathscr{C}$. Moreover, the notion of the p.d.f. of one random variable X can be extended to the notion of the p.d.f. of two or more random variables. Under certain restrictions on the space $\mathscr{A}$ and the function $f > 0$ on $\mathscr{A}$ (restrictions that will not be enumerated here), we say that the two random variables X and Y are of the discrete type or of the continuous type, and have a distribution of that type, according as the probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed as

$$P(A) = \Pr[(X, Y) \in A] = \sum_A\sum f(x, y),$$

or as

$$P(A) = \Pr[(X, Y) \in A] = \int_A\int f(x, y)\, dx\, dy.$$

In either case f is called the p.d.f. of the two random variables X and Y. Of necessity, $P(\mathscr{A}) = 1$ in each case.

We may extend the definition of a p.d.f. $f(x, y)$ over the entire xy-plane by using zero elsewhere. We shall do this consistently so that tedious, repetitious references to the space $\mathscr{A}$ can be avoided. Once this is done, we replace

$$\int_{\mathscr{A}}\int f(x, y)\, dx\, dy \quad \text{by} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy.$$

Similarly, after extending the definition of a p.d.f. of the discrete type, we replace

$$\sum_{\mathscr{A}}\sum f(x, y) \quad \text{by} \quad \sum_y\sum_x f(x, y).$$

In accordance with this convention (of extending the definition of a p.d.f.), it is seen that a point function f, whether in one or two variables, essentially satisfies the conditions of being a p.d.f. if (a) f

is defined and is nonnegative for all real values of its argument(s) and if (b) its integral [for the continuous type of random variable(s)], or its sum [for the discrete type of random variable(s)] over all real values of its arguments(s) is 1.

Finally, if a p.d.f. in one or more variables is explicitly defined, we can see by inspection whether the random variables are of the continuous or discrete type. For example, it seems obvious that the p.d.f.

$$f(x, y) = \frac{9}{4^{x+y}}, \qquad x = 1, 2, 3, \ldots, \quad y = 1, 2, 3, \ldots,$$
$$= 0 \qquad \text{elsewhere},$$

is a p.d.f. of two discrete-type random variables X and Y, whereas the p.d.f.

$$f(x, y) = 4xye^{-x^2-y^2}, \qquad 0 < x < \infty, \quad 0 < y < \infty,$$
$$= 0 \qquad \text{elsewhere},$$

is clearly a p.d.f. of two continuous-type random variables X and Y. In such cases it seems unnecessary to specify which of the two simpler types of random variables is under consideration.

Example 1. Let

$$f(x, y) = 6x^2y, \qquad 0 < x < 1, \quad 0 < y < 1,$$
$$= 0 \qquad \text{elsewhere},$$

be the p.d.f. of two random variables X and Y, which must be of the continuous type. We have, for instance,

$$\Pr(0 < X < \tfrac{3}{4}, \tfrac{1}{3} < Y < 2) = \int_{1/3}^{2}\int_{0}^{3/4} f(x, y)\,dx\,dy$$
$$= \int_{1/3}^{1}\int_{0}^{3/4} 6x^2y\,dx\,dy + \int_{1}^{2}\int_{0}^{3/4} 0\,dx\,dy$$
$$= \tfrac{3}{8} + 0 = \tfrac{3}{8}.$$

Note that this probability is the volume under the surface $f(x, y) = 6x^2y$ and above the rectangular set $\{(x, y) : 0 < x < \frac{3}{4}, \frac{1}{3} < y < 1\}$ in the xy-plane.

Let the random variables X and Y have the probability set function $P(A)$, where A is a two-dimensional set. If A is the unbounded set $\{(u, v) : u \le x, v \le y\}$, where x and y are real numbers, we have

$$P(A) = \Pr[(X, Y) \in A] = \Pr(X \le x, Y \le y).$$

This function of the point (x, y) is called the *distribution function* of X and Y and is denoted by

$$F(x, y) = \Pr(X \le x, Y \le y).$$

If X and Y are random variables of the continuous type that have p.d.f. $f(x, y)$, then

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv.$$

Accordingly, at points of continuity of $f(x, y)$, we have

$$\frac{\partial^2 F(x, y)}{\partial x\, \partial y} = f(x, y).$$

It is left as an exercise to show, in every case, that

$$\Pr(a < X \le b, c < Y \le d) = F(b, d) - F(b, c) - F(a, d) + F(a, c),$$

for all real constants $a < b$, $c < d$.

Consider next an experiment in which a person chooses at random a point (X, Y) from the unit square $\mathscr{C} = \mathscr{A} = \{(x, y): 0 < x < 1, 0 < y < 1\}$. Suppose that our interest is not in X or in Y but in $Z = X + Y$. Once a suitable probability model has been adopted, we shall see how to find the p.d.f. of Z. To be specific, let the nature of the random experiment be such that it is reasonable to *assume* that the distribution of probability over the unit square is uniform. Then the p.d.f. of X and Y may be written

$$\begin{aligned} f(x, y) &= 1, \qquad 0 < x < 1, \quad 0 < y < 1, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

and this describes the probability model. Now let the distribution function of Z be denoted by $G(z) = \Pr(X + Y \le z)$. Then

$$\begin{aligned} G(z) &= 0, \qquad z < 0, \\ &= \int_0^z \int_0^{z-x} dy\, dx = \frac{z^2}{2}, \qquad 0 \le z < 1, \\ &= 1 - \int_{z-1}^1 \int_{z-x}^1 dy\, dx = 1 - \frac{(2-z)^2}{2}, \qquad 1 \le z < 2, \\ &= 1, \qquad 2 \le z. \end{aligned}$$

Since $G'(z)$ exists for all values of z, the p.d.f. of Z may then be written

$$
\begin{aligned}
g(z) &= z, \qquad 0 < z < 1, \\
&= 2 - z, \qquad 1 \le z < 2, \\
&= 0 \qquad \text{elsewhere.}
\end{aligned}
$$

It is clear that a different choice of the p.d.f. $f(x, y)$ that describes the probability model will, in general, lead to a different p.d.f. of Z.

Let $f(x_1, x_2)$ be the p.d.f. of two random variables X_1 and X_2. From this point on, for emphasis and clarity, we shall call a p.d.f. or a distribution function a *joint* p.d.f. or a *joint* distribution function when more than one random variable is involved. Thus $f(x_1, x_2)$ is the joint p.d.f. of the random variables X_1 and X_2. Consider the event $a < X_1 < b, a < b$. This event can occur when and only when the event $a < X_1 < b, \ -\infty < X_2 < \infty$ occurs; that is, the two events are equivalent, so that they have the same probability. But the probability of the latter event has been defined and is given by

$$\Pr(a < X_1 < b, -\infty < X_2 < \infty) = \int_a^b \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2\, dx_1$$

for the continuous case, and by

$$\Pr(a < X_1 < b_1, -\infty < X_2 < \infty) = \sum_{a < x_1 < b} \sum_{x_2} f(x_1, x_2)$$

for the discrete case. Now each of

$$\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \qquad \text{and} \qquad \sum_{x_2} f(x_1, x_2)$$

is a function of x_1 alone, say $f_1(x_1)$. Thus, for every $a < b$, we have

$$
\begin{aligned}
\Pr(a < X_1 < b) &= \int_a^b f_1(x_1)\, dx_1 \qquad \text{(continuous case)}, \\
&= \sum_{a < x_1 < b} f_1(x_1) \qquad \text{(discrete case)},
\end{aligned}
$$

so that $f_1(x_1)$ is the p.d.f. of X_1 alone. Since $f_1(x_1)$ is found by summing (or integrating) the joint p.d.f. $f(x_1, x_2)$ over all x_2 for a fixed x_1, we can think of recording this sum in the "margin" of the

x_1x_2-plane. Accordingly, $f_1(x_1)$ is called the marginal p.d.f. of X_1. In like manner

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\,dx_1 \qquad \text{(continuous case)},$$

$$= \sum_{x_1} f(x_1, x_2) \qquad \text{(discrete case)},$$

is called the marginal p.d.f. of X_2.

Example 2. Consider a random experiment that consists of drawing at random one chip from a bowl containing 10 chips of the same shape and size. Each chip has an ordered pair of numbers on it: one with (1, 1), one with (2, 1), two with (3, 1), one with (1, 2), two with (2, 2), and three with (3, 2). Let the random variables X_1 and X_2 be defined as the respective first and second values of the ordered pair. Thus the joint p.d.f. $f(x_1, x_2)$ of X_1 and X_2 can be given by the following table, with $f(x_1, x_2)$ equal to zero elsewhere.

	x_1			
x_2	1	2	3	$f_2(x_2)$
1	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{4}{10}$
2	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{6}{10}$
$f_1(x_1)$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	

The joint probabilities have been summed in each row and each column and these sums recorded in the margins to give the marginal probability density functions of X_1 and X_2, respectively. Note that it is not necessary to have a formula for $f(x_1, x_2)$ to do this.

Example 3. Let X_1 and X_2 have the joint p.d.f.

$$f(x_1, x_2) = x_1 + x_2, \qquad 0 < x_1 < 1, \quad 0 < x_2 < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

The marginal p.d.f. of X_1 is

$$f_1(x_1) = \int_0^1 (x_1 + x_2)\,dx_2 = x_1 + \tfrac{1}{2}, \qquad 0 < x_1 < 1,$$

zero elsewhere, and the marginal p.d.f. of X_2 is

$$f_2(x_2) = \int_0^1 (x_1 + x_2)\,dx_1 = \tfrac{1}{2} + x_2, \qquad 0 < x_2 < 1,$$

zero elsewhere. A probability like $\Pr(X_1 \le \frac{1}{2})$ can be computed from either $f_1(x_1)$ or $f(x_1, x_2)$ because

$$\int_0^{1/2}\int_0^1 f(x_1, x_2)\,dx_2\,dx_1 = \int_0^{1/2} f_1(x_1)\,dx_1 = \tfrac{3}{8}.$$

However to find a probability like $\Pr(X_1 + X_2 \le 1)$, we must use the joint p.d.f. $f(x_1, x_2)$ as follows:

$$\int_0^1\int_0^{1-x_1}(x_1 + x_2)\,dx_2\,dx_1 = \int_0^1\left[x_1(1 - x_1) + \frac{(1 - x_1)^2}{2}\right]dx_1$$

$$= \int_0^1\left(\frac{1}{2} - \frac{1}{2}x_1^2\right)dx_1 = \frac{1}{3}.$$

This latter probability is the volume under the surface $f(x_1, x_2) = x_1 + x_2$ above the set $\{(x_1, x_2) : 0 < x_1, 0 < x_2, x_1 + x_2 \le 1\}$.

EXERCISES

2.1. Let $f(x_1, x_2) = 4x_1x_2$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere, be the p.d.f. of X_1 and X_2. Find $\Pr(0 < X_1 < \frac{1}{2}, \frac{1}{4} < X_2 < 1)$, $\Pr(X_1 = X_2)$, $\Pr(X_1 < X_2)$, and $\Pr(X_1 \le X_2)$.

Hint: Recall that $\Pr(X_1 = X_2)$ would be the volume under the surface $f(x_1, x_2) = 4x_1x_2$ and above the line segment $0 < x_1 = x_2 < 1$ in the x_1x_2-plane.

2.2. Let $A_1 = \{(x, y) : x \le 2, y \le 4\}$, $A_2 = \{(x, y) : x \le 2, y \le 1\}$, $A_3 = \{(x, y) : x \le 0, y \le 4\}$, and $A_4 = \{(x, y) : x \le 0, y \le 1\}$ be subsets of the space $\mathscr{A}$ of two random variables X and Y, which is the entire two-dimensional plane. If $P(A_1) = \frac{7}{8}$, $P(A_2) = \frac{4}{8}$, $P(A_3) = \frac{3}{8}$, and $P(A_4) = \frac{2}{8}$, find $P(A_5)$, where $A_5 = \{(x, y) : 0 < x \le 2, 1 < y \le 4\}$.

2.3. Let $F(x, y)$ be the distribution function of X and Y. Show that $\Pr(a < X \le b, c < Y \le d) = F(b, d) - F(b, c) - F(a, d) + F(a, c)$, for all real constants $a < b$, $c < d$.

2.4. Show that the function $F(x, y)$ that is equal to 1 provided that $x + 2y \ge 1$, and that is equal to zero provided that $x + 2y < 1$, cannot be a distribution function of two random variables.

Hint: Find four numbers $a < b$, $c < d$, so that

$$F(b, d) - F(a, d) - F(b, c) + F(a, c)$$

is less than zero.

2.5. Given that the nonnegative function $g(x)$ has the property that

$$\int_0^\infty g(x)\,dx = 1.$$

Show that

$$f(x_1, x_2) = [2g(\sqrt{x_1^2 + x_2^2})]/(\pi\sqrt{x_1^2 + x_2^2}),\ 0 < x_1 < \infty,\ 0 < x_2 < \infty,$$

zero elsewhere, satisfies the conditions of being a p.d.f. of two continuous-type random variables X_1 and X_2.

Hint: Use polar coordinates.

2.6. Let $f(x, y) = e^{-x-y}$, $0 < x < \infty$, $0 < y < \infty$, zero elsewhere, be the p.d.f. of X and Y. Then if $Z = X + Y$, compute $\Pr(Z \le 0)$, $\Pr(Z \le 6)$, and, more generally, $\Pr(Z \le z)$, for $0 < z < \infty$. What is the p.d.f. of Z?

2.7. Let X and Y have the p.d.f. $f(x, y) = 1$, $0 < x < 1$, $0 < y < 1$, zero elsewhere. Find the p.d.f. of the product $Z = XY$.

2.8. Let 13 cards be taken, at random and without replacement, from an ordinary deck of playing cards. If X is the number of spades in these 13 cards, find the p.d.f. of X. If, in addition, Y is the number of hearts in these 13 cards, find the probability $\Pr(X = 2, Y = 5)$. What is the joint p.d.f. of X and Y?

2.9. Let the random variables X_1 and X_2 have the joint p.d.f. described as follows:

(x_1, x_2)	(0, 0)	(0, 1)	(0, 2)	(1, 0)	(1, 1)	(1, 2)
$f(x_1, x_2)$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{12}$

and $f(x_1, x_2)$ is equal to zero elsewhere.

(a) Write these probabilities in a rectangular array as in Example 2, recording each marginal p.d.f. in the "margins."

(b) What is $\Pr(X_1 + X_2 = 1)$?

2.10. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = 15x_1^2x_2$, $0 < x_1 < x_2 < 1$, zero elsewhere. Find each marginal p.d.f. and compute $\Pr(X_1 + X_2 \le 1)$.

Hint: Graph the space of X_1 and X_2 and carefully choose the limits of integration in determining each marginal p.d.f.

2.2 Conditional Distributions and Expectations

We shall now discuss the notion of a conditional p.d.f. Let X_1 and X_2 denote random variables of the discrete type which have the joint p.d.f. $f(x_1, x_2)$ which is positive on $\mathscr{A}$ and is zero elsewhere. Let $f_1(x_1)$ and $f_2(x_2)$ denote, respectively, the marginal probability density functions of X_1 and X_2. Take A_1 to be the set $A_1 = \{(x_1, x_2) : x_1 = x_1', -\infty < x_2 < \infty\}$, where x_1' is such that $P(A_1) = \Pr(X_1 = x_1') = f_1(x_1') > 0$, and take A_2 to be the set

$A_2 = \{(x_1, x_2) : -\infty < x_1 < \infty, x_2 = x_2'\}$. Then, by definition, the conditional probability of the event A_2, given the event A_1, is

$$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{\Pr(X_1 = x_1', X_2 = x_2')}{\Pr(X_1 = x_1')} = \frac{f(x_1', x_2')}{f_1(x_1')}.$$

That is, if (x_1, x_2) is any point at which $f_1(x_1) > 0$, the conditional probability that $X_2 = x_2$, given that $X_1 = x_1$, is $f(x_1, x_2)/f_1(x_1)$. With x_1 held fast, and with $f_1(x_1) > 0$, this function of x_2 satisfies the conditions of being a p.d.f. of a discrete type of random variable X_2 because $f(x_1, x_2)/f_1(x_1)$ is nonnegative and

$$\sum_{x_2} \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{1}{f_1(x_1)} \sum_{x_2} f(x_1, x_2) = \frac{f_1(x_1)}{f_1(x_1)} = 1.$$

We now define the symbol $f_{2|1}(x_2|x_1)$ by the relation

$$f_{2|1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \qquad f_1(x_1) > 0,$$

and we call $f_{2|1}(x_2|x_1)$ the *conditional p.d.f.* of the discrete type of random variable X_2, given that the discrete type of random variable $X_1 = x_1$. In a similar manner we define the symbol $f_{1|2}(x_1|x_2)$ by the relation

$$f_{1|2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \qquad f_2(x_2) > 0,$$

and we call $f_{1|2}(x_1|x_2)$ the conditional p.d.f. of the discrete type of random variable X_1, given that the discrete type of random variable $X_2 = x_2$.

Now let X_1 and X_2 denote random variables of the continuous type that have the joint p.d.f. $f(x_1, x_2)$ and the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. We shall use the results of the preceding paragraph to motivate a definition of a conditional p.d.f. of a continuous type of random variable. When $f_1(x_1) > 0$, we define the symbol $f_{2|1}(x_2|x_1)$ by the relation

$$f_{2|1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}.$$

In this relation, x_1 is to be thought of as having a fixed (but any fixed) value for which $f_1(x_1) > 0$. It is evident that $f_{2|1}(x_2|x_1)$ is nonnegative and that

$$\begin{aligned}\int_{-\infty}^{\infty} f_{2|1}(x_2|x_1)\,dx_2 &= \int_{-\infty}^{\infty} \frac{f(x_1, x_2)}{f_1(x_1)}\,dx_2 \\ &= \frac{1}{f_1(x_1)} \int_{-\infty}^{\infty} f(x_1, x_2)\,dx_2 \\ &= \frac{1}{f_1(x_1)} f_1(x_1) = 1.\end{aligned}$$

That is, $f_{2|1}(x_2|x_1)$ has the properties of a p.d.f. of one continuous type of random variable. It is called the *conditional p.d.f.* of the continuous type of random variable X_2, given that the continuous type of random variable X_1 has the value x_1. When $f_2(x_2) > 0$, the conditional p.d.f. of the continuous type of random variable X_1, given that the continuous type of random variable X_2 has the value x_2, is defined by

$$f_{1|2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \qquad f_2(x_2) > 0.$$

Since each of $f_{2|1}(x_2|x_1)$ and $f_{1|2}(x_1|x_2)$ is a p.d.f. of one random variable (whether of the discrete or the continuous type), each has all the properties of such a p.d.f. Thus we can compute probabilities and mathematical expectations. If the random variables are of the continuous type, the probability

$$\Pr(a < X_2 < b|X_1 = x_1) = \int_a^b f_{2|1}(x_2|x_1)\,dx_2$$

is called "the conditional probability that $a < X_2 < b$, given that $X_1 = x_1$." If there is no ambiguity, this may be written in the form $\Pr(a < X_2 < b|x_1)$. Similarly, the conditional probability that $c < X_1 < d$, given $X_2 = x_2$, is

$$\Pr(c < X_1 < d|X_2 = x_2) = \int_c^d f_{1|2}(x_1|x_2)\,dx_1.$$

If $u(X_2)$ is a function of X_2, the expectation

$$E[u(X_2)|x_1] = \int_{-\infty}^{\infty} u(x_2) f_{2|1}(x_2|x_1)\,dx_2$$

is called the conditional expectation of $u(X_2)$, given that $X_1 = x_1$. In particular, if they do exist, then $E(X_2|x_1)$ is the mean and

$E\{[X_2 - E(X_2|x_1)]^2|x_1\}$ is the variance of the conditional distribution of X_2, given $X_1 = x_1$, which can be written more simply as var $(X_2|x_1)$. It is convenient to refer to these as the "conditional mean" and the "conditional variance" of X_2, given $X_1 = x_1$. Of course, we have

$$\text{var}(X_2|x_1) = E(X_2^2|x_1) - [E(X_2|x_1)]^2$$

from an earlier result. In like manner, the conditional expectation of $u(X_1)$, given $X_2 = x_2$, is given by

$$E[u(X_1)|x_2] = \int_{-\infty}^{\infty} u(x_1)f_{1|2}(x_1|x_2)\,dx_1.$$

With random variables of the discrete type, these conditional probabilities and conditional expectations are computed by using summation instead of integration. An illustrative example follows.

Example 1. Let X_1 and X_2 have the joint p.d.f.

$$\begin{aligned} f(x_1, x_2) &= 2, \qquad 0 < x_1 < x_2 < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Then the marginal probability density functions are, respectively,

$$\begin{aligned} f_1(x_1) &= \int_{x_1}^{1} 2\,dx_2 = 2(1 - x_1), \qquad 0 < x_1 < 1, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

and

$$\begin{aligned} f_2(x_2) &= \int_{0}^{x_2} 2\,dx_1 = 2x_2, \qquad 0 < x_2 < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

The conditional p.d.f. of X_1, given $X_2 = x_2$, $0 < x_2 < 1$, is

$$\begin{aligned} f_{1|2}(x_1|x_2) &= \frac{2}{2x_2} = \frac{1}{x_2}, \qquad 0 < x_1 < x_2, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Here the conditional mean and conditional variance of X_1, given $X_2 = x_2$, are, respectively,

$$\begin{aligned} E(X_1|x_2) &= \int_{-\infty}^{\infty} x_1 f_{1|2}(x_1|x_2)\,dx_1 \\ &= \int_{0}^{x_2} x_1\left(\frac{1}{x_2}\right) dx_1 \\ &= \frac{x_2}{2}, \qquad 0 < x_2 < 1, \end{aligned}$$

and

$$\operatorname{var}(X_1|x_2) = \int_0^{x_2} \left(x_1 - \frac{x_2}{2}\right)^2 \left(\frac{1}{x_2}\right) dx_1$$

$$= \frac{x_2^2}{12}, \qquad 0 < x_2 < 1.$$

Finally, we shall compare the values of

$$\Pr(0 < X_1 < \tfrac{1}{2}|X_2 = \tfrac{3}{4}) \qquad \text{and} \qquad \Pr(0 < X_1 < \tfrac{1}{2}).$$

We have

$$\Pr(0 < X_1 < \tfrac{1}{2}|X_2 = \tfrac{3}{4}) = \int_0^{1/2} f_{1|2}(x_1|\tfrac{3}{4})\, dx_1 = \int_0^{1/2} (\tfrac{4}{3})\, dx_1 = \tfrac{2}{3},$$

but

$$\Pr(0 < X_1 < \tfrac{1}{2}) = \int_0^{1/2} f_1(x_1)\, dx_1 = \int_0^{1/2} 2(1 - x_1)\, dx_1 = \tfrac{3}{4}.$$

Since $E(X_2|x_1)$ is a function of x_1, then $E(X_2|X_1)$ is a random variable with its own distribution, mean, and variance. Let us consider the following illustration of this.

Example 2. Let X_1 and X_2 have the joint p.d.f.

$$f(x_1, x_2) = 6x_2, \qquad 0 < x_2 < x_1 < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

Then the marginal p.d.f. of X_1 is

$$f_1(x_1) = \int_0^{x_1} 6x_2\, dx_2 = 3x_1^2, \qquad 0 < x_1 < 1,$$

zero elsewhere. The conditional p.d.f. of X_2, given $X_1 = x_1$, is

$$f_{2|1}(x_2|x_1) = \frac{6x_2}{3x_1^2} = \frac{2x_2}{x_1^2}, \qquad 0 < x_2 < x_1,$$

zero elsewhere, where $0 < x_1 < 1$. The conditional mean of X_2, given $X_1 = x_1$, is

$$E(X_2|x_1) = \int_0^{x_1} x_2\left(\frac{2x_2}{x_1^2}\right) dx_2 = \frac{2}{3}x_1, \qquad 0 < x_1 < 1.$$

Now $E(X_2|X_1) = 2X_1/3$ is a random variable, say Y. The distribution function of $Y = 2X_1/3$ is

$$G(y) = \Pr(Y \le y) = \Pr\left(X_1 \le \frac{3y}{2}\right), \qquad 0 \le y < \frac{2}{3}.$$

From the p.d.f. $f_1(x_1)$, we have

$$G(y) = \int_0^{3y/2} 3x_1^2\, dx_1 = \frac{27y^3}{8}, \qquad 0 \le y < \frac{2}{3}.$$

Of course, $G(y) = 0$, if $y < 0$, and $G(y) = 1$, if $\frac{2}{3} < y$. The p.d.f., mean, and variance of $Y = 2X_1/3$ are

$$g(y) = \frac{81y^2}{8}, \qquad 0 \le y < \frac{2}{3},$$

zero elsewhere,

$$E(Y) = \int_0^{2/3} y\left(\frac{81y^2}{8}\right) dy = \frac{1}{2},$$

and

$$\text{var}(Y) = \int_0^{2/3} y^2\left(\frac{81y^2}{8}\right) dy - \frac{1}{4} = \frac{1}{60}.$$

Since the marginal p.d.f. of X_2 is

$$f_2(x_2) = \int_{x_2}^1 6x_2\, dx_1 = 6x_2(1 - x_2), \qquad 0 < x_2 < 1,$$

zero elsewhere, it is easy to show that $E(X_2) = \frac{1}{2}$ and $\text{var}(X_2) = \frac{1}{20}$. That is, here

$$E(Y) = E[E(X_2|X_1)] = E(X_2)$$

and

$$\text{var}(Y) = \text{var}[E(X_2|X_1)] \le \text{var}(X_2).$$

Example 2 is excellent, as it provides us with the opportunity to apply many of these new definitions as well as review the distribution function technique for finding the distribution of a function of a random variable, namely $Y = 2X_1/3$. Moreover, the two observations at the end of Example 2 are no accident because it is true, in general, that

$$E[E(X_2|X_1)] = E(X_2) \quad \text{and} \quad \text{var}[E(X_2|X_1)] \le \text{var}(X_2).$$

To prove these two facts, we must first comment on the expectation of a function of two random variables, say $u(X_1, X_2)$. We do this for the continuous case, but the argument holds in the discrete case with summations replacing integrals. Of course, $Y = u(X_1, X_2)$ is a random variable and has a p.d.f., say $g(y)$, and

$$E(Y) = \int_{-\infty}^{\infty} y g(y)\, dy.$$

However, as before, it can be proved (Section 4.7) that $E(Y)$ equals

$$E[u(X_1, X_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1, x_2) f(x_1, x_2)\, dx_1\, dx_2.$$

We call $E[u(X_1, X_2)]$ the expectation (mathematical expectation or expected value) of $u(X_1, X_2)$, and it can be shown to be a linear operator as in the one-variable case. We also note that the expected value of X_2 can be found in two ways:

$$E(X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f(x_1, x_2)\, dx_1\, dx_2 = \int_{-\infty}^{\infty} x_2 f_2(x_2)\, dx_2,$$

the latter single integral being obtained from the double integral by integrating on x_1 first.

Example 3. Let X_1 and X_2 have the p.d.f.

$$f(x_1, x_2) = 8x_1x_2, \quad 0 < x_1 < x_2 < 1,$$
$$= 0 \quad \text{elsewhere.}$$

Then

$$E(X_1X_2^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2^2 f(x_1, x_2)\, dx_1\, dx_2$$
$$= \int_0^1\int_0^{x_2} 8x_1^2x_2^3\, dx_1\, dx_2$$
$$= \int_0^1 \tfrac{8}{3}x_2^6\, dx_2 = \tfrac{8}{21}.$$

In addition,

$$E(X_2) = \int_0^1\int_0^{x_2} x_2(8x_1x_2)\, dx_1\, dx_2 = \tfrac{4}{5}.$$

Since X_2 has the p.d.f. $f_2(x_2) = 4x_2^3$, $0 < x_2 < 1$, zero elsewhere, the latter expectation can be found by

$$E(X_2) = \int_0^1 x_2(4x_2^3)\, dx_2 = \tfrac{4}{5}.$$

Finally,

$$E(7X_1X_2^2 + 5X_2) = 7E(X_1X_2^2) + 5E(X_2)$$
$$= (7)(\tfrac{8}{21}) + (5)(\tfrac{4}{5}) = \tfrac{20}{3}.$$

We begin the proof of $E[E(X_2|X_1)] = E(X_2)$ and var $[E(X_2|X_1)] \le$ var (X_2) by noting that

$$E(X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f(x_1, x_2)\, dx_2\, dx_1$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x_2 \frac{f(x_1, x_2)}{f_1(x_1)} \, dx_2 \right] f_1(x_1) \, dx_1$$

$$= \int_{-\infty}^{\infty} E(X_2|x_1) f_1(x_1) \, dx_1$$

$$= E[E(X_2|X_1)],$$

which is the first result. Consider next, with $\mu_2 = E(X_2)$,

$$\begin{aligned} \text{var}(X_2) &= E[(X_2 - \mu_2)^2] \\ &= E\{[X_2 - E(X_2|X_1) + E(X_2|X_1) - \mu_2]^2\} \\ &= E\{[X_2 - E(X_2|X_1)]^2\} + E\{[E(X_2|X_1) - \mu_2]^2\} \\ &\quad + 2E\{[X_2 - E(X_2|X_1)][E(X_2|X_1) - \mu_2]\}. \end{aligned}$$

We shall show that the last term of the right-hand member of the immediately preceding equation is zero. It is equal to

$$2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x_2 - E(X_2|x_1)][E(X_2|x_1) - \mu_2] f(x_1, x_2) \, dx_2 \, dx_1$$

$$= 2 \int_{-\infty}^{\infty} [E(X_2|x_1) - \mu_2]$$

$$\times \left\{ \int_{-\infty}^{\infty} [x_2 - E(X_2|x_1)] \frac{f(x_1, x_2)}{f_1(x_1)} \, dx_2 \right\} f_1(x_1) \, dx_1.$$

But $E(X_2|x_1)$ is the conditional mean of X_2, given $X_1 = x_1$. Since the expression in the inner braces is equal to

$$E(X_2|x_1) - E(X_2|x_1) = 0,$$

the double integral is equal to zero. Accordingly, we have

$$\text{var}(X_2) = E\{[X_2 - E(X_2|X_1)]^2\} + E\{[E(X_2|X_1) - \mu_2]^2\}.$$

The first term in the right-hand member of this equation is nonnegative because it is the expected value of a nonnegative function, namely $[X_2 - E(X_2|X_1)]^2$. Since $E[E(X_2|X_1)] = \mu_2$, the second term will be the var $[E(X_2|X_1)]$. Hence we have

$$\text{var}(X_2) \geq \text{var}[E(X_2|X_1)],$$

which completes the proof.

Intuitively, this result could have this useful interpretation. Both the random variables X_2 and $E(X_2|X_1)$ have the same mean μ_2. If we did not know μ_2, we could use either of the two random variables to guess at the unknown μ_2. Since, however, var $(X_2) \geq$ var $[E(X_2|X_1)]$ we would put more reliance in $E(X_2|X_1)$ as a guess. That is, if we observe the pair (X_1, X_2) to be (x_1, x_2), we would prefer to use $E(X_2|x_1)$ to x_2 as a guess at the unknown μ_2. When studying the use of sufficient statistics in estimation in Chapter 7, we make use of this famous result, attributed to C. R. Rao and David Blackwell.

EXERCISES

2.11. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = x_1 + x_2$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere. Find the conditional mean and variance of X_2, given $X_1 = x_1$, $0 < x_1 < 1$.

2.12. Let $f_{1|2}(x_1|x_2) = c_1 x_1/x_2^2$, $0 < x_1 < x_2$, $0 < x_2 < 1$, zero elsewhere, and $f_2(x_2) = c_2 x_2^4$, $0 < x_2 < 1$, zero elsewhere, denote, respectively, the conditional p.d.f. of X_1, given $X_2 = x_2$, and the marginal p.d.f. of X_2. Determine:
(a) The constants c_1 and c_2.
(b) The joint p.d.f. of X_1 and X_2.
(c) $\Pr(\frac{1}{4} < X_1 < \frac{1}{2}|X_2 = \frac{5}{8})$.
(d) $\Pr(\frac{1}{4} < X_1 < \frac{1}{2})$.

2.13. Let $f(x_1, x_2) = 21x_1^2x_2^3$, $0 < x_1 < x_2 < 1$, zero elsewhere, be the joint p.d.f. of X_1 and X_2.
(a) Find the conditional mean and variance of X_1, given $X_2 = x_2$, $0 < x_2 < 1$.
(b) Find the distribution of $Y = E(X_1|X_2)$.
(c) Determine $E(Y)$ and var (Y) and compare these to $E(X_1)$ and var (X_1), respectively.

2.14. If X_1 and X_2 are random variables of the discrete type having p.d.f. $f(x_1, x_2) = (x_1 + 2x_2)/18$, $(x_1, x_2) = (1, 1), (1, 2), (2, 1), (2, 2)$, zero elsewhere, determine the conditional mean and variance of X_2, given $X_1 = x_1$, for $x_1 = 1$ or 2. Also compute $E(3X_1 - 2X_2)$.

2.15. Five cards are drawn at random and without replacement from a bridge deck. Let the random variables X_1, X_2, and X_3 denote, respectively, the number of spades, the number of hearts, and the number of diamonds that appear among the five cards.
(a) Determine the joint p.d.f. of X_1, X_2, and X_3.
(b) Find the marginal probability density functions of X_1, X_2, and X_3.
(c) What is the joint conditional p.d.f. of X_2 and X_3, given that $X_1 = 3$?

2.16. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$ described as follows:

(x_1, x_2)	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(2, 0)	(2, 1)
$f(x_1, x_2)$	$\frac{1}{18}$	$\frac{3}{18}$	$\frac{4}{18}$	$\frac{3}{18}$	$\frac{6}{18}$	$\frac{1}{18}$

and $f(x_1, x_2)$ is equal to zero elsewhere. Find the two marginal probability density functions and the two conditional means.

Hint: Write the probabilities in a rectangular array.

2.17. Let us choose at random a point from the interval (0, 1) and let the random variable X_1 be equal to the number which corresponds to that point. Then choose a point at random from the interval $(0, x_1)$, where x_1 is the experimental value of X_1; and let the random variable X_2 be equal to the number which corresponds to this point.

(a) Make assumptions about the marginal p.d.f. $f_1(x_1)$, and the conditional p.d.f. $f_{2|1}(x_2|x_1)$.

(b) Compute $\Pr(X_1 + X_2 \geq 1)$.

(c) Find the conditional mean $E(X_1|x_2)$.

2.18. Let $f(x)$ and $F(x)$ denote, respectively, the p.d.f. and the distribution function of the random variable X. The conditional p.d.f. of X, given $X > x_0$, x_0 a fixed number, is defined by $f(x|X > x_0) = f(x)/[1 - F(x_0)]$, $x_0 < x$, zero elsewhere. This kind of conditional p.d.f. finds application in a problem of time until death, given survival until time x_0.

(a) Show that $f(x|X > x_0)$ is a p.d.f.

(b) Let $f(x) = e^{-x}$, $0 < x < \infty$, and zero elsewhere. Compute $\Pr(X > 2|X > 1)$.

2.19. Let X and Y have the joint p.d.f. $f(x, y) = 6(1 - x - y)$, $0 < x$, $0 < y$, $x + y < 1$, and zero elsewhere. Compute $\Pr(2X + 3Y < 1)$ and $E(XY + 2X^2)$.

2.3 The Correlation Coefficient

Because the result that we obtain in this section is more familiar in terms of X and Y, we use X and Y rather than X_1 and X_2 as symbols for our two random variables. Let X and Y have joint p.d.f. $f(x, y)$. If $u(x, y)$ is a function of x and y, then $E[u(X, Y)]$ was defined, subject to its existence, in Section 2.2. The existence of all mathematical expectations will be assumed in this discussion. The means of X and Y, say μ_1 and μ_2, are obtained by taking $u(x, y)$ to be x and y, respectively; and the variances of X and Y, say σ_1^2 and σ_2^2, are obtained by setting the function $u(x, y)$ equal to $(x - \mu_1)^2$ and $(y - \mu_2)^2$, respectively. Consider the mathematical expectation

$$\begin{aligned} E[(X - \mu_1)(Y - \mu_2)] &= E(XY - \mu_2 X - \mu_1 Y + \mu_1\mu_2) \\ &= E(XY) - \mu_2 E(X) - \mu_1 E(Y) + \mu_1\mu_2 \\ &= E(XY) - \mu_1\mu_2. \end{aligned}$$

This number is called the *covariance* of X and Y and is often denoted by $\text{cov}(X, Y)$. If each of σ_1 and σ_2 is positive, the number

$$\rho = \frac{E[(X - \mu_1)(Y - \mu_2)]}{\sigma_1\sigma_2} = \frac{\text{cov}(X, Y)}{\sigma_1\sigma_2}$$

is called the *correlation coefficient* of X and Y. If the standard deviations are positive, the correlation coefficient of any two random variables is defined to be the covariance of the two random variables divided by the product of the standard deviations of the two random variables. It should be noted that the expected value of the product of two random variables is equal to the product of their expectations plus their covariance; that is, $E(XY) = \mu_1\mu_2 + \rho\sigma_1\sigma_2 = \mu_1\mu_2 + \text{cov}(X, Y)$.

Example 1. Let the random variables X and Y have the joint p.d.f.

$$\begin{aligned} f(x, y) &= x + y, \quad 0 < x < 1, \quad 0 < y < 1, \\ &= 0 \quad \text{elsewhere}. \end{aligned}$$

We shall compute the correlation coefficient of X and Y. When only two variables are under consideration, we shall denote the correlation coefficient by ρ. Now

$$\mu_1 = E(X) = \int_0^1 \int_0^1 x(x + y)\, dx\, dy = \frac{7}{12}$$

and

$$\sigma_1^2 = E(X^2) - \mu_1^2 = \int_0^1 \int_0^1 x^2(x + y)\, dx\, dy - \left(\frac{7}{12}\right)^2 = \frac{11}{144}.$$

Similarly,

$$\mu_2 = E(Y) = \frac{7}{12} \quad \text{and} \quad \sigma_2^2 = E(Y^2) - \mu_2^2 = \frac{11}{144}.$$

The covariance of X and Y is

$$E(XY) - \mu_1\mu_2 = \int_0^1 \int_0^1 xy(x + y)\, dx\, dy - \left(\frac{7}{12}\right)^2 = \frac{-1}{144}.$$

Accordingly, the correlation coefficient of X and Y is

$$\rho = \frac{-\frac{1}{144}}{\sqrt{(\frac{11}{144})(\frac{11}{144})}} = -\frac{1}{11}.$$

Remark. For certain kinds of distributions of two random variables, say X and Y, the correlation coefficient ρ proves to be a very useful characteristic of the distribution. Unfortunately, the formal definition of ρ does not reveal this fact. At this time we make some observations about ρ, some of which will be explored more fully at a later stage. It will soon be seen that if a joint distribution of two variables has a correlation coefficient (that is, if both of the variances are positive), then ρ satisfies $-1 \le \rho \le 1$. If $\rho = 1$, there is a line with equation $y = a + bx$, $b > 0$, the graph of which contains all of the probability of the distribution of X and Y. In this extreme case, we have $\Pr(Y = a + bX) = 1$. If $\rho = -1$, we have the same state of affairs except that $b < 0$. This suggests the following interesting question: When ρ does not have one of its extreme values, is there a line in the xy-plane such that the probability for X and Y tends to be concentrated in a band about this line? Under certain restrictive conditions this is in fact the case, and under those conditions we can look upon ρ as a measure of the intensity of the concentration of the probability for X and Y about that line.

Next, let $f(x, y)$ denote the joint p.d.f. of two random variables X and Y and let $f_1(x)$ denote the marginal p.d.f. of X. The conditional p.d.f. of Y, given $X = x$, is

$$f_{2|1}(y|x) = \frac{f(x, y)}{f_1(x)}$$

at points where $f_1(x) > 0$. Then the conditional mean of Y, given $X = x$, is given by

$$E(Y|x) = \int_{-\infty}^{\infty} y f_{2|1}(y|x)\, dy = \frac{\int_{-\infty}^{\infty} y f(x, y)\, dy}{f_1(x)},$$

when dealing with random variables of the continuous type. This conditional mean of Y, given $X = x$, is, of course, a function of x alone, say $u(x)$. In like vein, the conditional mean of X, given $Y = y$, is a function of y alone, say $v(y)$.

In case $u(x)$ is a linear function of x, say $u(x) = a + bx$, we say the conditional mean of Y is linear in x; or that Y has a linear conditional mean. When $u(x) = a + bx$, the constants a and b have simple values which will now be determined.

It will be assumed that neither σ_1^2 nor σ_2^2, the variances of X and Y, is zero. From

$$E(Y|x) = \frac{\int_{-\infty}^{\infty} yf(x, y)\, dy}{f_1(x)} = a + bx,$$

we have

$$\int_{-\infty}^{\infty} yf(x, y)\, dy = (a + bx)f_1(x). \tag{1}$$

If both members of Equation (1) are integrated on x, it is seen that

$$E(Y) = a + bE(X),$$

or

$$\mu_2 = a + b\mu_1, \tag{2}$$

where $\mu_1 = E(X)$ and $\mu_2 = E(Y)$. If both members of Equation (1) are first multiplied by x and then integrated on x, we have

$$E(XY) = aE(X) + bE(X^2),$$

or

$$\rho\sigma_1\sigma_2 + \mu_1\mu_2 = a\mu_1 + b(\sigma_1^2 + \mu_1^2), \tag{3}$$

where $\rho\sigma_1\sigma_2$ is the covariance of X and Y. The simultaneous solution of Equations (2) and (3) yields

$$a = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1 \qquad \text{and} \qquad b = \rho\frac{\sigma_2}{\sigma_1}.$$

That is,

$$u(x) = E(Y|x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

is the conditional mean of Y, given $X = x$, when the conditional mean of Y is linear in x. If the conditional mean of X, given $Y = y$, is linear in y, then that conditional mean is given by

$$v(y) = E(X|y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2).$$

We shall next investigate the variance of a conditional distribution under the assumption that the conditional mean is linear. The conditional variance of Y is given by

$$\operatorname{var}(Y|x) = \int_{-\infty}^{\infty} \left[y - \mu_2 - \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \right]^2 f_{2|1}(y|x)\, dy$$

$$= \frac{\displaystyle\int_{-\infty}^{\infty} \left[(y - \mu_2) - \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \right]^2 f(x, y)\, dy}{f_1(x)} \tag{4}$$

when the random variables are of the continuous type. This variance is nonnegative and is at most a function of x alone. If then, it is multiplied by $f_1(x)$ and integrated on x, the result obtained will be nonnegative. This result is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(y - \mu_2) - \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \right]^2 f(x, y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(y - \mu_2)^2 - 2\rho \frac{\sigma_2}{\sigma_1}(y - \mu_2)(x - \mu_1) + \rho^2 \frac{\sigma_2^2}{\sigma_1^2}(x - \mu_1)^2 \right] f(x, y)\, dy\, dx$$

$$= E[(Y - \mu_2)^2] - 2\rho \frac{\sigma_2}{\sigma_1} E[(X - \mu_1)(Y - \mu_2)] + \rho^2 \frac{\sigma_2^2}{\sigma_1^2} E[(X - \mu_1)^2]$$

$$= \sigma_2^2 - 2\rho \frac{\sigma_2}{\sigma_1} \rho\sigma_1\sigma_2 + \rho^2 \frac{\sigma_2^2}{\sigma_1^2}\sigma_1^2$$

$$= \sigma_2^2 - 2\rho^2\sigma_2^2 + \rho^2\sigma_2^2 = \sigma_2^2(1 - \rho^2) \ge 0.$$

That is, if the variance, Equation (4), is denoted by $k(x)$, then $E[k(X)] = \sigma_2^2(1 - \rho^2) \ge 0$. Accordingly, $\rho^2 \le 1$, or $-1 \le \rho \le 1$. It is left as an exercise to prove that $-1 \le \rho \le 1$ whether the conditional mean is or is not linear.

Suppose that the variance, Equation (4), is positive but not a function of x; that is, the variance is a constant $k > 0$. Now if k is multiplied by $f_1(x)$ and integrated on x, the result is k, so that $k = \sigma_2^2(1 - \rho^2)$. Thus, in this case, the variance of each conditional distribution of Y, given $X = x$, is $\sigma_2^2(1 - \rho^2)$. If $\rho = 0$, the variance of each conditional distribution of Y, given $X = x$, is σ_2^2, the variance of the marginal distribution of Y. On the other hand, if ρ^2 is near one, the variance of each conditional distribution of Y, given $X = x$, is relatively small, and there is a high concentration of the probability for this conditional distribution near the mean $E(Y|x) = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$.

It should be pointed out that if the random variables X and Y in the preceding discussion are taken to be of the discrete type, the results just obtained are valid.

Example 2. Let the random variables X and Y have the linear conditional means $E(Y|x) = 4x + 3$ and $E(X|y) = \frac{1}{16}y - 3$. In accordance with the general formulas for the linear conditional means, we see that $E(Y|x) = \mu_2$ if $x = \mu_1$ and $E(X|y) = \mu_1$ if $y = \mu_2$. Accordingly, in this special case, we have $\mu_2 = 4\mu_1 + 3$ and $\mu_1 = \frac{1}{16}\mu_2 - 3$ so that $\mu_1 = -\frac{15}{4}$ and $\mu_2 = -12$. The general formulas for the linear conditional means also show that the product of the coefficients of x and y, respectively, is equal to ρ^2 and that the quotient of these coefficients is equal to σ_2^2/σ_1^2. Here $\rho^2 = 4(\frac{1}{16}) = \frac{1}{4}$ with $\rho = \frac{1}{2}$ (not $-\frac{1}{2}$), and $\sigma_2^2/\sigma_1^2 = 64$. Thus, from the two linear conditional means, we are able to find the values of μ_1, μ_2, ρ, and σ_2/σ_1, but not the values of σ_1 and σ_2.

Example 3. To illustrate how the correlation coefficient measures the intensity of the concentration of the probability for X and Y about a line, let these random variables have a distribution that is uniform over the area depicted in Figure 2.1. That is, the joint p.d.f. of X and Y is

$$\begin{aligned} f(x, y) &= \frac{1}{4ah}, \qquad -a + bx < y < a + bx, \quad -h < x < h, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

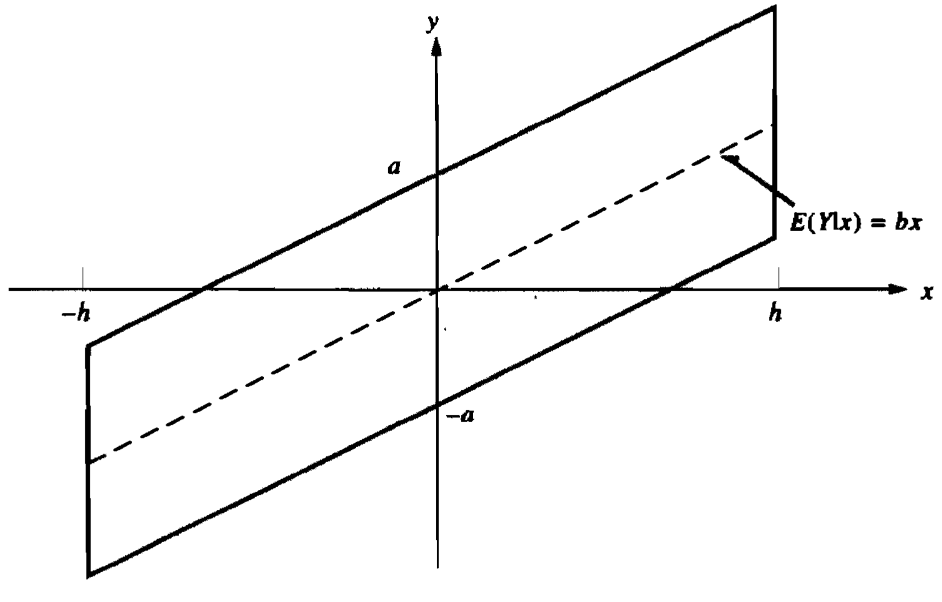

FIGURE 2.1

We assume here that $b \geq 0$, but the argument can be modified for $b \leq 0$. It is easy to show that the p.d.f. of X is uniform, namely

$$f_1(x) = \int_{-a+bx}^{a+bx} \frac{1}{4ah}\,dy = \frac{1}{2h}, \qquad -h < x < h,$$

$$= 0 \qquad \text{elsewhere.}$$

Thus the conditional p.d.f. of Y, given $X = x$, is uniform:

$$f_{2|1}(y|x) = \frac{1/4ah}{1/2h} = \frac{1}{2a}, \qquad -a + bx < y < a + bx,$$

$$= 0 \qquad \text{elsewhere.}$$

The conditional mean and variance are

$$E(Y|x) = bx \qquad \text{and} \qquad \text{var}\,(Y|x) = \frac{a^2}{3}.$$

From the general expressions for those characteristics we know that

$$b = \rho\frac{\sigma_2}{\sigma_1} \qquad \text{and} \qquad \frac{a^2}{3} = \sigma_2^2(1 - \rho^2).$$

In addition, we know that $\sigma_1^2 = h^2/3$. If we solve these three equations, we obtain an expression for the correlation coefficient, namely

$$\rho = \frac{bh}{\sqrt{a^2 + b^2h^2}}.$$

Referring to Figure 2.1, we note:

1. As a gets small (large), the straight line effect is more (less) intense and ρ is closer to 1 (zero).
2. As h gets large (small), the straight line effect is more (less) intense and ρ is closer to 1 (zero).
3. As b gets large (small), the straight line effect is more (less) intense and ρ is closer to 1 (zero).

This section will conclude with a definition and an illustrative example. Let $f(x, y)$ denote the joint p.d.f. of the two random variables X and Y. If $E(e^{t_1X + t_2Y})$ exists for $-h_1 < t_1 < h_1$, $-h_2 < t_2 < h_2$, where h_1 and h_2 are positive, it is denoted by $M(t_1, t_2)$ and is called the *moment-generating function* (m.g.f.) of the joint distribution of X and Y. As in the case of one random variable, the m.g.f. $M(t_1, t_2)$ completely determines the joint distribution of X and Y, and hence the marginal distributions of X and Y. In fact, the m.g.f. $M_1(t_1)$ of X is

$$M_1(t_1) = E(e^{t_1X}) = M(t_1, 0)$$

and the m.g.f. $M_2(t_2)$ of Y is

$$M_2(t_2) = E(e^{t_2 Y}) = M(0, t_2).$$

In addition, in the case of random variables of the continuous type,

$$\frac{\partial^{k+m} M(t_1, t_2)}{\partial t_1^k \, \partial t_2^m} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^m e^{t_1 x + t_2 y} f(x, y) \, dx \, dy,$$

so that

$$\left. \frac{\partial^{k+m} M(t_1, t_2)}{\partial t_1^k \, \partial t_2^m} \right|_{t_1 = t_2 = 0} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^m f(x, y) \, dx \, dy = E(X^k Y^m).$$

For instance, in a simplified notation which appears to be clear,

$$\mu_1 = E(X) = \frac{\partial M(0, 0)}{\partial t_1}, \qquad \mu_2 = E(Y) = \frac{\partial M(0, 0)}{\partial t_2},$$

$$\sigma_1^2 = E(X^2) - \mu_1^2 = \frac{\partial^2 M(0, 0)}{\partial t_1^2} - \mu_1^2,$$

$$\sigma_2^2 = E(Y^2) - \mu_2^2 = \frac{\partial^2 M(0, 0)}{\partial t_2^2} - \mu_2^2, \tag{5}$$

$$E[(X - \mu_1)(Y - \mu_2)] = \frac{\partial^2 M(0, 0)}{\partial t_1 \, \partial t_2} - \mu_1 \mu_2,$$

and from these we can compute the correlation coefficient ρ.

It is fairly obvious that the results of Equations (5) hold if X and Y are random variables of the discrete type. Thus the correlation coefficients may be computed by using the m.g.f. of the joint distribution if that function is readily available. An illustrative example follows. In this, we let $e^w = \exp(w)$.

Example 4. Let the continuous-type random variables X and Y have the joint p.d.f.

$$f(x, y) = e^{-y}, \qquad 0 < x < y < \infty,$$
$$= 0 \qquad \text{elsewhere.}$$

The m.g.f. of this joint distribution is

$$M(t_1, t_2) = \int_0^{\infty} \int_x^{\infty} \exp(t_1 x + t_2 y - y) \, dy \, dx$$

$$= \frac{1}{(1 - t_1 - t_2)(1 - t_2)},$$

provided that $t_1 + t_2 < 1$ and $t_2 < 1$. For this distribution, Equations (5) become

$$\begin{aligned} \mu_1 &= 1, \qquad \mu_2 = 2, \\ \sigma_1^2 &= 1, \qquad \sigma_2^2 = 2, \\ E[(X-\mu_1)(Y-\mu_2)] &= 1. \end{aligned} \tag{6}$$

Verification of results of Equations (6) is left as an exercise. If, momentarily, we accept these results, the correlation coefficient of X and Y is $\rho = 1/\sqrt{2}$. Furthermore, the moment-generating functions of the marginal distributions of X and Y are, respectively,

$$M(t_1, 0) = \frac{1}{1-t_1}, \qquad t_1 < 1,$$

$$M(0, t_2) = \frac{1}{(1-t_2)^2}, \qquad t_2 < 1.$$

These moment-generating functions are, of course, respectively, those of the marginal probability density functions,

$$f_1(x) = \int_x^\infty e^{-y}\,dy = e^{-x}, \qquad 0 < x < \infty,$$

zero elsewhere, and

$$f_2(y) = e^{-y}\int_0^y dx = ye^{-y}, \qquad 0 < y < \infty,$$

zero elsewhere.

EXERCISES

2.20. Let the random variables X and Y have the joint p.d.f.
(a) $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 0), (1, 1), (2, 2)$, zero elsewhere.
(b) $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 2), (1, 1), (2, 0)$, zero elsewhere.
(c) $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 0), (1, 1), (2, 0)$, zero elsewhere.
In each case compute the correlation coefficient of X and Y.

2.21. Let X and Y have the joint p.d.f. described as follows:

(x, y)	$(1, 1)$	$(1, 2)$	$(1, 3)$	$(2, 1)$	$(2, 2)$	$(2, 3)$
$f(x, y)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$

and $f(x, y)$ is equal to zero elsewhere. (a) Find the means μ_1 and μ_2, the variances σ_1^2 and σ_2^2, and the correlation coefficient ρ. (b) Compute $E(Y|X = 1)$, $E(Y|X = 2)$, and the line $\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$. Do the points $[k, E(Y|X = k)]$, $k = 1, 2$, lie on this line?

2.22. Let $f(x, y) = 2, 0 < x < y, 0 < y < 1$, zero elsewhere, be the joint p.d.f. of X and Y. Show that the conditional means are, respectively, $(1 + x)/2$, $0 < x < 1$, and $y/2$, $0 < y < 1$. Show that the correlation coefficient of X and Y is $\rho = \frac{1}{2}$.

2.23. Show that the variance of the conditional distribution of Y, given $X = x$, in Exercise 2.22, is $(1 - x)^2/12$, $0 < x < 1$, and that the variance of the conditional distribution of X, given $Y = y$, is $y^2/12$, $0 < y < 1$.

2.24. Verify the results of Equations (6) of this section.

2.25. Let X and Y have the joint p.d.f. $f(x, y) = 1$, $-x < y < x$, $0 < x < 1$, zero elsewhere. Show that, on the set of positive probability density, the graph of $E(Y|x)$ is a straight line, whereas that of $E(X|y)$ is not a straight line.

2.26. If the correlation coefficient ρ of X and Y exists, show that $-1 \le \rho \le 1$.
Hint: Consider the discriminant of the nonnegative quadratic function $h(v) = E\{[(X - \mu_1) + v(Y - \mu_2)]^2\}$, where v is real and is not a function of X nor of Y.

2.27. Let $\psi(t_1, t_2) = \ln M(t_1, t_2)$, where $M(t_1, t_2)$ is the m.g.f. of X and Y. Show that

$$\frac{\partial \psi(0, 0)}{\partial t_i}, \quad \frac{\partial^2 \psi(0, 0)}{\partial t_i^2}, \quad i = 1, 2,$$

and

$$\frac{\partial^2 \psi(0, 0)}{\partial t_1 \, \partial t_2}$$

yield the means, the variances, and the covariance of the two random variables. Use this result to find the means, the variances, and the covariance of X and Y of Example 4.

2.4 Independent Random Variables

Let X_1 and X_2 denote random variables of either the continuous or the discrete type which have the joint p.d.f. $f(x_1, x_2)$ and marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. In accordance with the definition of the conditional p.d.f. $f_{2|1}(x_2|x_1)$, we may write the joint p.d.f. $f(x_1, x_2)$ as

$$f(x_1, x_2) = f_{2|1}(x_2|x_1)f_1(x_1).$$

Suppose that we have an instance where $f_{2|1}(x_2|x_1)$ does not depend upon x_1. Then the marginal p.d.f. of X_2 is, for random variables of the continuous type,

$$f_2(x_2) = \int_{-\infty}^{\infty} f_{2|1}(x_2|x_1)f_1(x_1)\,dx_1$$

$$= f_{2|1}(x_2|x_1)\int_{-\infty}^{\infty} f_1(x_1)\,dx_1$$

$$= f_{2|1}(x_2|x_1).$$

Accordingly,

$$f_2(x_2) = f_{2|1}(x_2|x_1) \qquad \text{and} \qquad f(x_1, x_2) = f_1(x_1)f_2(x_2),$$

when $f_{2|1}(x_2|x_1)$ does not depend upon x_1. That is, if the conditional distribution of X_2, given $X_1 = x_1$, is independent of any assumption about x_1, then $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. These considerations motivate the following definition.

Definition 2. Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$ and the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. The random variables X_1 and X_2 are said to be independent if, and only if, $f(x_1, x_2) \equiv f_1(x_1)f_2(x_2)$. Random variables that are not independent are said to be dependent.

Remarks. Two comments should be made about the preceding definition. First, the product of two positive functions $f_1(x_1)f_2(x_2)$ means a function that is positive on a product space. That is, if $f_1(x_1)$ and $f_2(x_2)$ are positive on, and only on, the respective spaces $\mathscr{A}_1$ and $\mathscr{A}_2$, then the product of $f_1(x_1)$ and $f_2(x_2)$ is positive on, and only on, the product space $\mathscr{A} = \{(x_1, x_2) : x_1 \in \mathscr{A}_1, x_2 \in \mathscr{A}_2\}$. For instance, if $\mathscr{A}_1 = \{x_1 : 0 < x_1 < 1\}$ and $\mathscr{A}_2 = \{x_2 : 0 < x_2 < 3\}$, then $\mathscr{A} = \{(x_1, x_2) : 0 < x_1 < 1, 0 < x_2 < 3\}$. The second remark pertains to the identity. The identity in Definition 2 should be interpreted as follows. There may be certain points $(x_1, x_2) \in \mathscr{A}$ at which $f(x_1, x_2) \neq f_1(x_1)f_2(x_2)$. However, if A is the set of points (x_1, x_2) at which the equality does not hold, then $P(A) = 0$. In the subsequent theorems and the subsequent generalizations, a product of nonnegative functions and an identity should be interpreted in an analogous manner.

Example 1. Let the joint p.d.f. of X_1 and X_2 be

$$f(x_1, x_2) = x_1 + x_2, \qquad 0 < x_1 < 1, \quad 0 < x_2 < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

It will be shown that X_1 and X_2 are dependent. Here the marginal probability density functions are

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 = \int_0^1 (x_1 + x_2)\, dx_2 = x_1 + \tfrac{1}{2}, \qquad 0 < x_1 < 1,$$

$$= 0 \qquad \text{elsewhere,}$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 = \int_0^1 (x_1 + x_2)\, dx_1 = \tfrac{1}{2} + x_2, \qquad 0 < x_2 < 1,$$

$$= 0 \qquad \text{elsewhere.}$$

Since $f(x_1, x_2) \not\equiv f_1(x_1)f_2(x_2)$, the random variables X_1 and X_2 are dependent

The following theorem makes it possible to assert, without computing the marginal probability density functions, that the random variables X_1 and X_2 of Example 1 are dependent.

Theorem 1. *Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$. Then X_1 and X_2 are independent if and only if $f(x_1, x_2)$ can be written as a product of a nonnegative function of x_1 alone and a nonnegative function of x_2 alone. That is,*

$$f(x_1, x_2) \equiv g(x_1)h(x_2),$$

where $g(x_1) > 0$, $x_1 \in \mathscr{A}_1$, zero elsewhere, and $h(x_2) > 0$, $x_2 \in \mathscr{A}_2$, zero elsewhere.

Proof. If X_1 and X_2 are independent, then $f(x_1, x_2) \equiv f_1(x_1)f_2(x_2)$, where $f_1(x_1)$ and $f_2(x_2)$ are the marginal probability density functions of X_1 and X_2, respectively. Thus the condition $f(x_1, x_2) \equiv g(x_1)h(x_2)$ is fulfilled.

Conversely, if $f(x_1, x_2) \equiv g(x_1)h(x_2)$, then, for random variables of the continuous type, we have

$$f_1(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\, dx_2 = g(x_1)\int_{-\infty}^{\infty} h(x_2)\, dx_2 = c_1 g(x_1)$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\, dx_1 = h(x_2)\int_{-\infty}^{\infty} g(x_1)\, dx_1 = c_2 h(x_2),$$

where c_1 and c_2 are constants, not functions of x_1 or x_2. Moreover, $c_1c_2 = 1$ because

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1)h(x_2)\,dx_1\,dx_2 = \left[\int_{-\infty}^{\infty} g(x_1)\,dx_1\right]\left[\int_{-\infty}^{\infty} h(x_2)\,dx_2\right]$$
$$= c_2c_1.$$

These results imply that

$$f(x_1, x_2) \equiv g(x_1)h(x_2) \equiv c_1g(x_1)c_2h(x_2) \equiv f_1(x_1)f_2(x_2).$$

Accordingly, X_1 and X_2 are independent.

If we now refer to Example 1, we see that the joint p.d.f.

$$f(x_1, x_2) = x_1 + x_2, \qquad 0 < x_1 < 1, \quad 0 < x_2 < 1,$$
$$= 0 \qquad \text{elsewhere,}$$

cannot be written as the product of a nonnegative function of x_1 alone and a nonnegative function of x_2 alone. Accordingly, X_1 and X_2 are dependent.

Example 2. Let the p.d.f. of the random variables X_1 and X_2 be $f(x_1, x_2) = 8x_1x_2$, $0 < x_1 < x_2 < 1$, zero elsewhere. The formula $8x_1x_2$ might suggest to some that X_1 and X_2 are independent. However, if we consider the space $\mathscr{A} = \{(x_1, x_2) : 0 < x_1 < x_2 < 1\}$, we see that it is not a product space. This should make it clear that, in general, X_1 and X_2 must be dependent if the space of positive probability density of X_1 and X_2 is bounded by a curve that is neither a horizontal nor a vertical line.

We now give a theorem that frequently simplifies the calculations of probabilities of events which involve independent variables.

Theorem 2. *If X_1 and X_2 are independent random variables with marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively, then*

$$\Pr(a < X_1 < b, c < X_2 < d) = \Pr(a < X_1 < b)\Pr(c < X_2 < d)$$

for every $a < b$ and $c < d$, where a, b, c, and d are constants.

Proof. From the independence of X_1 and X_2, the joint p.d.f. of X_1 and X_2 is $f_1(x_1)f_2(x_2)$. Accordingly, in the continuous case,

$$\begin{aligned}\Pr(a < X_1 < b, c < X_2 < d) &= \int_a^b \int_c^d f_1(x_1)f_2(x_2)\,dx_2\,dx_1 \\ &= \left[\int_a^b f_1(x_1)\,dx_1\right]\left[\int_c^d f_2(x_2)\,dx_2\right] \\ &= \Pr(a < X_1 < b)\Pr(c < X_2 < d);\end{aligned}$$

or, in the discrete case,

$$\begin{aligned}\Pr(a < X_1 < b, c < X_2 < d) &= \sum_{a<x_1<b}\ \sum_{c<x_2<d} f_1(x_1)f_2(x_2) \\ &= \left[\sum_{a<x_1<b} f_1(x_1)\right]\left[\sum_{c<x_2<d} f_2(x_2)\right] \\ &= \Pr(a < X_1 < b)\Pr(c < X_2 < d),\end{aligned}$$

as was to be shown.

Example 3. In Example 1, X_1 and X_2 were found to be dependent. There, in general,

$$\Pr(a < X_1 < b, c < X_2 < d) \neq \Pr(a < X_1 < b)\Pr(c < X_2 < d).$$

For instance,

$$\Pr(0 < X_1 < \tfrac{1}{2}, 0 < X_2 < \tfrac{1}{2}) = \int_0^{1/2}\int_0^{1/2} (x_1 + x_2)\,dx_1\,dx_2 = \tfrac{1}{8},$$

whereas

$$\Pr(0 < X_1 < \tfrac{1}{2}) = \int_0^{1/2} (x_1 + \tfrac{1}{2})\,dx_1 = \tfrac{3}{8}$$

and

$$\Pr(0 < X_2 < \tfrac{1}{2}) = \int_0^{1/2} (\tfrac{1}{2} + x_2)\,dx_2 = \tfrac{3}{8}.$$

Not merely are calculations of some probabilities usually simpler when we have independent random variables, but many expectations, including certain moment-generating functions, have comparably simpler computations. The following result will prove so useful that we state it in the form of a theorem.

Theorem 3. *Let the independent random variables X_1 and X_2 have the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. The expected value of the product of a function $u(X_1)$ of X_1 alone and a function $v(X_2)$ of X_2 alone is, subject to their existence, equal to* *the product of the expected value of $u(X_1)$ and the expected value of $v(X_2)$; that is,*

$$E[u(X_1)v(X_2)] = E[u(X_1)]E[v(X_2)].$$

Proof. The independence of X_1 and X_2 implies that the joint p.d.f. of X_1 and X_2 is $f_1(x_1)f_2(x_2)$. Thus we have, by definition of expectation, in the continuous case,

$$\begin{aligned} E[u(X_1)v(X_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1)v(x_2)f_1(x_1)f_2(x_2)\,dx_1\,dx_2 \\ &= \left[\int_{-\infty}^{\infty} u(x_1)f_1(x_1)\,dx_1\right]\left[\int_{-\infty}^{\infty} v(x_2)f_2(x_2)\,dx_2\right] \\ &= E[u(X_1)]E[v(X_2)]; \end{aligned}$$

or, in the discrete case,

$$\begin{aligned} E[u(X_1)v(X_2)] &= \sum_{x_2}\sum_{x_1} u(x_1)v(x_2)f_1(x_1)f_2(x_2) \\ &= \left[\sum_{x_1} u(x_1)f_1(x_1)\right]\left[\sum_{x_2} v(x_2)f_2(x_2)\right] \\ &= E[u(X_1)]E[v(X_2)], \end{aligned}$$

as stated in the theorem.

Example 4. Let X and Y be two independent random variables with means μ_1 and μ_2 and positive variances σ_1^2 and σ_2^2, respectively. We shall show that the independence of X and Y implies that the correlation coefficient of X and Y is zero. This is true because the covariance of X and Y is equal to

$$E[(X - \mu_1)(Y - \mu_2)] = E(X - \mu_1)E(Y - \mu_2) = 0.$$

We shall now prove a very useful theorem about independent random variables. The proof of the theorem relies heavily upon our assertion that an m.g.f., when it exists, is unique and that it uniquely determines the distribution of probability.

Theorem 4. *Let X_1 and X_2 denote random variables that have the joint p.d.f. $f(x_1, x_2)$ and the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. Furthermore, let $M(t_1, t_2)$ denote the m.g.f. of the distribution. Then X_1 and X_2 are independent if and only if*

$$M(t_1, t_2) = M(t_1, 0)M(0, t_2).$$

Proof. If X_1 and X_2 are independent, then

$$\begin{aligned} M(t_1, t_2) &= E(e^{t_1X_1 + t_2X_2}) \\ &= E(e^{t_1X_1}e^{t_2X_2}) \\ &= E(e^{t_1X_1})E(e^{t_2X_2}) \\ &= M(t_1, 0)M(0, t_2). \end{aligned}$$

Thus the independence of X_1 and X_2 implies that the m.g.f. of the joint distribution factors into the product of the moment-generating functions of the two marginal distributions.

Suppose next that the m.g.f. of the joint distribution of X_1 and X_2 is given by $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$. Now X_1 has the unique m.g.f. which, in the continuous case, is given by

$$M(t_1, 0) = \int_{-\infty}^{\infty} e^{t_1x_1}f_1(x_1)\, dx_1.$$

Similarly, the unique m.g.f. of X_2, in the continuous case, is given by

$$M(0, t_2) = \int_{-\infty}^{\infty} e^{t_2x_2}f_2(x_2)\, dx_2.$$

Thus we have

$$\begin{aligned} M(t_1, 0)M(0, t_2) &= \left[\int_{-\infty}^{\infty} e^{t_1x_1}f_1(x_1)\, dx_1\right]\left[\int_{-\infty}^{\infty} e^{t_2x_2}f_2(x_2)\, dx_2\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x_1 + t_2x_2}f_1(x_1)f_2(x_2)\, dx_1\, dx_2. \end{aligned}$$

We are given that $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$; so

$$M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x_1 + t_2x_2}f_1(x_1)f_2(x_2)\, dx_1\, dx_2.$$

But $M(t_1, t_2)$ is the m.g.f. of X_1 and X_2. Thus also

$$M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x_1 + t_2x_2}f(x_1, x_2)\, dx_1\, dx_2.$$

The uniqueness of the m.g.f. implies that the two distributions of probability that are described by $f_1(x_1)f_2(x_2)$ and $f(x_1, x_2)$ are the same. Thus

$$f(x_1, x_2) \equiv f_1(x_1)f_2(x_2).$$

That is, if $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$, then X_1 and X_2 are independent. This completes the proof when the random variables are of the continuous type. With random variables of the discrete type, the proof is made by using summation instead of integration.

EXERCISES

2.28. Show that the random variables X_1 and X_2 with joint p.d.f. $f(x_1, x_2) = 12x_1x_2(1 - x_2)$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere, are independent.

2.29. If the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = 2e^{-x_1 - x_2}$, $0 < x_1 < x_2$, $0 < x_2 < \infty$, zero elsewhere, show that X_1 and X_2 are dependent.

2.30. Let $f(x_1, x_2) = \frac{1}{16}$, $x_1 = 1, 2, 3, 4$, and $x_2 = 1, 2, 3, 4$, zero elsewhere, be the joint p.d.f. of X_1 and X_2. Show that X_1 and X_2 are independent.

2.31. Find $\Pr(0 < X_1 < \frac{1}{3}, 0 < X_2 < \frac{1}{3})$ if the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = 4x_1(1 - x_2)$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere.

2.32. Find the probability of the union of the events $a < X_1 < b$, $-\infty < X_2 < \infty$ and $-\infty < X_1 < \infty$, $c < X_2 < d$ if X_1 and X_2 are two independent variables with $\Pr(a < X_1 < b) = \frac{2}{3}$ and $\Pr(c < X_2 < d) = \frac{5}{8}$.

2.33. If $f(x_1, x_2) = e^{-x_1 - x_2}$, $0 < x_1 < \infty$, $0 < x_2 < \infty$, zero elsewhere, is the joint p.d.f. of the random variables X_1 and X_2, show that X_1 and X_2 are independent and that $M(t_1, t_2) = (1 - t_1)^{-1}(1 - t_2)^{-1}$, $t_2 < 1$, $t_1 < 1$. Also show that

$$E(e^{t(X_1 + X_2)}) = (1 - t)^{-2}, \quad t < 1.$$

Accordingly, find the mean and the variance of $Y = X_1 + X_2$.

2.34. Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = 1/\pi$, $(x_1 - 1)^2 + (x_2 + 2)^2 < 1$, zero elsewhere. Find $f_1(x_1)$ and $f_2(x_2)$. Are X_1 and X_2 independent?

2.35. Let X and Y have the joint p.d.f. $f(x, y) = 3x$, $0 < y < x < 1$, zero elsewhere. Are X and Y independent? If not, find $E(X|y)$.

2.36. Suppose that a man leaves for work between 8:00 A.M. and 8:30 A.M. and takes between 40 and 50 minutes to get to the office. Let X denote the time of departure and let Y denote the time of travel. If we assume that these random variables are independent and uniformly distributed, find the probability that he arrives at the office before 9:00 A.M.

2.5 Extension to Several Random Variables

The notions about two random variables can be extended immediately to n random variables. We make the following definition of the space of n random variables.

Definition 3. Consider a random experiment with the sample space $\mathscr{C}$. Let the random variable X_i assign to each element $c \in \mathscr{C}$ one and only one real number $X_i(c) = x_i$, $i = 1, 2, \ldots, n$. The *space* of these random variables is the set of ordered n-tuples $\mathscr{A} = \{(x_1, x_2, \ldots, x_n) : x_1 = X_1(c), \ldots, x_n = X_n(c), c \in \mathscr{C}\}$. Furthermore, let A be a subset of $\mathscr{A}$. Then $\Pr[(X_1, \ldots, X_n) \in A] = P(C)$, where $C = \{c : c \in \mathscr{C} \text{ and } [X_1(c), X_2(c), \ldots, X_n(c)] \in A\}$.

Again we should make the comment that $\Pr[(X_1, \ldots, X_n) \in A]$ could be denoted by the probability set function $P_{X_1,\ldots,X_n}(A)$. But, if there is no chance of misunderstanding, it will be written simply as $P(A)$. We say that the n random variables $X_1, X_2, \ldots, X_n$ are of the discrete type or of the continuous type, and have a distribution of that type, according as the probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed as

$$P(A) = \Pr[(X_1, \ldots, X_n) \in A] = \sum \cdot \underset{A}{\cdot} \cdot \sum f(x_1, \ldots, x_n),$$

or as

$$P(A) = \Pr[(X_1, \ldots, X_n) \in A] = \int \cdots \underset{A}{\int} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

In accordance with the convention of extending the definition of a p.d.f., it is seen that a point function f essentially satisfies the conditions of being a p.d.f. if (a) f is defined and is nonnegative for all real values of its argument(s) and if (b) its integral [for the continuous type of random variable(s)], or its sum [for the discrete type of random variable(s)] over all real values of its argument(s) is 1.

The distribution function of the n random variables $X_1, X_2, \ldots, X_n$ is the point function

$$F(x_1, x_2, \ldots, x_n) = \Pr(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n).$$

An illustrative example follows.

Example 1. Let $f(x, y, z) = e^{-(x+y+z)}$, $0 < x, y, z < \infty$, zero elsewhere, be the p.d.f. of the random variables X, Y, and Z. Then the distribution function of X, Y, and Z is given by

$$\begin{aligned} F(x, y, z) &= \Pr(X \le x, Y \le y, Z \le z) \\ &= \int_0^z \int_0^y \int_0^x e^{-u-v-w}\, du\, dv\, dw \\ &= (1 - e^{-x})(1 - e^{-y})(1 - e^{-z}), \qquad 0 \le x, y, z < \infty, \end{aligned}$$

and is equal to zero elsewhere. Incidentally, except for a set of probability measure zero, we have

$$\frac{\partial^3 F(x, y, z)}{\partial x\, \partial y\, \partial z} = f(x, y, z).$$

Let $X_1, X_2, \ldots, X_n$ be random variables having joint p.d.f. $f(x_1, x_2, \ldots, x_n)$ and let $u(X_1, X_2, \ldots, X_n)$ be a function of these variables such that the n-fold integral

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \quad (1)$$

exists, if the random variables are of the continuous type, or such that the n-fold sum

$$\sum_{x_n} \cdots \sum_{x_1} u(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) \quad (2)$$

exists if the random variables are of the discrete type. The n-fold integral (or the n-fold sum, as the case may be) is called the *expectation*, denoted by $E[u(X_1, X_2, \ldots, X_n)]$, of the function $u(X_1, X_2, \ldots, X_n)$. In Section 4.7 we show this expectation to be equal to $E(Y)$, where $Y = u(X_1, X_2, \ldots, X_n)$. Of course, E is a linear operator.

We shall now discuss the notions of marginal and conditional probability density functions from the point of view of n random variables. All of the preceding definitions can be directly generalized to the case of n variables in the following manner. Let the random variables $X_1, X_2, \ldots, X_n$ have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$. If the random variables are of the continuous type, then by an argument similar to the two-variable case, we have for every $a < b$,

$$\Pr(a < X_1 < b) = \int_a^b f_1(x_1)\, dx_1,$$

where $f_1(x_1)$ is defined by the $(n-1)$-fold integral

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_2 \cdots dx_n.$$

Therefore, $f_1(x_1)$ is the p.d.f. of the one random variable X_1 and $f_1(x_1)$ is called the marginal p.d.f. of X_1. The marginal probability density functions $f_2(x_2), \ldots, f_n(x_n)$ of $X_2, \ldots, X_n$, respectively, are similar $(n-1)$-fold integrals.

Up to this point, each marginal p.d.f. has been a p.d.f. of one random variable. It is convenient to extend this terminology to joint probability density functions, which we shall do now. Here let $f(x_1, x_2, \ldots, x_n)$ be the joint p.d.f. of the n random variables $X_1, X_2, \ldots, X_n$, just as before. Now, however, let us take any group of $k < n$ of these random variables and let us find the joint p.d.f. of them. This joint p.d.f. is called the marginal p.d.f. of this particular group of k variables. To fix the ideas, take $n = 6$, $k = 3$, and let us select the group X_2, X_4, X_5. Then the marginal p.d.f. of X_2, X_4, X_5 is the joint p.d.f. of this particular group of three variables, namely,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4, x_5, x_6)\, dx_1\, dx_3\, dx_6,$$

if the random variables are of the continuous type.

Next we extend the definition of a conditional p.d.f. If $f_1(x_1) > 0$, the symbol $f_{2,\ldots,n|1}(x_2, \ldots, x_n|x_1)$ is defined by the relation

$$f_{2,\ldots,n|1}(x_2, \ldots, x_n|x_1) = \frac{f(x_1, x_2, \ldots, x_n)}{f_1(x_1)},$$

and $f_{2,\ldots,n|1}(x_2, \ldots, x_n|x_1)$ is called the *joint conditional p.d.f.* of $X_2, \ldots, X_n$, given $X_1 = x_1$. The joint conditional p.d.f. of any $n - 1$ random variables, say $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, given $X_i = x_i$, is defined as the joint p.d.f. of $X_1, X_2, \ldots, X_n$ divided by the marginal p.d.f. $f_i(x_i)$, provided that $f_i(x_i) > 0$. More generally, the joint conditional p.d.f. of $n - k$ of the random variables, for given values of the remaining k variables, is defined as the joint p.d.f. of the n variables divided by the marginal p.d.f. of the particular group of k variables, provided that the latter p.d.f. is positive. We remark that there are many other conditional probability density functions; for instance, see Exercise 2.18.

Because a conditional p.d.f. is a p.d.f. of a certain number of random variables, the expectation of a function of these random variables has been defined. To emphasize the fact that a conditional p.d.f. is under consideration, such expectations are called conditional expectations. For instance, the conditional expectation of $u(X_2, \ldots, X_n)$ given $X_1 = x_1$, is, for random variables of the continuous type, given by

$$E[u(X_2, \ldots, X_n)|x_1] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_2, \ldots, x_n)$$

$$\times f_{2,\ldots,n|1}(x_2, \ldots, x_n|x_1)\, dx_2 \cdots dx_n,$$

provided $f_1(x_1) > 0$ and the integral converges (absolutely). If the random variables are of the discrete type, conditional expectations are, of course, computed by using sums instead of integrals.

Let the random variables $X_1, X_2, \ldots, X_n$ have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$ and the marginal probability density functions $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, respectively. The definition of the independence of X_1 and X_2 is generalized to the mutual independence of $X_1, X_2, \ldots, X_n$ as follows: The random variables $X_1, X_2, \ldots, X_n$ are said to be *mutually independent* if and only if

$$f(x_1, x_2, \ldots, x_n) \equiv f_1(x_1)f_2(x_2)\cdots f_n(x_n).$$

It follows immediately from this definition of the mutual independence of $X_1, X_2, \ldots, X_n$ that

$$\begin{aligned}\Pr(a_1 < X_1 < b_1, a_2 < X_2 < b_2, \ldots, a_n < X_n < b_n)\\ &= \Pr(a_1 < X_1 < b_1)\Pr(a_2 < X_2 < b_2)\cdots\Pr(a_n < X_n < b_n)\\ &= \prod_{i=1}^{n} \Pr(a_i < X_i < b_i),\end{aligned}$$

where the symbol $\prod_{i=1}^{n} \varphi(i)$ is defined to be

$$\prod_{i=1}^{n} \varphi(i) = \varphi(1)\varphi(2)\cdots\varphi(n).$$

The theorem that

$$E[u(X_1)v(X_2)] = E[u(X_1)]E[v(X_2)]$$

for independent random variables X_1 and X_2 becomes, for mutually independent random variables $X_1, X_2, \ldots, X_n$,

$$E[u_1(X_1)u_2(X_2)\cdots u_n(X_n)] = E[u_1(X_1)]E[u_2(X_2)]\cdots E[u_n(X_n)],$$

or

$$E\left[\prod_{i=1}^{n} u_i(X_i)\right] = \prod_{i=1}^{n} E[u_i(X_i)].$$

The moment-generating function of the joint distribution of n random variables $X_1, X_2, \ldots, X_n$ is defined as follows. Let

$$E[\exp(t_1X_1 + t_2X_2 + \cdots + t_nX_n)]$$

exist for $-h_i < t_i < h_i$, $i = 1, 2, \ldots, n$, where each h_i is positive. This expectation is denoted by $M(t_1, t_2, \ldots, t_n)$ and it is called the m.g.f. of the joint distribution of $X_1, \ldots, X_n$ (or simply the m.g.f. of $X_1, \ldots, X_n$). As in the cases of one and two variables, this m.g.f. is unique and uniquely determines the joint distribution of the n

variables (and hence all marginal distributions). For example, the m.g.f. of the marginal distribution of X_i is $M(0, \ldots, 0, t_i, 0, \ldots, 0)$, $i = 1, 2, \ldots, n$; that of the marginal distribution of X_i and X_j is $M(0, \ldots, 0, t_i, 0, \ldots, 0, t_j, 0, \ldots, 0)$; and so on. Theorem 4 of this chapter can be generalized, and the factorization

$$M(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} M(0, \ldots, 0, t_i, 0, \ldots, 0)$$

is a necessary and sufficient condition for the mutual independence of $X_1, X_2, \ldots, X_n$.

Remark. If X_1, X_2, and X_3 are mutually independent, they are *pairwise independent* (that is, X_i and X_j, $i \neq j$, where $i, j = 1, 2, 3$, are independent). However, the following example, due to S. Bernstein, shows that pairwise independence does not necessarily imply mutual independence. Let X_1, X_2, and X_3 have the joint p.d.f.

$$\begin{aligned} f(x_1, x_2, x_3) &= \tfrac{1}{4}, \qquad (x_1, x_2, x_3) \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 1)\}, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

The joint p.d.f. of X_i and X_j, $i \neq j$, is

$$\begin{aligned} f_{ij}(x_i, x_j) &= \tfrac{1}{4}, \qquad (x_i, x_j) \in \{(0, 0), (1, 0), (0, 1), (1, 1)\}, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

whereas the marginal p.d.f. of X_i is

$$\begin{aligned} f_i(x_i) &= \tfrac{1}{2}, \qquad x_i = 0, 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Obviously, if $i \neq j$, we have

$$f_{ij}(x_i, x_j) \equiv f_i(x_i) f_j(x_j),$$

and thus X_i and X_j are independent. However,

$$f(x_1, x_2, x_3) \not\equiv f_1(x_1) f_2(x_2) f_3(x_3).$$

Thus X_1, X_2, and X_3 are not mutually independent.

Example 2. Let X_1, X_2, and X_3 be three mutually independent random variables and let each have the p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. The joint p.d.f. of X_1, X_2, X_3 is $f(x_1)f(x_2)f(x_3) = 8x_1x_2x_3$, $0 < x_i < 1$, $i = 1, 2, 3$, zero elsewhere. Then, for illustration, the expected value of $5X_1X_2^3 + 3X_2X_3^4$ is

$$\int_0^1 \int_0^1 \int_0^1 (5x_1x_2^3 + 3x_2x_3^4) 8x_1x_2x_3 \, dx_1 \, dx_2 \, dx_3 = 2.$$

Let Y be the maximum of X_1, X_2, and X_3. Then, for instance, we have

$$\Pr(Y \le \tfrac{1}{2}) = \Pr(X_1 \le \tfrac{1}{2}, X_2 \le \tfrac{1}{2}, X_3 \le \tfrac{1}{2})$$
$$= \int_0^{1/2}\int_0^{1/2}\int_0^{1/2} 8x_1x_2x_3\,dx_1\,dx_2\,dx_3$$
$$= (\tfrac{1}{2})^6 = \tfrac{1}{64}.$$

In a similar manner, we find that the distribution function of Y is

$$\begin{aligned} G(y) = \Pr(Y \le y) &= 0, &\quad y < 0 \\ &= y^6, &\quad 0 \le y < 1, \\ &= 1, &\quad 1 \le y. \end{aligned}$$

Accordingly, the p.d.f. of Y is

$$\begin{aligned} g(y) &= 6y^5, &\quad 0 < y < 1, \\ &= 0 &\quad \text{elsewhere.} \end{aligned}$$

Remark. Unless there is a possible misunderstanding between *mutual* and *pairwise* independence, we usually drop the modifier *mutual*. Accordingly, using this practice in Example 2, we say that X_1, X_2, X_3 are independent random variables, meaning that they are mutually independent. Occasionally, for emphasis, we use *mutually independent* so that the reader is reminded that this is different from *pairwise independence*.

EXERCISES

2.37. Let X, Y, Z have joint p.d.f. $f(x, y, z) = 2(x + y + z)/3$, $0 < x < 1$, $0 < y < 1$, $0 < z < 1$, zero elsewhere.

(a) Find the marginal probability density functions.

(b) Compute $\Pr(0 < X < \tfrac{1}{2}, 0 < Y < \tfrac{1}{2}, 0 < Z < \tfrac{1}{2})$ and $\Pr(0 < X < \tfrac{1}{2}) = \Pr(0 < Y < \tfrac{1}{2}) = \Pr(0 < Z < \tfrac{1}{2})$.

(c) Are X, Y, and Z independent?

(d) Calculate $E(X^2YZ + 3XY^4Z^2)$.

(e) Determine the distribution function of X, Y, and Z.

(f) Find the conditional distribution of X and Y, given $Z = z$, and evaluate $E(X + Y|z)$.

(g) Determine the conditional distribution of X, given $Y = y$ and $Z = z$, and compute $E(X|y, z)$.

2.38. Let $f(x_1, x_2, x_3) = \exp[-(x_1 + x_2 + x_3)]$, $0 < x_1 < \infty$, $0 < x_2 < \infty$, $0 < x_3 < \infty$, zero elsewhere, be the joint p.d.f. of X_1, X_2, X_3.

(a) Compute $\Pr(X_1 < X_2 < X_3)$ and $\Pr(X_1 = X_2 < X_3)$.

(b) Determine the m.g.f. of X_1, X_2, and X_3. Are these random variables independent?

2.39. Let X_1, X_2, X_3, and X_4 be four independent random variables, each with p.d.f. $f(x) = 3(1-x)^2$, $0 < x < 1$, zero elsewhere. If Y is the minimum of these four variables, find the distribution function and the p.d.f. of Y.

2.40. A fair die is cast at random three independent times. Let the random variable X_i be equal to the number of spots that appear on the ith trial, $i = 1, 2, 3$. Let the random variable Y be equal to $\max(X_i)$. Find the distribution function and the p.d.f. of Y.

Hint: $\Pr(Y \le y) = \Pr(X_i \le y, i = 1, 2, 3)$.

2.41. Let $M(t_1, t_2, t_3)$ be the m.g.f. of the random variables X_1, X_2, and X_3 of Bernstein's example, described in the remark preceding Example 2 of this section. Show that $M(t_1, t_2, 0) = M(t_1, 0, 0)M(0, t_2, 0)$, $M(t_1, 0, t_3) = M(t_1, 0, 0)M(0, 0, t_3)$, $M(0, t_2, t_3) = M(0, t_2, 0)M(0, 0, t_3)$, but $M(t_1, t_2, t_3) \ne M(t_1, 0, 0)M(0, t_2, 0)\,M(0, 0, t_3)$. Thus X_1, X_2, X_3 are pairwise independent but not mutually independent.

2.42. Let X_1, X_2, and X_3 be three random variables with means, variances, and correlation coefficients, denoted by μ_1, μ_2, μ_3; $\sigma_1^2, \sigma_2^2, \sigma_3^2$; and $\rho_{12}, \rho_{13}, \rho_{23}$, respectively. If $E(X_1 - \mu_1 | x_2, x_3) = b_2(x_2 - \mu_2) + b_3(x_3 - \mu_3)$, where b_2 and b_3 are constants, determine b_2 and b_3 in terms of the variances and the correlation coefficients.

ADDITIONAL EXERCISES

2.43. Find $\Pr[X_1 X_2 \le 2]$, where X_1 and X_2 are independent and each has the distribution with p.d.f. $f(x) = 1$, $1 < x < 2$, zero elsewhere.

2.44. Let the joint p.d.f. of X and Y be given by $f(x, y) = \dfrac{2}{(1 + x + y)^3}$, $0 < x < \infty$, $0 < y < \infty$, zero elsewhere.

(a) Compute the marginal p.d.f. of X and the conditional p.d.f. of Y, given $X = x$.

(b) For a fixed $X = x$, compute $E(1 + x + Y|x)$ and use the result to compute $E(Y|x)$.

2.45. Let X_1, X_2, X_3 be independent and each have a distribution with p.d.f. $f(x) = \exp(-x)$, $0 < x < \infty$, zero elsewhere. Evaluate:

(a) $\Pr(X_1 < X_2 | X_1 < 2X_2)$.

(b) $\Pr(X_1 < X_2 < X_3 | X_3 < 1)$.

2.46. Let X and Y be random variables with space consisting of the four points: $(0, 0)$, $(1, 1)$, $(1, 0)$, $(1, -1)$. Assign positive probabilities to these four points so that the correlation coefficient is equal to zero. Are X and Y independent?

2.47. Two line segments, each of length 2 units, are placed along the x-axis. The midpoint of the first is between $x = 0$ and $x = 14$ and that of the second is between $x = 6$ and $x = 20$. Assuming independence and uniform distributions for these midpoints, find the probability that the line segments overlap.

2.48. Let X and Y have the joint p.d.f. $f(x, y) = \frac{1}{7}$, $(x, y) = (0, 0), (1, 0), (0, 1), (1, 1), (2, 1), (1, 2), (2, 2)$, and zero elsewhere. Find the correlation coefficient ρ.

2.49. Let X_1 and X_2 have the joint p.d.f. described by the following table:

(x_1, x_2)	(0, 0)	(0, 1)	(0, 2)	(1, 1)	(1, 2)	(2, 2)
$f(x_1, x_2)$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{3}{12}$	$\frac{4}{12}$	$\frac{1}{12}$

Find $f_1(x_1)$, $f_2(x_2)$, μ_1, μ_2, σ_1^2, σ_2^2, and ρ.

2.50. If the discrete random variables X_1 and X_2 have joint p.d.f. $f(x_1, x_2) = (3x_1 + x_2)/24$, $(x_1, x_2) = (1, 1), (1, 2), (2, 1), (2, 2)$, zero elsewhere, find the conditional mean $E(X_2|x_1)$, when $x_1 = 1$.

2.51. Let X and Y have the joint p.d.f. $f(x, y) = 21x^2y^3$, $0 < x < y < 1$, zero elsewhere. Find the conditional mean $E(Y|x)$ of Y, given $X = x$.

2.52. Let X_1 and X_2 have the p.d.f. $f(x_1, x_2) = x_1 + x_2$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere. Evaluate $\Pr(X_1/X_2 \leq 2)$.

2.53. Cast a fair die and let $X = 0$ if 1, 2, or 3 spots appear, let $X = 1$ if 4 or 5 spots appear, and let $X = 2$ if 6 spots appear. Do this two independent times, obtaining X_1 and X_2. Calculate $\Pr(|X_1 - X_2| = 1)$.

2.54. Let $\sigma_1^2 = \sigma_2^2 = \sigma^2$ be the common variance of X_1 and X_2 and let ρ be the correlation coefficient of X_1 and X_2. Show that

$$\Pr[|(X_1 - \mu_1) + (X_2 - \mu_2)| \geq k\sigma] \leq \frac{2(1 + \rho)}{k^2}.$$

CHAPTER 3

Some Special Distributions

3.1 The Binomial and Related Distributions

In Chapter 1 we introduced the *uniform distribution* and the *hypergeometric distribution*. In this chapter we discuss some other important distributions of random variables frequently used in statistics. We begin with the binomial and related distributions.

A *Bernoulli experiment* is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, say, success or failure (e.g., female or male, life or death, nondefective or defective). A sequence of *Bernoulli trials* occurs when a Bernoulli experiment is performed several independent times so that the probability of success, say p, remains the same from trial to trial. That is, in such a sequence, we let p denote the probability of success on each trial.

Let X be a random variable associated with a Bernoulli trial by defining it as follows:

$$X(\text{success}) = 1 \qquad \text{and} \qquad X(\text{failure}) = 0.$$

That is, the two outcomes, success and failure, are denoted by one and zero, respectively. The p.d.f. of X can be written as

$$f(x) = p^x(1-p)^{1-x}, \qquad x = 0, 1,$$

and we say that X has a *Bernoulli distribution.* The expected value of X is

$$\mu = E(X) = \sum_{x=0}^{1} xp^x(1-p)^{1-x} = (0)(1-p) + (1)(p) = p,$$

and the variance of X is

$$\sigma^2 = \text{var}(X) = \sum_{x=0}^{1} (x-p)^2 p^x(1-p)^{1-x}$$

$$= p^2(1-p) + (1-p)^2 p = p(1-p).$$

It follows that the standard deviation of X is $\sigma = \sqrt{p(1-p)}$.

In a sequence of n Bernoulli trials, we shall let X_i denote the Bernoulli random variable associated with the ith trial. An observed sequence of n Bernoulli trials will then be an n-tuple of zeros and ones. In such a sequence of Bernoulli trials, we are often interested in the total number of successes and not in the order of their occurrence. If we let the random variable X equal the number of observed successes in n Bernoulli trials, the possible values of X are $0, 1, 2, \ldots, n$. If x successes occur, where $x = 0, 1, 2, \ldots, n$, then $n - x$ failures occur. The number of ways of selecting x positions for the x successes in the n trials is

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}.$$

Since the trials are independent and since the probabilities of success and failure on each trial are, respectively, p and $1 - p$, the probability of each of these ways is $p^x(1-p)^{n-x}$. Thus the p.d.f. of X, say $f(x)$, is the sum of the probabilities of these $\binom{n}{x}$ mutually exclusive events; that is,

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n,$$

$$= 0 \qquad \text{elsewhere.}$$

Recall, if n is a positive integer, that

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} b^x a^{n-x}.$$

Thus it is clear that $f(x) \geq 0$ and that

$$\sum_x f(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= [(1-p)+p]^n = 1.$$

That is, $f(x)$ satisfies the conditions of being a p.d.f. of a random variable X of the discrete type. A random variable X that has a p.d.f. of the form of $f(x)$ is said to have a *binomial distribution*, and any such $f(x)$ is called a *binomial p.d.f.* A binomial distribution will be denoted by the symbol $b(n, p)$. The constants n and p are called the *parameters* of the binomial distribution. Thus, if we say that X is $b(5, \frac{1}{3})$, we mean that X has the binomial p.d.f.

$$f(x) = \binom{5}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{5-x}, \qquad x = 0, 1, \ldots, 5,$$

$$= 0 \qquad \text{elsewhere.}$$

The m.g.f. of a binomial distribution is easily found. It is

$$M(t) = \sum_x e^{tx} f(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x}$$

$$= [(1-p) + pe^t]^n$$

for all real values of t. The mean μ and the variance σ^2 of X may be computed from $M(t)$. Since

$$M'(t) = n[(1-p)+pe^t]^{n-1}(pe^t)$$

and

$$M''(t) = n[(1-p)+pe^t]^{n-1}(pe^t) + n(n-1)[(1-p)+pe^t]^{n-2}(pe^t)^2,$$

it follows that

$$\mu = M'(0) = np$$

and

$$\sigma^2 = M''(0) - \mu^2 = np + n(n-1)p^2 - (np)^2 = np(1-p).$$

Example 1. Let X be the number of heads (successes) in $n = 7$ independent tosses of an unbiased coin. The p.d.f. of X is

$$f(x) = \binom{7}{x}\left(\frac{1}{2}\right)^x\left(1 - \frac{1}{2}\right)^{7-x}, \qquad x = 0, 1, 2, \ldots, 7,$$

$$= 0 \qquad \text{elsewhere.}$$

Then X has the m.g.f.

$$M(t) = (\tfrac{1}{2} + \tfrac{1}{2}e^t)^7,$$

has mean $\mu = np = \frac{7}{2}$, and has variance $\sigma^2 = np(1 - p) = \frac{7}{4}$. Furthermore, we have

$$\Pr(0 \le X \le 1) = \sum_{x=0}^{1} f(x) = \frac{1}{128} + \frac{7}{128} = \frac{8}{128}$$

and

$$\Pr(X = 5) = f(5)$$

$$= \frac{7!}{5!\,2!}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^2 = \frac{21}{128}.$$

Example 2. If the m.g.f. of a random variable X is

$$M(t) = (\tfrac{2}{3} + \tfrac{1}{3}e^t)^5,$$

then X has a binomial distribution with $n = 5$ and $p = \frac{1}{3}$; that is, the p.d.f. of X is

$$f(x) = \binom{5}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{5-x}, \qquad x = 0, 1, 2, \ldots, 5,$$

$$= 0 \qquad \text{elsewhere.}$$

Here $\mu = np = \frac{5}{3}$ and $\sigma^2 = np(1 - p) = \frac{10}{9}$.

Example 3. If Y is $b(n, \frac{1}{3})$, then $\Pr(Y \ge 1) = 1 - \Pr(Y = 0) = 1 - (\frac{2}{3})^n$. Suppose that we wish to find the smallest value of n that yields $\Pr(Y \ge 1) > 0.80$. We have $1 - (\frac{2}{3})^n > 0.80$ and $0.20 > (\frac{2}{3})^n$. Either by inspection or by use of logarithms, we see that $n = 4$ is the solution. That is, the probability of at least one success throughout $n = 4$ independent repetitions of a random experiment with probability of success $p = \frac{1}{3}$ is greater than 0.80.

Example 4. Let the random variable Y be equal to the number of successes throughout n independent repetitions of a random experiment with probability p of success. That is, Y is $b(n, p)$. The ratio Y/n is called the relative frequency of success. For every $\epsilon > 0$, we have

$$\Pr\left(\left|\frac{Y}{n} - p\right| \ge \epsilon\right) = \Pr(|Y - np| \ge \epsilon n)$$

$$= \Pr\left(|Y - \mu| \ge \epsilon\sqrt{\frac{n}{p(1-p)}}\,\sigma\right),$$

where $\mu = np$ and $\sigma^2 = np(1 - p)$. In accordance with Chebyshev's inequality with $k = \epsilon\sqrt{n/p(1 - p)}$, we have

$$\Pr\left(|Y - \mu| \geq \epsilon\sqrt{\frac{n}{p(1-p)}}\,\sigma\right) \leq \frac{p(1-p)}{n\epsilon^2}$$

and hence

$$\Pr\left(\left|\frac{Y}{n} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Now, for every fixed $\epsilon > 0$, the right-hand member of the preceding inequality is close to zero for sufficiently large n. That is,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{Y}{n} - p\right| \geq \epsilon\right) = 0$$

and

$$\lim_{n\to\infty} \Pr\left(\left|\frac{Y}{n} - p\right| < \epsilon\right) = 1.$$

Since this is true for every fixed $\epsilon > 0$, we see, in a certain sense, that the relative frequency of success is for large values of n, close to the probability p of success. This result is one form of the *law of large numbers*. It was alluded to in the initial discussion of probability in Chapter 1 and will be considered again, along with related concepts, in Chapter 5.

Example 5. Let the independent random variables X_1, X_2, X_3 have the same distribution function $F(x)$. Let Y be the middle value of X_1, X_2, X_3. To determine the distribution function of Y, say $G(y) = \Pr(Y \leq y)$, we note that $Y \leq y$ if and only if at least two of the random variables X_1, X_2, X_3 are less than or equal to y. Let us say that the ith "trial" is a success if $X_i \leq y$, $i = 1, 2, 3$; here each "trial" has the probability of success $F(y)$. In this terminology, $G(y) = \Pr(Y \leq y)$ is then the probability of at least two successes in three independent trials. Thus

$$G(y) = \binom{3}{2}[F(y)]^2[1 - F(y)] + [F(y)]^3.$$

If $F(x)$ is a continuous type of distribution function so that the p.d.f. of X is $F'(x) = f(x)$, then the p.d.f. of Y is

$$g(y) = G'(y) = 6[F(y)][1 - F(y)]f(y).$$

Example 6. Consider a sequence of independent repetitions of a random experiment with constant probability p of success. Let the random variable Y denote the total number of failures in this sequence before the rth success;

that is, $Y + r$ is equal to the number of trials necessary to produce exactly r successes. Here r is a fixed positive integer. To determine the p.d.f. of Y, let y be an element of $\{y : y = 0, 1, 2, \dots\}$. Then, by the multiplication rule of probabilities, $\Pr(Y = y) = g(y)$ is equal to the product of the probability

$$\binom{y+r-1}{r-1} p^{r-1}(1-p)^y$$

of obtaining exactly $r - 1$ successes in the first $y + r - 1$ trials and the probability p of a success on the $(y + r)$th trial. Thus the p.d.f. $g(y)$ of Y is given by

$$g(y) = \binom{y+r-1}{r-1} p^r(1-p)^y, \qquad y = 0, 1, 2, \dots,$$

$$= 0 \qquad \text{elsewhere.}$$

A distribution with a p.d.f. of the form $g(y)$ is called a *negative binomial distribution*; and any such $g(y)$ is called a negative binomial p.d.f. The distribution derives its name from the fact that $g(y)$ is a general term in the expansion of $p^r[1 - (1 - p)]^{-r}$. It is left as an exercise to show that the m.g.f. of this distribution is $M(t) = p^r[1 - (1 - p)e^t]^{-r}$, for $t < -\ln(1 - p)$. If $r = 1$, then Y has the p.d.f.

$$g(y) = p(1-p)^y, \qquad y = 0, 1, 2, \dots,$$

zero elsewhere, and the m.g.f. $M(t) = p[1 - (1 - p)e^t]^{-1}$. In this special case, $r = 1$, we say that Y has a *geometric distribution*.

The binomial distribution is generalized to the multinomial distribution as follows. Let a random experiment be repeated n independent times. On each repetition, the experiment terminates in but one of k mutually exclusive and exhaustive ways, say $C_1, C_2, \dots, C_k$. Let p_i be the probability that the outcome is an element of C_i and let p_i remain constant throughout the n independent repetitions, $i = 1, 2, \dots, k$. Define the random variable X_i to be equal to the number of outcomes that are elements of C_i, $i = 1, 2, \dots, k - 1$. Furthermore, let $x_1, x_2, \dots, x_{k-1}$ be nonnegative integers so that $x_1 + x_2 + \cdots + x_{k-1} \le n$. Then the probability that exactly x_1 terminations of the experiment are in $C_1, \dots,$ exactly x_{k-1} terminations are in C_{k-1}, and hence exactly $n - (x_1 + \cdots + x_{k-1})$ terminations are in C_k is

$$\frac{n!}{x_1! \cdots x_{k-1}!\, x_k!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} p_k^{x_k},$$

where x_k is merely an abbreviation for $n - (x_1 + \cdots + x_{k-1})$. This is the *multinomial p.d.f.* of $k-1$ random variables $X_1, X_2, \ldots, X_{k-1}$ of the discrete type. To see that this is correct, note that the number of distinguishable arrangements of $x_1 C_1$'s, $x_2 C_2$'s, $\ldots$, $x_k C_k$'s is

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-2}}{x_{k-1}} = \frac{n!}{x_1!\,x_2!\cdots x_k!}$$

and that the probability of each of these distinguishable arrangements is

$$p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}.$$

Hence the product of these two latter expressions gives the correct probability, which is in agreement with the formula for the multinomial p.d.f.

When $k=3$, we often let $X = X_1$ and $Y = X_2$; then $n - X - Y = X_3$. We say that X and Y have a *trinomial distribution.* The joint p.d.f. of X and Y is

$$f(x, y) = \frac{n!}{x!\,y!\,(n-x-y)!}\,p_1^x p_2^y p_3^{n-x-y},$$

where x and y are nonnegative integers with $x + y \le n$, and p_1, p_2, and p_3 are positive proper fractions with $p_1 + p_2 + p_3 = 1$; and let $f(x, y) = 0$ elsewhere. Accordingly, $f(x, y)$ satisfies the conditions of being a joint p.d.f. of two random variables X and Y of the discrete type; that is, $f(x, y)$ is nonnegative and its sum over all points (x, y) at which $f(x, y)$ is positive is equal to $(p_1 + p_2 + p_3)^n = 1$.

If n is a positive integer and a_1, a_2, a_3 are fixed constants, we have

$$\sum_{x=0}^{n}\sum_{y=0}^{n-x}\frac{n!}{x!\,y!\,(n-x-y)!}\,a_1^x a_2^y a_3^{n-x-y}$$

$$= \sum_{x=0}^{n}\frac{n!\,a_1^x}{x!\,(n-x)!}\sum_{y=0}^{n-x}\frac{(n-x)!}{y!\,(n-x-y)!}\,a_2^y a_3^{n-x-y}$$

$$= \sum_{x=0}^{n}\frac{n!}{x!\,(n-x)!}\,a_1^x(a_2+a_3)^{n-x}$$

$$= (a_1 + a_2 + a_3)^n. \qquad (1)$$

Consequently, the m.g.f. of a trinomial distribution, in accordance with Equation (1), is given by

$$M(t_1, t_2) = \sum_{x=0}^{n}\sum_{y=0}^{n-x}\frac{n!}{x!\,y!\,(n-x-y)!}\,(p_1e^{t_1})^x(p_2e^{t_2})^y p_3^{n-x-y}$$

$$= (p_1e^{t_1} + p_2e^{t_2} + p_3)^n,$$

for all real values of t_1 and t_2. The moment-generating functions of the marginal distributions of X and Y are, respectively,

$$M(t_1, 0) = (p_1e^{t_1} + p_2 + p_3)^n = [(1 - p_1) + p_1e^{t_1}]^n$$

and

$$M(0, t_2) = (p_1 + p_2e^{t_2} + p_3)^n = [(1 - p_2) + p_2e^{t_2}]^n.$$

We see immediately, from Theorem 4, Section 2.4, that X and Y are dependent random variables. In addition, X is $b(n, p_1)$ and Y is $b(n, p_2)$. Accordingly, the means and the variances of X and Y are, respectively, $\mu_1 = np_1$, $\mu_2 = np_2$, $\sigma_1^2 = np_1(1 - p_1)$, and $\sigma_2^2 = np_2(1 - p_2)$.

Consider next the conditional p.d.f. of Y, given $X = x$. We have

$$f_{2|1}(y|x) = \frac{(n - x)!}{y!\,(n-x-y)!}\left(\frac{p_2}{1-p_1}\right)^y\left(\frac{p_3}{1-p_1}\right)^{n-x-y}, \quad y=0, 1, \dots, n-x,$$

$$= 0 \qquad \text{elsewhere.}$$

Thus the conditional distribution of Y, given $X = x$, is $b[n - x, p_2/(1 - p_1)]$. Hence the conditional mean of Y, given $X = x$, is the linear function

$$E(Y|x) = (n - x)\left(\frac{p_2}{1 - p_1}\right).$$

We also find that the conditional distribution of X, given $Y = y$, is $b[n - y, p_1/(1 - p_2)]$ and thus

$$E(X|y) = (n - y)\left(\frac{p_1}{1 - p_2}\right).$$

Now recall (Example 2, Section 2.3) that the square of the correlation coefficient, say ρ^2, is equal to the product of $-p_2/(1 - p_1)$ and $-p_1/(1 - p_2)$, the coefficients of x and y in the respective conditional means. Since both of these coefficients are negative (and thus ρ is negative), we have

$$\rho = -\sqrt{\frac{p_1p_2}{(1 - p_1)(1 - p_2)}}.$$

In general, the m.g.f. of a multinomial distribution is given by

$$M(t_1, \dots, t_{k-1}) = (p_1e^{t_1} + \cdots + p_{k-1}e^{t_{k-1}} + p_k)^n$$

for all real values of $t_1, t_2, \ldots, t_{k-1}$. Thus each one-variable marginal p.d.f. is binomial, each two-variable marginal p.d.f. is trinomial, and so on.

EXERCISES

3.1. If the m.g.f. of a random variable X is $(\frac{1}{3} + \frac{2}{3}e^t)^5$, find Pr ($X = 2$ or 3).

3.2. The m.g.f. of a random variable X is $(\frac{2}{3} + \frac{1}{3}e^t)^9$. Show that

$$\Pr(\mu - 2\sigma < X < \mu + 2\sigma) = \sum_{x=1}^{5} \binom{9}{x}\left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{9-x}.$$

3.3. If X is $b(n, p)$, show that

$$E\left(\frac{X}{n}\right) = p \quad \text{and} \quad E\left[\left(\frac{X}{n} - p\right)^2\right] = \frac{p(1-p)}{n}.$$

3.4. Let the independent random variables X_1, X_2, X_3 have the same p.d.f. $f(x) = 3x^2, 0 < x < 1$, zero elsewhere. Find the probability that exactly two of these three variables exceed $\frac{1}{2}$.

3.5. Let Y be the number of successes in n independent repetitions of a random experiment having the probability of success $p = \frac{2}{3}$. If $n = 3$, compute Pr $(2 \le Y)$; if $n = 5$, compute Pr $(3 \le Y)$.

3.6. Let Y be the number of successes throughout n independent repetitions of a random experiment having probability of success $p = \frac{1}{4}$. Determine the smallest value of n so that Pr $(1 \le Y) \ge 0.70$.

3.7. Let the independent random variables X_1 and X_2 have binomial distributions with parameters $n_1 = 3, p_1 = \frac{2}{3}$ and $n_2 = 4, p_2 = \frac{1}{2}$, respectively. Compute Pr $(X_1 = X_2)$.

Hint: List the four mutually exclusive ways that $X_1 = X_2$ and compute the probability of each.

3.8. Toss two nickels and three dimes at random. Make appropriate assumptions and compute the probability that there are more heads showing on the nickels than on the dimes.

3.9. Let $X_1, X_2, \ldots, X_{k-1}$ have a multinomial distribution.

(a) Find the m.g.f. of $X_2, X_3, \ldots, X_{k-1}$.

(b) What is the p.d.f. of $X_2, X_3, \ldots, X_{k-1}$?

(c) Determine the conditional p.d.f. of X_1, given that

$$X_2 = x_2, \ldots, X_{k-1} = x_{k-1}.$$

(d) What is the conditional expectation $E(X_1|x_2, \ldots, x_{k-1})$?

3.10. Let X be $b(2, p)$ and let Y be $b(4, p)$. If Pr $(X \ge 1) = \frac{5}{9}$, find Pr $(Y \ge 1)$.

3.11. If $x = r$ is the unique mode of a distribution that is $b(n, p)$, show that

$$(n + 1)p - 1 < r < (n + 1)p.$$

Hint: Determine the values of x for which the ratio $f(x + 1)/f(x) > 1$.

3.12. Let X have a binomial distribution with parameters n and $p = \frac{1}{3}$. Determine the smallest integer n can be such that $\Pr(X \geq 1) \geq 0.85$.

3.13. Let X have the p.d.f. $f(x) = (\frac{1}{3})(\frac{2}{3})^x$, $x = 0, 1, 2, 3, \ldots$, zero elsewhere. Find the conditional p.d.f. of X, given that $X \geq 3$.

3.14. One of the numbers $1, 2, \ldots, 6$ is to be chosen by casting an unbiased die. Let this random experiment be repeated five independent times. Let the random variable X_1 be the number of terminations in the set $\{x : x = 1, 2, 3\}$ and let the random variable X_2 be the number of terminations in the set $\{x : x = 4, 5\}$. Compute $\Pr(X_1 = 2, X_2 = 1)$.

3.15. Show that the m.g.f. of the negative binomial distribution is $M(t) = p^r[1 - (1 - p)e^t]^{-r}$. Find the mean and the variance of this distribution.

Hint: In the summation representing $M(t)$, make use of the MacLaurin's series for $(1 - w)^{-r}$.

3.16. Let X_1 and X_2 have a trinomial distribution. Differentiate the moment-generating function to show that their covariance is $-np_1p_2$.

3.17. If a fair coin is tossed at random five independent times, find the conditional probability of five heads relative to the hypothesis that there are at least four heads.

3.18. Let an unbiased die be cast at random seven independent times. Compute the conditional probability that each side appears at least once relative to the hypothesis that side 1 appears exactly twice.

3.19. Compute the measures of skewness and kurtosis of the binomial distribution $b(n, p)$.

3.20. Let

$$f(x_1, x_2) = \binom{x_1}{x_2}\left(\frac{1}{2}\right)^{x_1}\left(\frac{x_1}{15}\right), \quad \begin{array}{l} x_2 = 0, 1, \ldots, x_1, \\ x_1 = 1, 2, 3, 4, 5, \end{array}$$

zero elsewhere, be the joint p.d.f. of X_1 and X_2. Determine:

(a) $E(X_2)$.

(b) $u(x_1) = E(X_2|x_1)$.

(c) $E[u(X_1)]$.

Compare the answers of parts (a) and (c).

Hint: Note that $E(X_2) = \sum_{x_1=1}^{5} \sum_{x_2=0}^{x_1} x_2 f(x_1, x_2)$ and use the fact that $\sum_{y=0}^{n} y \binom{n}{y} (\frac{1}{2})^n = n/2$. Why?

3.21. Three fair dice are cast. In 10 independent casts, let X be the number of times all three faces are alike and let Y be the number of times only two faces are alike. Find the joint p.d.f. of X and Y and compute $E(6XY)$.

3.2 The Poisson Distribution

Recall that the series

$$1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{m^x}{x!}$$

converges, for all values of m, to e^m. Consider the function $f(x)$ defined by

$$\begin{aligned} f(x) &= \frac{m^x e^{-m}}{x!}, \qquad x = 0, 1, 2, \ldots, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

where $m > 0$. Since $m > 0$, then $f(x) \geq 0$ and

$$\sum_x f(x) = \sum_{x=0}^{\infty} \frac{m^x e^{-m}}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{m^x}{x!} = e^{-m} e^m = 1;$$

that is, $f(x)$ satisfies the conditions of being a p.d.f. of a discrete type of random variable. A random variable that has a p.d.f. of the form $f(x)$ is said to have a *Poisson distribution*, and any such $f(x)$ is called a *Poisson p.d.f.*

Remarks. Experience indicates that the Poisson p.d.f. may be used in a number of applications with quite satisfactory results. For example, let the random variable X denote the number of alpha particles emitted by a radioactive substance that enter a prescribed region during a prescribed interval of time. With a suitable value of m, it is found that X may be assumed to have a Poisson distribution. Again let the random variable X denote the number of defects on a manufactured article, such as a refrigerator door. Upon examining many of these doors, it is found, with an appropriate value of m, that X may be said to have a Poisson distribution. The number of automobile accidents in some unit of time (or the number of insurance claims in some unit of time) is often assumed to be a random variable which has a Poisson distribution. Each of these instances can be thought of as a process that generates a number of changes (accidents, claims, etc.) in a fixed interval (of time or space, etc.). If a process leads to a Poisson distribution, that process is called a *Poisson process*. Some assumptions that ensure a Poisson process will now be enumerated.

Let $g(x, w)$ denote the probability of x changes in each interval of length w. Furthermore, let the symbol $o(h)$ represent any function such that $\lim_{h \to 0} [o(h)/h] = 0$; for example, $h^2 = o(h)$ and $o(h) + o(h) = o(h)$. The Poisson postulates are the following:

1. $g(1, h) = \lambda h + o(h)$, where λ is a positive constant and $h > 0$.
2. $\sum_{x=2}^{\infty} g(x, h) = o(h)$.
3. The numbers of changes in nonoverlapping intervals are independent.

Postulates 1 and 3 state, in effect, that the probability of one change in a short interval h is independent of changes in other nonoverlapping intervals and is approximately proportional to the length of the interval. The substance of postulate 2 is that the probability of two or more changes in the same short interval h is essentially equal to zero. If $x = 0$, we take $g(0, 0) = 1$. In accordance with postulates 1 and 2, the probability of at least one change in an interval of length h is $\lambda h + o(h) + o(h) = \lambda h + o(h)$. Hence the probability of zero changes in this interval of length h is $1 - \lambda h - o(h)$. Thus the probability $g(0, w + h)$ of zero changes in an interval of length $w + h$ is, in accordance with postulate 3, equal to the product of the probability $g(0, w)$ of zero changes in an interval of length w and the probability $[1 - \lambda h - o(h)]$ of zero changes in a nonoverlapping interval of length h. That is,

$$g(0, w + h) = g(0, w)[1 - \lambda h - o(h)].$$

Then

$$\frac{g(0, w + h) - g(0, w)}{h} = -\lambda g(0, w) - \frac{o(h)g(0, w)}{h}.$$

If we take the limit as $h \to 0$, we have

$$D_w[g(0, w)] = -\lambda g(0, w).$$

The solution of this differential equation is

$$g(0, w) = ce^{-\lambda w}.$$

The condition $g(0, 0) = 1$ implies that $c = 1$; so

$$g(0, w) = e^{-\lambda w}.$$

If x is a positive integer, we take $g(x, 0) = 0$. The postulates imply that

$$g(x, w + h) = [g(x, w)][1 - \lambda h - o(h)] + [g(x - 1, w)][\lambda h + o(h)] + o(h).$$

Accordingly, we have

$$\frac{g(x, w + h) - g(x, w)}{h} = -\lambda g(x, w) + \lambda g(x - 1, w) + \frac{o(h)}{h}$$

and

$$D_w[g(x, w)] = -\lambda g(x, w) + \lambda g(x - 1, w),$$

for $x = 1, 2, 3, \ldots$. It can be shown, by mathematical induction, that the solutions to these differential equations, with boundary conditions $g(x, 0) = 0$ for $x = 1, 2, 3, \ldots$, are, respectively,

$$g(x, w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}, \qquad x = 1, 2, 3, \ldots .$$

Hence the number of changes X in an interval of length w has a Poisson distribution with parameter $m = \lambda w$.

The m.g.f. of a Poisson distribution is given by

$$M(t) = \sum_x e^{tx} f(x) = \sum_{x=0}^{\infty} e^{tx} \frac{m^x e^{-m}}{x!}$$

$$= e^{-m} \sum_{x=0}^{\infty} \frac{(me^t)^x}{x!}$$

$$= e^{-m} e^{me^t} = e^{m(e^t - 1)}$$

for all real values of t. Since

$$M'(t) = e^{m(e^t - 1)}(me^t)$$

and

$$M''(t) = e^{m(e^t - 1)}(me^t) + e^{m(e^t - 1)}(me^t)^2,$$

then

$$\mu = M'(0) = m$$

and

$$\sigma^2 = M''(0) - \mu^2 = m + m^2 - m^2 = m.$$

That is, a Poisson distribution has $\mu = \sigma^2 = m > 0$. On this account, a Poisson p.d.f. is frequently written

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}, \qquad x = 0, 1, 2, \ldots,$$

$$= 0 \qquad \text{elsewhere.}$$

Thus the parameter m in a Poisson p.d.f. is the mean μ. Table I in Appendix B gives approximately the distribution for various values of the parameter $m = \mu$.

Example 1. Suppose that X has a Poisson distribution with $\mu = 2$. Then the p.d.f. of X is

$$f(x) = \frac{2^x e^{-2}}{x!}, \qquad x = 0, 1, 2, \dots,$$
$$= 0 \qquad \text{elsewhere.}$$

The variance of this distribution is $\sigma^2 = \mu = 2$. If we wish to compute $\Pr(1 \le X)$, we have

$$\Pr(1 \le X) = 1 - \Pr(X = 0)$$
$$= 1 - f(0) = 1 - e^{-2} = 0.865,$$

approximately, by Table I of Appendix B.

Example 2. If the m.g.f. of a random variable X is

$$M(t) = e^{4(e^t - 1)},$$

then X has a Poisson distribution with $\mu = 4$. Accordingly, by way of example,

$$\Pr(X = 3) = \frac{4^3 e^{-4}}{3!} = \frac{32}{3} e^{-4};$$

or, by Table I,

$$\Pr(X = 3) = \Pr(X \le 3) - \Pr(X \le 2) = 0.433 - 0.238 = 0.195.$$

Example 3. Let the probability of exactly one blemish in 1 foot of wire be about $\frac{1}{1000}$ and let the probability of two or more blemishes in that length be, for all practical purposes, zero. Let the random variable X be the number of blemishes in 3000 feet of wire. If we assume the independence of the numbers of blemishes in nonoverlapping intervals, then the postulates of the Poisson process are approximated, with $\lambda = \frac{1}{1000}$ and $w = 3000$. Thus X has an approximate Poisson distribution with mean $3000(\frac{1}{1000}) = 3$. For example, the probability that there are exactly five blemishes in 3000 feet of wire is

$$\Pr(X = 5) = \frac{3^5 e^{-3}}{5!}$$

and by Table I,

$$\Pr(X = 5) = \Pr(X \le 5) - \Pr(X \le 4) = 0.101,$$

approximately.

EXERCISES

3.22. If the random variable X has a Poisson distribution such that $\Pr(X = 1) = \Pr(X = 2)$, find $\Pr(X = 4)$.

3.23. The m.g.f. of a random variable X is $e^{4(e^t - 1)}$. Show that $\Pr(\mu - 2\sigma < X < \mu + 2\sigma) = 0.931$.

3.24. In a lengthy manuscript, it is discovered that only 13.5 percent of the pages contain no typing errors. If we assume that the number of errors per page is a random variable with a Poisson distribution, find the percentage of pages that have exactly one error.

3.25. Let the p.d.f. $f(x)$ be positive on and only on the nonnegative integers. Given that $f(x) = (4/x)f(x - 1)$, $x = 1, 2, 3, \ldots$. Find $f(x)$.

Hint: Note that $f(1) = 4f(0)$, $f(2) = (4^2/2!)f(0)$, and so on. That is, find each $f(x)$ in terms of $f(0)$ and then determine $f(0)$ from

$$1 = f(0) + f(1) + f(2) + \cdots.$$

3.26. Let X have a Poisson distribution with $\mu = 100$. Use Chebyshev's inequality to determine a lower bound for $\Pr(75 < X < 125)$.

3.27. Given that $g(x, 0) = 0$ and that

$$D_w[g(x, w)] = -\lambda g(x, w) + \lambda g(x - 1, w)$$

for $x = 1, 2, 3, \ldots$. If $g(0, w) = e^{-\lambda w}$, show, by mathematical induction, that

$$g(x, w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}, \qquad x = 1, 2, 3, \ldots.$$

3.28. Let the number of chocolate drops in a certain type of cookie have a Poisson distribution. We want the probability that a cookie of this type contains at least two chocolate drops to be greater than 0.99. Find the smallest value that the mean of the distribution can take.

3.29. Compute the measures of skewness and kurtosis of the Poisson distribution with mean μ.

3.30. On the average a grocer sells 3 of a certain article per week. How many of these should he have in stock so that the chance of his running out within a week will be less than 0.01? Assume a Poisson distribution.

3.31. Let X have a Poisson distribution. If $\Pr(X = 1) = \Pr(X = 3)$, find the mode of the distribution.

3.32. Let X have a Poisson distribution with mean 1. Compute, if it exists, the expected value $E(X!)$.

3.33. Let X and Y have the joint p.d.f. $f(x, y) = e^{-2}/[x!(y-x)!]$, $y = 0, 1, 2, \ldots$; $x = 0, 1, \ldots, y$, zero elsewhere.

(a) Find the m.g.f. $M(t_1, t_2)$ of this joint distribution.

(b) Compute the means, the variances, and the correlation coefficient of X and Y.

(c) Determine the conditional mean $E(X|y)$.

Hint: Note that

$$\sum_{x=0}^{y} [\exp(t_1 x)]y!/[x!(y-x)!] = [1 + \exp(t_1)]^y.$$

Why?

3.3 The Gamma and Chi-Square Distributions

In this section we introduce the gamma and chi-square distributions. It is proved in books on advanced calculus that the integral

$$\int_0^\infty y^{\alpha-1}e^{-y}\,dy$$

exists for $\alpha > 0$ and that the value of the integral is a positive number. The integral is called the gamma function of α, and we write

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy.$$

If $\alpha = 1$, clearly

$$\Gamma(1) = \int_0^\infty e^{-y}\,dy = 1.$$

If $\alpha > 1$, an integration by parts shows that

$$\Gamma(\alpha) = (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}\,dy = (\alpha-1)\Gamma(\alpha-1).$$

Accordingly, if α is a positive integer greater than 1,

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\cdots(3)(2)(1)\Gamma(1) = (\alpha-1)!.$$

Since $\Gamma(1) = 1$, this suggests that we take $0! = 1$, as we have done.

In the integral that defines $\Gamma(\alpha)$, let us introduce a new variable x by writing $y = x/\beta$, where $\beta > 0$. Then

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

or, equivalently,

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}\, dx.$$

Since $\alpha > 0$, $\beta > 0$, and $\Gamma(\alpha) > 0$, we see that

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad 0 < x < \infty, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

is a p.d.f. of a random variable of the continuous type. A random variable X that has a p.d.f. of this form is said to have a *gamma distribution* with parameters α and β; and any such $f(x)$ is called a *gamma-type p.d.f.*

Remark. The gamma distribution is frequently the probability model for waiting times; for instance, in life testing, the waiting time until "death" is the random variable which frequently has a gamma distribution. To see this, let us assume the postulates of a Poisson process and let the interval of length w be a time interval. Specifically, let the random variable W be the time that is needed to obtain exactly k changes (possibly deaths), where k is a fixed positive integer. Then the distribution function of W is

$$G(w) = \Pr(W \le w) = 1 - \Pr(W > w).$$

However, the event $W > w$, for $w > 0$, is equivalent to the event in which there are less than k changes in a time interval of length w. That is, if the random variable X is the number of changes in an interval of length w, then

$$\Pr(W > w) = \sum_{x=0}^{k-1} \Pr(X = x) = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!}.$$

It is left as an exercise to verify that

$$\int_{\lambda w}^\infty \frac{z^{k-1} e^{-z}}{(k-1)!}\, dz = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!}.$$

If, momentarily, we accept this result, we have, for $w > 0$,

$$G(w) = 1 - \int_{\lambda w}^\infty \frac{z^{k-1} e^{-z}}{\Gamma(k)}\, dz = \int_0^{\lambda w} \frac{z^{k-1} e^{-z}}{\Gamma(k)}\, dz,$$

and for $w \le 0$, $G(w) = 0$. If we change the variable of integration in the integral that defines $G(w)$ by writing $z = \lambda y$, then

$$G(w) = \int_0^w \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma(k)}\, dy, \qquad w > 0,$$

and $G(w) = 0$, $w \le 0$. Accordingly, the p.d.f. of W is

$$g(w) = G'(w) = \frac{\lambda^k w^{k-1} e^{-\lambda w}}{\Gamma(k)}, \qquad 0 < w < \infty,$$
$$= 0 \qquad \text{elsewhere.}$$

That is, W has a gamma distribution with $\alpha = k$ and $\beta = 1/\lambda$. If W is the waiting time until the first change, that is, if $k = 1$, the p.d.f. of W is

$$g(w) = \lambda e^{-\lambda w}, \qquad 0 < w < \infty,$$
$$= 0 \qquad \text{elsewhere,}$$

and W is said to have an *exponential distribution* with mean $\beta = 1/\lambda$.

We now find the m.g.f. of a gamma distribution. Since

$$M(t) = \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}\, dx$$
$$= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x(1-\beta t)/\beta}\, dx,$$

we may set $y = x(1 - \beta t)/\beta$, $t < 1/\beta$, or $x = \beta y/(1 - \beta t)$, to obtain

$$M(t) = \int_0^\infty \frac{\beta/(1-\beta t)}{\Gamma(\alpha)\beta^\alpha} \left(\frac{\beta y}{1-\beta t}\right)^{\alpha-1} e^{-y}\, dy.$$

That is,

$$M(t) = \left(\frac{1}{1-\beta t}\right)^\alpha \int_0^\infty \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy$$
$$= \frac{1}{(1-\beta t)^\alpha}, \qquad t < \frac{1}{\beta}.$$

Now

$$M'(t) = (-\alpha)(1 - \beta t)^{-\alpha-1}(-\beta)$$

and

$$M''(t) = (-\alpha)(-\alpha - 1)(1 - \beta t)^{-\alpha-2}(-\beta)^2.$$

Hence, for a gamma distribution, we have

$$\mu = M'(0) = \alpha\beta$$

and

$$\sigma^2 = M''(0) - \mu^2 = \alpha(\alpha + 1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

Example 1. Let the waiting time W have a gamma p.d.f. with $\alpha = k$ and $\beta = 1/\lambda$. Accordingly, $E(W) = k/\lambda$. If $k = 1$, then $E(W) = 1/\lambda$; that is, the expected waiting time for $k = 1$ changes is equal to the reciprocal of λ.

Example 2. Let X be a random variable such that

$$E(X^m) = \frac{(m+3)!}{3!}3^m, \qquad m = 1, 2, 3, \ldots .$$

Then the m.g.f. of X is given by the series

$$M(t) = 1 + \frac{4!\,3}{3!\,1!}t + \frac{5!\,3^2}{3!\,2!}t^2 + \frac{6!\,3^3}{3!\,3!}t^3 + \cdots .$$

This, however, is the Maclaurin's series for $(1 - 3t)^{-4}$, provided that $-1 < 3t < 1$. Accordingly, X has a gamma distribution with $\alpha = 4$ and $\beta = 3$.

Remark. The gamma distribution is not only a good model for waiting times, but one for many nonnegative random variables of the continuous type. For illustration, the distribution of certain incomes could be modeled satisfactorily by the gamma distribution, since the two parameters α and β provide a great deal of flexibility. Several gamma probability density functions are depicted in Figure 3.1.

Let us now consider the special case of the gamma distribution in which $\alpha = r/2$, where r is a positive integer, and $\beta = 2$. A random variable X of the continuous type that has the p.d.f.

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(r/2)2^{r/2}}x^{r/2-1}e^{-x/2}, \qquad 0 < x < \infty, \\ &= 0 \qquad \text{elsewhere,} \end{aligned}$$

and the m.g.f.

$$M(t) = (1 - 2t)^{-r/2}, \qquad t < \tfrac{1}{2},$$

is said to have a *chi-square distribution*, and any $f(x)$ of this form is called a *chi-square p.d.f.* The mean and the variance of a chi-square distribution are $\mu = \alpha\beta = (r/2)2 = r$ and $\sigma^2 = \alpha\beta^2 = (r/2)2^2 = 2r$, respectively. For no obvious reason, we call the parameter r the number of degrees of freedom of the chi-square distribution (or of the chi-square p.d.f.). Because the chi-square distribution has an important role in statistics and occurs so frequently, we write, for brevity, that X is $\chi^2(r)$ to mean that the random variable X has a chi-square distribution with r degrees of freedom.

Example 3. If X has the p.d.f.

$$f(x) = \tfrac{1}{4}xe^{-x/2}, \qquad 0 < x < \infty,$$
$$= 0 \quad \text{elsewhere,}$$

then X is $\chi^2(4)$. Hence $\mu = 4$, $\sigma^2 = 8$, and $M(t) = (1 - 2t)^{-2}$, $t < \frac{1}{2}$.

Example 4. If X has the m.g.f. $M(t) = (1 - 2t)^{-8}$, $t < \frac{1}{2}$, then X is $\chi^2(16)$.

If the random variable X is $\chi^2(r)$, then, with $c_1 < c_2$, we have

$$\Pr(c_1 \le X \le c_2) = \Pr(X \le c_2) - \Pr(X \le c_1),$$

since $\Pr(X = c_1) = 0$. To compute such a probability, we need the value of an integral like

$$\Pr(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1}e^{-w/2}\,dw.$$

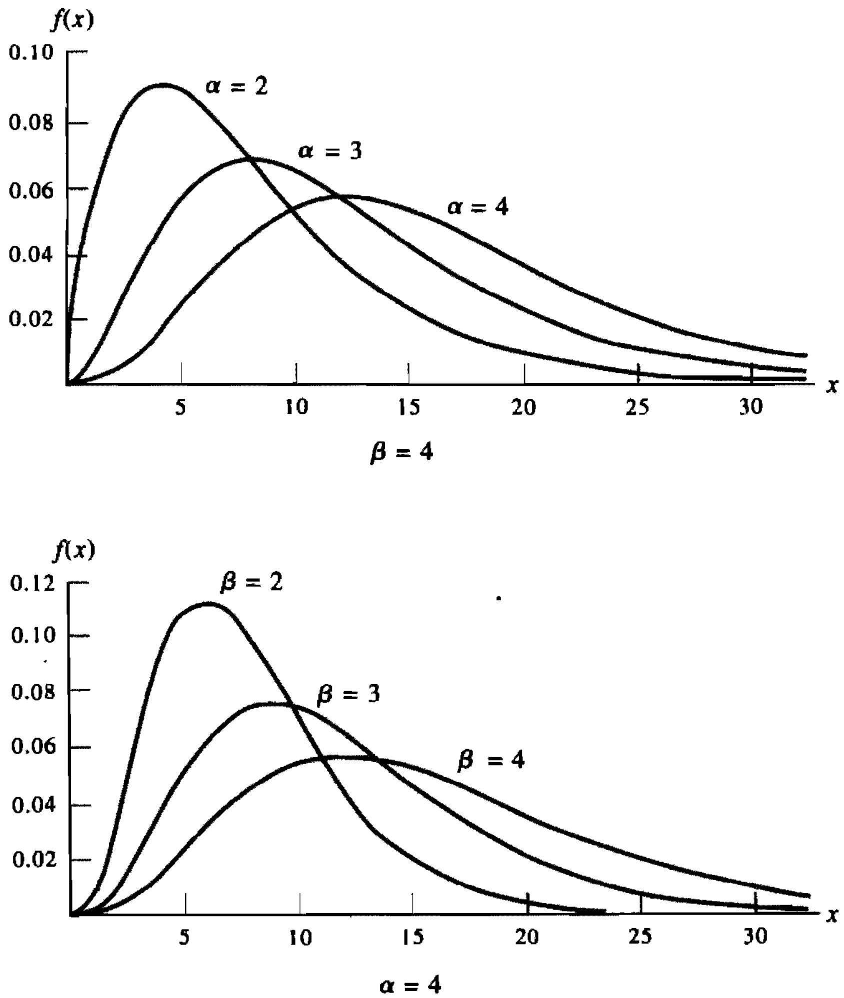

FIGURE 3.1

Tables of this integral for selected values of r and x have been prepared and are partially reproduced in Table II in Appendix B.

Example 5. Let X be $\chi^2(10)$. Then, by Table II of Appendix B, with $r = 10$,

$$\begin{aligned} \Pr(3.25 \le X \le 20.5) &= \Pr(X \le 20.5) - \Pr(X \le 3.25) \\ &= 0.975 - 0.025 = 0.95. \end{aligned}$$

Again, by way of example, if $\Pr(a < X) = 0.05$, then $\Pr(X \le a) = 0.95$, and thus $a = 18.3$ from Table II with $r = 10$.

Example 6. Let X have a gamma distribution with $\alpha = r/2$, where r is a positive integer, and $\beta > 0$. Define the random variable $Y = 2X/\beta$. We seek the p.d.f. of Y. Now the distribution function of Y is

$$G(y) = \Pr(Y \le y) = \Pr\left(X \le \frac{\beta y}{2}\right).$$

If $y \le 0$, then $G(y) = 0$; but if $y > 0$, then

$$G(y) = \int_0^{\beta y/2} \frac{1}{\Gamma(r/2)\beta^{r/2}} x^{r/2 - 1} e^{-x/\beta}\, dx.$$

Accordingly, the p.d.f. of Y is

$$\begin{aligned} g(y) = G'(y) &= \frac{\beta/2}{\Gamma(r/2)\beta^{r/2}} (\beta y/2)^{r/2 - 1} e^{-y/2} \\ &= \frac{1}{\Gamma(r/2)2^{r/2}} y^{r/2 - 1} e^{-y/2} \end{aligned}$$

if $y > 0$. That is, Y is $\chi^2(r)$.

EXERCISES

3.34. If $(1 - 2t)^{-6}$, $t < \frac{1}{2}$, is the m.g.f. of the random variable X, find $\Pr(X < 5.23)$.

3.35. If X is $\chi^2(5)$, determine the constants c and d so that $\Pr(c < X < d) = 0.95$ and $\Pr(X < c) = 0.025$.

3.36. If X has a gamma distribution with $\alpha = 3$ and $\beta = 4$, find $\Pr(3.28 < X < 25.2)$.

Hint: Consider the probability of the equivalent event $1.64 < Y < 12.6$, where $Y = 2X/4 = X/2$.

3.37. Let X be a random variable such that $E(X^m) = (m + 1)!\, 2^m$, $m = 1, 2, 3, \ldots$. Determine the m.g.f. and the distribution of X.

3.38. Show that

$$\int_\mu^\infty \frac{1}{\Gamma(k)} z^{k-1}e^{-z}\,dz = \sum_{x=0}^{k-1} \frac{\mu^x e^{-\mu}}{x!}, \qquad k = 1, 2, 3, \ldots .$$

This demonstrates the relationship between the distribution functions of the gamma and Poisson distributions.

Hint: Either integrate by parts $k - 1$ times or simply note that the "antiderivative" of $z^{k-1}e^{-z}$ is

$$-z^{k-1}e^{-z} - (k-1)z^{k-2}e^{-z} - \cdots - (k-1)!\,e^{-z}$$

by differentiating the latter expression.

3.39. Let X_1, X_2, and X_3 be independent random variables, each with p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Find the distribution of $Y = \text{minimum}\ (X_1, X_2, X_3)$.

Hint: $\Pr(Y \le y) = 1 - \Pr(Y > y) = 1 - \Pr(X_i > y, i = 1, 2, 3)$.

3.40. Let X have a gamma distribution with p.d.f.

$$f(x) = \frac{1}{\beta^2} x e^{-x/\beta}, \qquad 0 < x < \infty,$$

zero elsewhere. If $x = 2$ is the unique mode of the distribution, find the parameter β and $\Pr(X < 9.49)$.

3.41. Compute the measures of skewness and kurtosis of a gamma distribution with parameters α and β.

3.42. Let X have a gamma distribution with parameters α and β. Show that $\Pr(X \ge 2\alpha\beta) \le (2/e)^\alpha$.

Hint: Use the result of Exercise 1.115.

3.43. Give a reasonable definition of a chi-square distribution with zero degrees of freedom.

Hint: Work with the m.g.f. of a distribution that is $\chi^2(r)$ and let $r = 0$.

3.44. In the Poisson postulates on page 127, let λ be a nonnegative function of w, say $\lambda(w)$, such that $D_w[g(0, w)] = -\lambda(w)g(0, w)$. Suppose that $\lambda(w) = krw^{r-1}$, $r \ge 1$.

(a) Find $g(0, w)$ noting that $g(0, 0) = 1$.

(b) Let W be the time that is needed to obtain exactly one change. Then find the distribution function of W, namely $G(w) = \Pr(W \le w) = 1 - \Pr(W > w) = 1 - g(0, w)$, $0 \le w$, and then find the p.d.f. of W. This p.d.f. is that of the *Weibull distribution*, which is used in the study of breaking strengths of materials.

3.45. Let X have a Poisson distribution with parameter m. If m is an experimental value of a random variable having a gamma distribution with $\alpha = 2$ and $\beta = 1$, compute $\Pr(X = 0, 1, 2)$.

3.46. Let X have the uniform distribution with p.d.f. $f(x) = 1, 0 < x < 1$, zero elsewhere. Find the distribution function of $Y = -2 \ln X$. What is the p.d.f. of Y?

3.47. Find the uniform distribution of the continuous type that has the same mean and the same variance as those of a chi-square distribution with 8 degrees of freedom.

3.4 The Normal Distribution

Consider the integral

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

This integral exists because the integrand is a positive continuous function which is bounded by an integrable function; that is,

$$0 < \exp\left(-\frac{y^2}{2}\right) < \exp(-|y| + 1), \qquad -\infty < y < \infty,$$

and

$$\int_{-\infty}^{\infty} \exp(-|y| + 1)\, dy = 2e.$$

To evaluate the integral I, we note that $I > 0$ and that I^2 may be written

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2 + z^2}{2}\right) dy\, dz.$$

This iterated integral can be evaluated by changing to polar coordinates. If we set $y = r\cos\theta$ and $z = r\sin\theta$, we have

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r\, dr\, d\theta$$

$$= \int_0^{2\pi} d\theta = 2\pi.$$

Accordingly, $I = \sqrt{2\pi}$ and

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = 1.$$

If we introduce a new variable of integration, say x, by writing

$$y = \frac{x - a}{b}, \qquad b > 0,$$

the preceding integral becomes

$$\int_{-\infty}^{\infty} \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2b^2}\right] dx = 1.$$

Since $b > 0$, this implies that

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

satisfies the conditions of being a p.d.f. of a continuous type of random variable. A random variable of the continuous type that has a p.d.f. of the form of $f(x)$ is said to have a *normal distribution*, and any $f(x)$ of this form is called a normal p.d.f.

We can find the m.g.f. of a normal distribution as follows. In

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2b^2}\right] dx \\ &= \int_{-\infty}^{\infty} \frac{1}{b\sqrt{2\pi}} \exp\left(-\frac{-2b^2tx + x^2 - 2ax + a^2}{2b^2}\right) dx \end{aligned}$$

we complete the square in the exponent. Thus $M(t)$ becomes

$$\begin{aligned} M(t) &= \exp\left[-\frac{a^2 - (a + b^2t)^2}{2b^2}\right] \int_{-\infty}^{\infty} \frac{1}{b\sqrt{2\pi}} \\ &\qquad \times \exp\left[-\frac{(x - a - b^2t)^2}{2b^2}\right] dx \\ &= \exp\left(at + \frac{b^2t^2}{2}\right) \end{aligned}$$

because the integrand of the last integral can be thought of as a normal p.d.f. with a replaced by $a + b^2t$, and hence it is equal to 1.

The mean μ and variance σ^2 of a normal distribution will be calculated from $M(t)$. Now

$$M'(t) = M(t)(a + b^2t)$$

and

$$M''(t) = M(t)(b^2) + M(t)(a + b^2t)^2.$$

Thus

$$\mu = M'(0) = a$$

and

$$\sigma^2 = M''(0) - \mu^2 = b^2 + a^2 - a^2 = b^2.$$

This permits us to write a normal p.d.f. in the form of

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

a form that shows explicitly the values of μ and σ^2. The m.g.f. $M(t)$ can be written

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Example 1. If X has the m.g.f.

$$M(t) = e^{2t + 32t^2},$$

then X has a normal distribution with $\mu = 2$, $\sigma^2 = 64$.

The normal p.d.f. occurs so frequently in certain parts of statistics that we denote it, for brevity, by $N(\mu, \sigma^2)$. Thus, if we say that the random variable X is $N(0, 1)$, we mean that X has a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, so that the p.d.f. of X is

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

If we say that X is $N(5, 4)$, we mean that X has a normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 4$, so that the p.d.f. of X is

$$f(x) = \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{(x-5)^2}{2(4)}\right], \qquad -\infty < x < \infty.$$

Moreover, if

$$M(t) = e^{t^2/2},$$

then X is $N(0, 1)$.

The graph of

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

is seen (1) to be symmetric about a vertical axis through $x = \mu$, (2) to have its maximum of $1/(\sigma\sqrt{2\pi})$ at $x = \mu$, and (3) to have the x-axis as a horizontal asymptote. It should also be verified that (4) there are points of inflection at $x = \mu \pm \sigma$.

Remark. Each of the special distributions considered thus far has been "justified" by some derivation that is based upon certain concepts found in elementary probability theory. Such a motivation for the normal distribution is not given at this time; a motivation is presented in Chapter 5. However, the normal distribution is one of the more widely used distributions in applications of statistical methods. Variables that are often assumed to be random variables having normal distributions (with appropriate values of μ and σ) are the diameter of a hole made by a drill press, the score on a test, the yield of a grain on a plot of ground, and the length of a newborn child.

We now prove a very useful theorem.

Theorem 1. *If the random variable* X *is* $N(\mu, \sigma^2)$, $\sigma^2 > 0$, *then the random variable* $W = (X - \mu)/\sigma$ *is* $N(0, 1)$.

Proof. The distribution function $G(w)$ of W is, since $\sigma > 0$,

$$G(w) = \Pr\left(\frac{X-\mu}{\sigma} \le w\right) = \Pr(X \le w\sigma + \mu).$$

That is,

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

If we change the variable of integration by writing $y = (x - \mu)/\sigma$, then

$$G(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

Accordingly, the p.d.f. $g(w) = G'(w)$ of the continuous-type random variable W is

$$g(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2}, \qquad -\infty < w < \infty.$$

Thus W is $N(0, 1)$, which is the desired result (see also Exercise 3.100).

This fact considerably simplifies the calculations of probabilities concerning normally distributed variables, as will be seen presently.

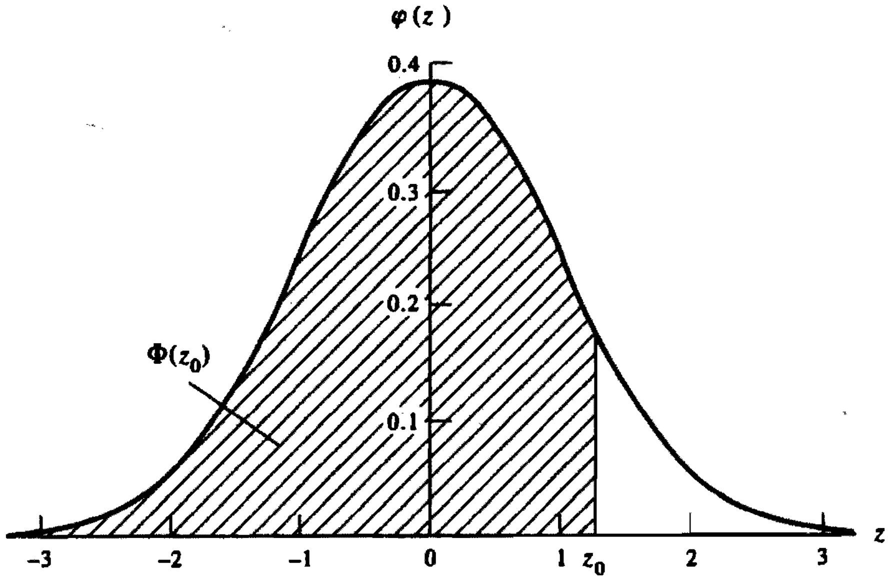

FIGURE 3.2

Suppose that X is $N(\mu, \sigma^2)$. Then, with $c_1 < c_2$ we have, since $\Pr(X = c_1) = 0$,

$$\begin{aligned}\Pr(c_1 < X < c_2) &= \Pr(X < c_2) - \Pr(X < c_1)\\ &= \Pr\left(\frac{X-\mu}{\sigma} < \frac{c_2-\mu}{\sigma}\right) - \Pr\left(\frac{X-\mu}{\sigma} < \frac{c_1-\mu}{\sigma}\right)\\ &= \int_{-\infty}^{(c_2-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\,dw - \int_{-\infty}^{(c_1-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\,dw\end{aligned}$$

because $W = (X - \mu)/\sigma$ is $N(0, 1)$. That is, probabilities concerning X, which is $N(\mu, \sigma^2)$, can be expressed in terms of probabilities concerning W, which is $N(0, 1)$.

An integral such as

$$\int_{-\infty}^{k} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\,dw$$

cannot be evaluated by the fundamental theorem of calculus because an "antiderivative" of $e^{-w^2/2}$ is not expressible as an elementary function. Instead, tables of the approximate value of this integral for various values of k have been prepared and are partially reproduced in Table III in Appendix B. We use the notation

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\,dw.$$

Moreover, we say that $\Phi(z)$ and its derivative $\Phi'(z) = \varphi(z)$ are, respectively, the distribution function and p.d.f. of a *standard normal distribution* $N(0, 1)$. These are depicted in Figure 3.2.

To summarize, we have shown that if X is $N(\mu, \sigma^2)$, then

$$\Pr(c_1 < X < c_2) = \Pr\left(\frac{X-\mu}{\sigma} < \frac{c_2-\mu}{\sigma}\right) - \Pr\left(\frac{X-\mu}{\sigma} < \frac{c_1-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{c_2-\mu}{\sigma}\right) - \Phi\left(\frac{c_1-\mu}{\sigma}\right).$$

It is left as an exercise to show that $\Phi(-x) = 1 - \Phi(x)$.

Example 2. Let X be $N(2, 25)$. Then, by Table III,

$$\Pr(0 < X < 10) = \Phi\left(\frac{10-2}{5}\right) - \Phi\left(\frac{0-2}{5}\right)$$
$$= \Phi(1.6) - \Phi(-0.4)$$
$$= 0.945 - (1 - 0.655) = 0.600$$

and

$$\Pr(-8 < X < 1) = \Phi\left(\frac{1-2}{5}\right) - \Phi\left(\frac{-8-2}{5}\right)$$
$$= \Phi(-0.2) - \Phi(-2)$$
$$= (1 - 0.579) - (1 - 0.977) = 0.398.$$

Example 3. Let X be $N(\mu, \sigma^2)$. Then, by Table III,

$$\Pr(\mu - 2\sigma < X < \mu + 2\sigma) = \Phi\left(\frac{\mu + 2\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right)$$
$$= \Phi(2) - \Phi(-2)$$
$$= 0.977 - (1 - 0.977) = 0.954.$$

Example 4. Suppose that 10 percent of the probability for a certain distribution that is $N(\mu, \sigma^2)$ is below 60 and that 5 percent is above 90. What are the values of μ and σ? We are given that the random variable X is $N(\mu, \sigma^2)$ and that $\Pr(X \le 60) = 0.10$ and $\Pr(X \le 90) = 0.95$. Thus $\Phi[(60 - \mu)/\sigma] = 0.10$ and $\Phi[(90 - \mu)/\sigma] = 0.95$. From Table III we have

$$\frac{60-\mu}{\sigma} = -1.282, \qquad \frac{90-\mu}{\sigma} = 1.645.$$

These conditions require that $\mu = 73.1$ and $\sigma = 10.2$ approximately.

Remark. In this chapter we have illustrated three types of *parameters* associated with distributions. The mean μ of $N(\mu, \sigma^2)$ is called a *location* *parameter* because changing its value simply changes the location of the middle of the normal p.d.f.; that is, the graph of the p.d.f. looks exactly the same except for a shift in location. The *standard deviation* σ of $N(\mu, \sigma^2)$ is called a *scale parameter* because changing its value changes the spread of the distribution. That is, a small value of σ requires the graph of the normal p.d.f. to be tall and narrow, while a large value of σ requires it to spread out and not be so tall. No matter what the values of μ and σ, however, the graph of the normal p.d.f. will be that familiar "bell shape." Incidentally, the β of the gamma distribution is also a scale parameter. On the other hand, the α of the gamma distribution is called a *shape parameter*, as changing its value modifies the shape of the graph of the p.d.f. as can be seen by referring to Figure 3.1. The parameters p and μ of the binomial and Poisson distributions, respectively, are also shape parameters.

We close this section with an important theorem.

Theorem 2. *If the random variable* X is $N(\mu, \sigma^2)$, $\sigma^2 > 0$, *then the random variable* $V = (X - \mu)^2/\sigma^2$ *is* $\chi^2(1)$.

Proof. Because $V = W^2$, where $W = (X - \mu)/\sigma$ is $N(0, 1)$, the distribution function $G(v)$ of V is, for $v \ge 0$,

$$G(v) = \Pr(W^2 \le v) = \Pr(-\sqrt{v} \le W \le \sqrt{v}).$$

That is,

$$G(v) = 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\,dw, \qquad 0 \le v,$$

and

$$G(v) = 0, \qquad v < 0.$$

If we change the variable of integration by writing $w = \sqrt{y}$, then

$$G(v) = \int_0^v \frac{1}{\sqrt{2\pi}\sqrt{y}} e^{-y/2}\,dy, \qquad 0 \le v.$$

Hence the p.d.f. $g(v) = G'(v)$ of the continuous-type random variable V is

$$g(v) = \frac{1}{\sqrt{\pi}\sqrt{2}} v^{1/2-1} e^{-v/2}, \qquad 0 < v < \infty,$$

$$= 0 \qquad \text{elsewhere.}$$

Since $g(v)$ is a p.d.f. and hence

$$\int_0^\infty g(v)\,dv = 1,$$

it must be that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and thus V is $\chi^2(1)$.

EXERCISES

3.48. If

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw,$$

show that $\Phi(-z) = 1 - \Phi(z)$.

3.49. If X is $N(75, 100)$, find $\Pr(X < 60)$ and $\Pr(70 < X < 100)$.

3.50. If X is $N(\mu, \sigma^2)$, find b so that $\Pr[-b < (X - \mu)/\sigma < b] = 0.90$.

3.51. Let X be $N(\mu, \sigma^2)$ so that $\Pr(X < 89) = 0.90$ and $\Pr(X < 94) = 0.95$. Find μ and σ^2.

3.52. Show that the constant c can be selected so that $f(x) = c2^{-x^2}$, $-\infty < x < \infty$, satisfies the conditions of a normal p.d.f.

Hint: Write $2 = e^{\ln 2}$.

3.53. If X is $N(\mu, \sigma^2)$, show that $E(|X - \mu|) = \sigma\sqrt{2/\pi}$.

3.54. Show that the graph of a p.d.f. $N(\mu, \sigma^2)$ has points of inflection at $x = \mu - \sigma$ and $x = \mu + \sigma$.

3.55. Evaluate $\int_2^3 \exp[-2(x - 3)^2]\, dx$.

3.56. Determine the ninetieth percentile of the distribution, which is $N(65, 25)$.

3.57. If $e^{3t + 8t^2}$ is the m.g.f. of the random variable X, find $\Pr(-1 < X < 9)$.

3.58. Let the random variable X have the p.d.f.

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \qquad 0 < x < \infty, \quad \text{zero elsewhere.}$$

Find the mean and variance of X.

Hint: Compute $E(X)$ directly and $E(X^2)$ by comparing that integral with the integral representing the variance of a variable that is $N(0, 1)$.

3.59. Let X be $N(5, 10)$. Find $\Pr[0.04 < (X - 5)^2 < 38.4]$.

3.60. If X is $N(1, 4)$, compute the probability $\Pr(1 < X^2 < 9)$.

3.61. If X is $N(75, 25)$, find the conditional probability that X is greater than 80 relative to the hypothesis that X is greater than 77. See Exercise 2.18.

3.62. Let X be a random variable such that $E(X^{2m}) = (2m)!/(2^m m!)$, $m = 1, 2, 3, \ldots$ and $E(X^{2m-1}) = 0$, $m = 1, 2, 3, \ldots$. Find the m.g.f. and the p.d.f. of X.

3.63. Let the mutually independent random variables X_1, X_2, and X_3 be $N(0, 1)$, $N(2, 4)$, and $N(-1, 1)$, respectively. Compute the probability that exactly two of these three variables are less than zero.

3.64. Compute the measures of skewness and kurtosis of a distribution which is $N(\mu, \sigma^2)$.

3.65. Let the random variable X have a distribution that is $N(\mu, \sigma^2)$.
(a) Does the random variable $Y = X^2$ also have a normal distribution?
(b) Would the random variable $Y = aX + b$, a and b nonzero constants, have a normal distribution?
Hint: In each case, first determine $\Pr(Y \le y)$.

3.66. Let the random variable X be $N(\mu, \sigma^2)$. What would this distribution be if $\sigma^2 = 0$?
Hint: Look at the m.g.f. of X for $\sigma^2 > 0$ and investigate its limit as $\sigma^2 \to 0$.

3.67. Let $\varphi(x)$ and $\Phi(x)$ be the p.d.f. and distribution function of a standard normal distribution. Let Y have a *truncated* distribution with p.d.f. $g(y) = \varphi(y)/[\Phi(b) - \Phi(a)]$, $a < y < b$, zero elsewhere. Show that $E(Y)$ is equal to $[\varphi(a) - \varphi(b)]/[\Phi(b) - \Phi(a)]$.

3.68. Let $f(x)$ and $F(x)$ be the p.d.f. and the distribution function of a distribution of the continuous type such that $f'(x)$ exists for all x. Let the mean of the truncated distribution that has p.d.f. $g(y) = f(y)/F(b)$, $-\infty < y < b$, zero elsewhere, be equal to $-f(b)/F(b)$ for all real b. Prove that $f(x)$ is a p.d.f. of a standard normal distribution.

3.69. Let X and Y be independent random variables, each with a distribution that is $N(0, 1)$. Let $Z = X + Y$. Find the integral that represents the distribution function $G(z) = \Pr(X + Y \le z)$ of Z. Determine the p.d.f. of Z.
Hint: We have that $G(z) = \int_{-\infty}^{\infty} H(x, z)\, dx$, where

$$H(x, z) = \int_{-\infty}^{z-x} \frac{1}{2\pi} \exp\left[-(x^2 + y^2)/2\right] dy.$$

Find $G'(z)$ by evaluating $\int_{-\infty}^{\infty} [\partial H(x, z)/\partial z]\, dx$.

3.5 The Bivariate Normal Distribution

Remark. If the reader with an adequate background in matrix algebra so chooses, this section can be omitted at this point and Section 4.10 can be considered later. If this decision is made, only an example in Section 4.7 and a few exercises need be skipped because the bivariate normal distribution would not be known. Many statisticians, however, find it easier to remember the multivariate (including the bivariate) normal p.d.f. and m.g.f. using matrix notation that is used in Section 4.10. Moreover, that section provides an excellent example of a transformation (in particular, an orthogonal one) and a good illustration of the moment-generating function technique; these are two of the major concepts introduced in Chapter 4.

Let us investigate the function

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-q/2}, \qquad -\infty < x < \infty, \qquad -\infty < y < \infty,$$

where, with $\sigma_1 > 0$, $\sigma_2 > 0$, and $-1 < \rho < 1$,

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

At this point we do not know that the constants μ_1, μ_2, σ_1^2, σ_2^2, and ρ are those respective parameters of a distribution. As a matter of fact, we do not know that $f(x, y)$ has the properties of a joint p.d.f. It will be shown that:

1. $f(x, y)$ is a joint p.d.f.
2. X is $N(\mu_1, \sigma_1^2)$ and Y is $N(\mu_2, \sigma_2^2)$.
3. ρ is the correlation coefficient of X and Y.

A joint p.d.f. of this form is called a *bivariate normal p.d.f.*, and the random variables X and Y are said to have a *bivariate normal distribution*.

That the nonnegative function $f(x, y)$ is actually a joint p.d.f. can be seen as follows. Define $f_1(x)$ by

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

Now

$$(1-\rho^2)q = \left[\left(\frac{y-\mu_2}{\sigma_2}\right) - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right]^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2$$

$$= \left(\frac{y-b}{\sigma_2}\right)^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2,$$

where $b = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$. Thus

$$f_1(x) = \frac{\exp[-(x-\mu_1)^2/2\sigma_1^2]}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp\{-(y-b)^2/[2\sigma_2^2(1-\rho^2)]\}}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}}\, dy.$$

For the purpose of integration, the integrand of the integral in this expression for $f_1(x)$ may be considered a normal p.d.f. with mean b and variance $\sigma_2^2(1-\rho^2)$. Thus this integral is equal to 1 and

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right], \qquad -\infty < x < \infty.$$

Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dy\, dx = \int_{-\infty}^{\infty} f_1(x)\, dx = 1,$$

the nonnegative function $f(x, y)$ is a joint p.d.f. of two continuous-type random variables X and Y. Accordingly, the function $f_1(x)$ is the marginal p.d.f. of X, and X is seen to be $N(\mu_1, \sigma_1^2)$. In like manner, we see that Y is $N(\mu_2, \sigma_2^2)$.

Moreover, from the development above, we note that

$$f(x, y) = f_1(x)\left(\frac{1}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}} \exp\left[-\frac{(y-b)^2}{2\sigma_2^2(1-\rho^2)}\right]\right),$$

where $b = \mu_2 + \rho(\sigma_2/\sigma_1)(x-\mu_1)$. Accordingly, the second factor in the right-hand member of the equation above is the conditional p.d.f. of Y, given that $X = x$. That is, the conditional p.d.f. of Y, given $X = x$, is itself normal with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x-\mu_1)$ and variance $\sigma_2^2(1-\rho^2)$. Thus, with a bivariate normal distribution, the conditional mean of Y, given that $X = x$, is linear in x and is given by

$$E(Y|x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

Since the coefficient of x in this linear conditional mean $E(Y|x)$ is $\rho\sigma_2/\sigma_1$, and since σ_1 and σ_2 represent the respective standard deviations, the number ρ is, in fact, the correlation coefficient of X and Y. This follows from the result, established in Section 2.3, that the coefficient of x in a general linear conditional mean $E(Y|x)$ is the product of the correlation coefficient and the ratio σ_2/σ_1.

Although the mean of the conditional distribution of Y, given $X = x$, depends upon x (unless $\rho = 0$), the variance $\sigma_2^2(1-\rho^2)$ is the same for all real values of x. Thus, by way of example, given that $X = x$, the conditional probability that Y is within $(2.576)\sigma_2\sqrt{1-\rho^2}$ units of the conditional mean is 0.99, whatever the value of x may be. In this sense, most of the probability for the distribution of X and Y lies in the band

$$\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \pm (2.576)\sigma_2\sqrt{1 - \rho^2}$$

about the graph of the linear conditional mean. For every fixed positive σ_2, the width of this band depends upon ρ. Because the band is narrow when ρ^2 is nearly 1, we see that ρ does measure the intensity of the concentration of the probability for X and Y about the linear conditional mean. This is the fact to which we alluded in the remark of Section 2.3.

In a similar manner we can show that the conditional distribution of X, given $Y = y$, is the normal distribution

$$N\left[\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right].$$

Example 1. Let us assume that in a certain population of married couples the height X_1 of the husband and the height X_2 of the wife have a bivariate normal distribution with parameters $\mu_1 = 5.8$ feet, $\mu_2 = 5.3$ feet, $\sigma_1 = \sigma_2 = 0.2$ foot, and $\rho = 0.6$. The conditional p.d.f. of X_2, given $X_1 = 6.3$, is normal with mean $5.3 + (0.6)(6.3 - 5.8) = 5.6$ and standard deviation $(0.2)\sqrt{(1 - 0.36)} = 0.16$. Accordingly, given that the height of the husband is 6.3 feet, the probability that his wife has a height between 5.28 and 5.92 feet is

$$\Pr(5.28 < X_2 < 5.92|X_1 = 6.3) = \Phi(2) - \Phi(-2) = 0.954.$$

The interval (5.28, 5.92) could be thought of as a 95.4 percent *prediction interval* for the wife's height, given $X_1 = 6.3$.

The m.g.f. of a bivariate normal distribution can be determined as follows. We have

$$M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x + t_2y}f(x, y)\; dx\, dy$$

$$= \int_{-\infty}^{\infty} e^{t_1x}f_1(x)\left[\int_{-\infty}^{\infty} e^{t_2y}f_{2|1}(y|x)\, dy\right] dx$$

for all real values of t_1 and t_2. The integral within the brackets is the m.g.f. of the conditional p.d.f. $f_{2|1}(y|x)$. Since $f_{2|1}(y|x)$ is a normal p.d.f. with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$ and variance $\sigma_2^2(1 - \rho^2)$, then

$$\int_{-\infty}^{\infty} e^{t_2y}f_{2|1}(y|x)\, dy = \exp\left\{t_2\left[\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right] + \frac{t_2^2\sigma_2^2(1 - \rho^2)}{2}\right\}.$$

Accordingly, $M(t_1, t_2)$ can be written in the form

$$\exp\left\{t_2\mu_2 - t_2\rho\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{t_2^2\sigma_2^2(1-\rho^2)}{2}\right\}\int_{-\infty}^{\infty}\exp\left[\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right)x\right]f_1(x)\,dx.$$

But $E(e^{tX}) = \exp[\mu_1 t + (\sigma_1^2 t^2)/2]$ for all real values of t. Accordingly, if we set $t = t_1 + t_2\rho(\sigma_2/\sigma_1)$, we see that $M(t_1, t_2)$ is given by

$$\exp\left\{t_2\mu_2 - t_2\rho\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{t_2^2\sigma_2^2(1-\rho^2)}{2} + \mu_1\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right) + \sigma_1^2\frac{\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right)^2}{2}\right\}$$

or, equivalently,

$$M(t_1, t_2) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}\right).$$

It is interesting to note that if, in this m.g.f. $M(t_1, t_2)$, the correlation coefficient ρ is set equal to zero, then

$$M(t_1, t_2) = M(t_1, 0)M(0, t_2).$$

Thus X and Y are independent when $\rho = 0$. If, conversely,

$$M(t_1, t_2) \equiv M(t_1, 0)M(0, t_2),$$

we have $e^{\rho\sigma_1\sigma_2 t_1 t_2} = 1$. Since each of σ_1 and σ_2 is positive, then $\rho = 0$. Accordingly, we have the following theorem.

Theorem 3. *Let X and Y have a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. Then X and Y are independent if and only if $\rho = 0$.*

As a matter of fact, if any two random variables are independent and have positive standard deviations, we have noted in Example 4 of Section 2.4 that $\rho = 0$. However, $\rho = 0$ does not in general imply that two variables are independent; this can be seen in Exercises 2.20 (c) and 2.25. The importance of Theorem 3 lies in the fact that we now know when and only when two random variables that have a bivariate normal distribution are independent.

EXERCISES

3.70. Let X and Y have a bivariate normal distribution with respective parameters $\mu_X = 2.8$, $\mu_Y = 110$, $\sigma_X^2 = 0.16$, $\sigma_Y^2 = 100$, and $\rho = 0.6$. Compute:
(a) $\Pr(106 < Y < 124)$.
(b) $\Pr(106 < Y < 124 | X = 3.2)$.

3.71. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 3$, $\mu_2 = 1$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, and $\rho = \frac{3}{5}$. Determine the following probabilities:
(a) $\Pr(3 < Y < 8)$.
(b) $\Pr(3 < Y < 8 | X = 7)$.
(c) $\Pr(-3 < X < 3)$.
(d) $\Pr(-3 < X < 3 | Y = -4)$.

3.72. If $M(t_1, t_2)$ is the m.g.f. of a bivariate normal distribution, compute the covariance by using the formula

$$\frac{\partial^2 M(0, 0)}{\partial t_1\, \partial t_2} - \frac{\partial M(0, 0)}{\partial t_1} \frac{\partial M(0, 0)}{\partial t_2}.$$

Now let $\psi(t_1, t_2) = \ln M(t_1, t_2)$. Show that $\partial^2 \psi(0, 0)/\partial t_1\, \partial t_2$ gives this covariance directly.

3.73. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 5$, $\mu_2 = 10$, $\sigma_1^2 = 1$, $\sigma_2^2 = 25$, and $\rho > 0$. If $\Pr(4 < Y < 16 | X = 5) = 0.954$, determine ρ.

3.74. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1^2 = 9$, $\sigma_2^2 = 4$, and $\rho = 0.6$. Find the shortest interval for which 0.90 is the conditional probability that Y is in this interval, given that $X = 22$.

3.75. Say the correlation coefficient between the heights of husbands and wives is 0.70 and the mean male height is 5 feet 10 inches with standard deviation 2 inches, and the mean female height is 5 feet 4 inches with standard deviation $1\frac{1}{2}$ inches. Assuming a bivariate normal distribution, what is the best guess of the height of a woman whose husband's height is 6 feet? Find a 95 percent prediction interval for her height.

3.76. Let

$$f(x, y) = (1/2\pi) \exp\left[-\tfrac{1}{2}(x^2 + y^2)\right]\left\{1 + xy \exp\left[-\tfrac{1}{2}(x^2 + y^2 - 2)\right]\right\},$$

where $-\infty < x < \infty$, $-\infty < y < \infty$. If $f(x, y)$ is a joint p.d.f., it is not a normal bivariate p.d.f. Show that $f(x, y)$ actually is a joint p.d.f. and that each marginal p.d.f. is normal. Thus the fact that each marginal p.d.f. is normal does not imply that the joint p.d.f. is bivariate normal.

3.77. Let X, Y, and Z have the joint p.d.f.

$$\left(\frac{1}{2\pi}\right)^{3/2} \exp\left(-\frac{x^2+y^2+z^2}{2}\right)\left[1 + xyz \exp\left(-\frac{x^2+y^2+z^2}{2}\right)\right],$$

where $-\infty < x < \infty$, $-\infty < y < \infty$, and $-\infty < z < \infty$. While X, Y, and Z are obviously dependent, show that X, Y, and Z are pairwise independent and that each pair has a bivariate normal distribution.

3.78. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$, and correlation coefficient ρ. Find the distribution of the random variable $Z = aX + bY$ in which a and b are nonzero constants.

Hint: Write $G(z) = \Pr(Z \le z)$ as an iterated integral and compute $G'(z) = g(z)$ by differentiating under the first integral sign and then evaluating the resulting integral by completing the square in the exponent.

ADDITIONAL EXERCISES

3.79. Let X have a binomial distribution with parameters $n = 288$ and $p = \frac{1}{3}$. Use Chebyshev's inequality to determine a lower bound for $\Pr(76 < X < 116)$.

3.80. Let $f(x) = \dfrac{e^{-\mu}\mu^x}{x!}$, $x = 0, 1, 2, \ldots$, zero elsewhere. Find the values of μ so that $x = 1$ is the unique mode; that is, $f(0) < f(1)$ and $f(1) > f(2) > f(3) > \cdots$.

3.81. Let X and Y be two independent binomial variables with parameters $n = 4$, $p = \frac{1}{2}$ and $n = 3$, $p = \frac{2}{3}$, respectively. Determine $\Pr(X - Y = 3)$.

3.82. Let X and Y be two independent binomial variables, both with parameters n and $p = \frac{1}{2}$. Show that

$$\Pr(X - Y = 0) = \frac{(2n)!}{n!\,n!\,(2^{2n})}.$$

3.83. Two people toss a coin five independent times each. Find the probability that they will obtain the same number of heads.

3.84. Color blindness appears in 1 percent of the people in a certain population. How large must a sample with replacement be if the probability of its containing at least one color-blind person is to be at least 0.95? Assume a binomial distribution $b(n, p = 0.01)$ and find n.

3.85. Assume that the number X of hours of sunshine per day in a certain place has a chi-square distribution with 10 degrees of freedom. The profit of a certain outdoor activity depends upon the number of hours of sunshine through the formula

$$\text{profit} = 1000(1 - e^{-X/10}).$$

Find the expected level of the profit.

3.86. Place five similar balls (each either red or blue) in a bowl at random as follows: A coin is flipped 5 independent times and a red ball is placed in the bowl for each head and a blue ball for each tail. The bowl is then taken and two balls are selected at random without replacement. Given that each of those two balls is red, compute the conditional probability that 5 red balls were placed in the bowl at random.

3.87. If a die is rolled four independent times, what is the probability of one four, two fives, and one six, given that at least one six is produced?

3.88. Let the p.d.f. $f(x)$ be positive on, and only on, the integers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, so that $f(x) = [(11 - x)/x]\, f(x - 1)$, $x = 1, 2, 3, \ldots, 10$. Find $f(x)$.

3.89. Let X and Y have a bivariate normal distribution with $\mu_1 = 5$, $\mu_2 = 10$, $\sigma_1^2 = 1$, $\sigma_2^2 = 25$, and $\rho = \frac{4}{5}$. Compute $\Pr(7 < Y < 19 | x = 5)$.

3.90. Say that Jim has three cents and that Bill has seven cents. A coin is tossed ten independent times. For each head that appears, Bill pays Jim two cents, and for each tail that appears, Jim pays Bill one cent. What is the probability that neither person is in debt after the ten trials?

3.91. If $E(X^r) = [(r + 1)!](2^r)$, $r = 1, 2, 3, \ldots$, find the m.g.f. and p.d.f. of X.

3.92. For a biased coin, say that the probability of exactly two heads in three independent tosses is $\frac{4}{9}$. What is the probability of exactly six heads in nine independent tosses of this coin?

3.93. It is discovered that 75 percent of the pages of a certain book contain no errors. If we assume that the number of errors per page follows a Poisson distribution, find the percentage of pages that have exactly one error.

3.94. Let X have a Poisson distribution with double mode at $x = 1$ and $x = 2$. Find $\Pr[X = 0]$.

3.95. Let X and Y be jointly normally distributed with $\mu_X = 20$, $\mu_Y = 40$, $\sigma_X = 3$, $\sigma_Y = 2$, $\rho = 0.6$. Find a symmetric interval about the conditional mean, so that the probability is 0.90 that Y lies in that interval given that X equals 25.

3.96. Let $f(x) = \binom{10}{x} p^x(1-p)^{10-x}$, $x = 0, 1, \ldots, 10$, zero elsewhere. Find the values of p, so that $f(0) \geq f(1) \geq \cdots \geq f(10)$.

3.97. Let $f(x, y)$ be a bivariate normal p.d.f. and let c be a positive constant so that $c < (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1}$. Show that $c = f(x, y)$ defines an ellipse in the xy-plane.

3.98. Let $f_1(x, y)$ and $f_2(x, y)$ be two bivariate normal probability density functions, each having means equal to zero and variances equal to 1. The respective correlation coefficients are ρ and $-\rho$. Consider the joint distribution of X and Y defined by the joint p.d.f. $[f_1(x, y) + f_2(x, y)]/2$. Show that the two marginal distributions are both $N(0, 1)$, X and Y are dependent, and $E(XY) = 0$ and hence the correlation coefficient of X and Y is zero.

3.99. Let X be $N(\mu, \sigma^2)$. Define the random variable $Y = e^X$ and find its p.d.f. by differentiating $G(y) = \Pr(e^X \leq y) = \Pr(X \leq \ln y)$. This is the p.d.f. of a *lognormal distribution*.

3.100. In the proof of Theorem 1 of Section 3.4, we could let

$$G(w) = \Pr(X \leq w\sigma + \mu) = F(w\sigma + \mu),$$

where F and $F' = f$ are the distribution function and p.d.f. of X, respectively. Then, by the chain rule,

$$g(w) = G'(w) = [F'(w\sigma + \mu)]\sigma.$$

Show that the right-hand member is the p.d.f. of a standard normal distribution; thus this provides another proof of Theorem 1.

CHAPTER 4

Distributions of Functions of Random Variables

4.1 Sampling Theory

Let $X_1, X_2, \ldots, X_n$ denote n random variables that have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$. These variables may or may not be independent. Problems such as the following are very interesting in themselves; but more important, their solutions often provide the basis for making statistical inferences. Let Y be a random variable that is defined by a function of $X_1, X_2, \ldots, X_n$, say $Y = u(X_1, X_2, \ldots, X_n)$. Once the p.d.f. $f(x_1, x_2, \ldots, x_n)$ is given, can we find the p.d.f. of Y? In some of the preceding chapters, we have solved a few of these problems. Among them are the following two. If $n = 1$ and if X_1 is $N(\mu, \sigma^2)$, then $Y = (X_1 - \mu)/\sigma$ is $N(0, 1)$. Let n be a positive integer and let the random variables X_i, $i = 1, 2, \ldots, n$, be independent, each having the same p.d.f. $f(x) = p^x(1-p)^{1-x}$, $x = 0, 1$, and zero elsewhere. If $Y = \sum_1^n X_i$, then Y is $b(n, p)$. It should be observed that $Y = u(X_1) = (X_1 - \mu)/\sigma$ is a function of X_1 that depends upon the two parameters of the normal distribution; whereas $Y = u(X_1, X_2, \ldots, X_n) = \sum_1^n X_i$ does not depend upon p, the parameter of the common p.d.f. of the X_i, $i = 1, 2, \ldots, n$. The distinction that we make between these functions is brought out in the following definition.

Definition 1. A function of one or more random variables that does not depend upon any *unknown* parameter is called a *statistic.*

In accordance with this definition, the random variable $Y = \sum_1^n X_i$ discussed above is a statistic. But the random variable $Y = (X_1 - \mu)/\sigma$ is not a statistic unless μ and σ are known numbers. It should be noted that, although a statistic does not depend upon any unknown parameter, the *distribution* of the statistic may very well depend upon unknown parameters.

Remark. We remark, for the benefit of the more advanced reader, that a statistic is usually defined to be a measurable function of the random variables. In this book, however, we wish to minimize the use of measure theoretic terminology, so we have suppressed the modifier "measurable." It is quite clear that a statistic is a random variable. In fact, some probabilists avoid the use of the word "statistic" altogether, and they refer to a measurable function of random variables as a random variable. We decided to use the word "statistic" because the reader will encounter it so frequently in books and journals.

We can motivate the study of the distribution of a statistic in the following way. Let a random variable X be defined on a sample space $\mathscr{C}$ and let the space of X be denoted by $\mathscr{A}$. In many situations confronting us, the distribution of X is not completely known. For instance, we may know the distribution except for the value of an unknown parameter. To obtain more information about this distribution (or the unknown parameter), we shall repeat under identical conditions the random experiment n independent times. Let the random variable X_i be a function of the ith outcome, $i = 1, 2, \ldots, n$. Then we call $X_1, X_2, \ldots, X_n$ the *observations* of a random sample from the distribution under consideration. Suppose that we can define a statistic $Y = u(X_1, X_2, \ldots, X_n)$ whose p.d.f. is found to be $g(y)$. Perhaps this p.d.f. shows that there is a great probability that Y has a value close to the unknown parameter. Once the experiment has been repeated in the manner indicated and we have $X_1 = x_1, \ldots, X_n = x_n$, then $y = u(x_1, x_2, \ldots, x_n)$ is a known number. It is to be hoped that this known number can in some manner be used to elicit information about the unknown parameter. Thus a statistic may prove to be useful.

Remarks. Let the random variable X be defined as the diameter of a hole to be drilled by a certain drill press and let it be assumed that X has a normal distribution. Past experience with many drill presses makes this assumption plausible; but the assumption does not specify the mean μ nor the variance σ^2 of this normal distribution. The only way to obtain information about μ and σ^2 is to have recourse to experimentation. Thus we shall drill a number, say $n = 20$, of these holes whose diameters will be $X_1, X_2, \ldots, X_{20}$. Then $X_1, X_2, \ldots, X_{20}$ is a random sample from the normal distribution under consideration. Once the holes have been drilled and the diameters measured, the 20 numbers may be used, as will be seen later, to elicit information about μ and σ^2.

The term "random sample" is now defined in a more formal manner.

Definition 2. Let $X_1, X_2, \ldots, X_n$ denote n independent random variables, each of which has the same but possibly unknown p.d.f. $f(x)$; that is, the probability density functions of $X_1, X_2, \ldots, X_n$ are, respectively, $f_1(x_1) = f(x_1)$, $f_2(x_2) = f(x_2), \ldots, f_n(x_n) = f(x_n)$, so that the joint p.d.f. is $f(x_1)f(x_2) \cdots f(x_n)$. The random variables $X_1, X_2, \ldots, X_n$ are then said to constitute a *random sample* from a distribution that has p.d.f. $f(x)$. That is, the observations of a random sample are *independent and identically distributed* (often abbreviated i.i.d.).

Later we shall define what we mean by a random sample from a distribution of more than one random variable.

Sometimes it is convenient to refer to a random sample of size n from a given distribution and, as has been remarked, to refer to $X_1, X_2, \ldots, X_n$ as the observations of the random sample. A reexamination of Example 2 of Section 2.5 reveals that we found the p.d.f. of the statistic, which is the maximum of the observations of a random sample of size $n = 3$, from a distribution with p.d.f.

$f(x) = 2x$, $0 < x < 1$, zero elsewhere. In Section 3.1 we found the p.d.f. of the statistic, which is the sum of the observations of a random sample of size n from a distribution that has p.d.f. $f(x) = p^x(1-p)^{1-x}$, $x = 0, 1$, zero elsewhere. This fact was also referred to at the beginning of this section.

In this book, most of the statistics that we shall encounter will be functions of the observations of a random sample from a given distribution. Next, we define two important statistics of this type.

Definition 3. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a given distribution. The statistic

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is called the *mean* of the random sample, and the statistic

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n} = \sum_{i=1}^{n} \frac{X_i^2}{n} - \bar{X}^2$$

is called the *variance* of the random sample.

Remarks. Many writers do not define the variance of a random sample as we have done but, instead, they take $S^2 = \sum_{1}^{n} (X_i - \bar{X})^2/(n-1)$. There are good reasons for doing this. But a certain price has to be paid, as we shall indicate. Let $x_1, x_2, \ldots, x_n$ denote experimental values of the random variable X that has the p.d.f. $f(x)$ and the distribution function $F(x)$. Thus we may look upon $x_1, x_2, \ldots, x_n$ as the experimental values of a random sample of size n from the given distribution. The *distribution of the sample* is then defined to be the distribution obtained by assigning a probability of $1/n$ to each of the points $x_1, x_2, \ldots, x_n$. This is a distribution of the discrete type. The corresponding distribution function will be denoted by $F_n(x)$ and it is a step function. If we let f_x denote the number of sample values that are less than or equal to x, then $F_n(x) = f_x/n$, so that $F_n(x)$ gives the relative frequency of the event $X \le x$ in the set of n observations. The function $F_n(x)$ is often called the "empirical distribution function" and it has a number of uses.

Because the distribution of the sample is a discrete distribution, the mean and the variance have been defined and are, respectively, $\sum_{1}^{n} x_i/n = \bar{x}$ and $\sum_{1}^{n} (x_i - \bar{x})^2/n = s^2$. Thus, if one finds the distribution of the sample and the associated empirical distribution function to be useful concepts, it would seem logically inconsistent to define the variance of a random sample in any way other than we have.

We have also defined $\bar{X}$ and S^2 only for observations that are i.i.d., that is, when $X_1, X_2, \ldots, X_n$ denote a random sample. However, statisticians often use these symbols, $\bar{X}$ and S^2, even if the assumption of independence is dropped. For example, suppose that $X_1, X_2, \ldots, X_n$ were the observations taken at random from a finite collection of numbers *without replacement*. These observations could be thought of as a sample and its mean $\bar{X}$ and variance S^2 computed; yet $X_1, X_2, \ldots, X_n$ are dependent. Moreover, the n observations could simply be some values, not necessarily taken from a distribution, and we could compute the mean $\bar{X}$ and the variance S^2 associated with these n values. If we do these things, however, we must recognize the conditions under which the observations were obtained, and we cannot make the same statements that are associated with the mean and the variance of what we call a random sample.

Random sampling distribution theory means the general problem of finding distributions of functions of the observations of a random sample. Up to this point, the only method, other than direct probabilistic arguments, of finding the distribution of a function of one or more random variables is the *distribution function technique*. That is, if $X_1, X_2, \ldots, X_n$ are random variables, the distribution of $Y = u(X_1, X_2, \ldots, X_n)$ is determined by computing the distribution function of Y,

$$G(y) = \Pr [u(X_1, X_2, \ldots, X_n) \le y].$$

Even in what superficially appears to be a very simple problem, this can be quite tedious. This fact is illustrated in the next paragraph.

Let X_1, X_2, X_3 denote a random sample of size 3 from a standard normal distribution. Let Y denote the statistic that is the sum of the squares of the sample observations. The distribution function of Y is

$$G(y) = \Pr (X_1^2 + X_2^2 + X_3^2 \le y).$$

If $y < 0$, then $G(y) = 0$. However, if $y \ge 0$, then

$$G(y) = \iiint_A \frac{1}{(2\pi)^{3/2}} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \right] dx_1\, dx_2\, dx_3,$$

where A is the set of points (x_1, x_2, x_3) interior to, or on the surface of, a sphere with center at $(0, 0, 0)$ and radius equal to $\sqrt{y}$. This is not a simple integral. We might hope to make progress by changing to spherical coordinates:

$$x_1 = \rho \cos\theta \sin\varphi, \qquad x_2 = \rho \sin\theta \sin\varphi, \qquad x_3 = \rho\cos\varphi,$$

where $\rho \geq 0$, $0 \leq \theta < 2\pi$, $0 \leq \varphi \leq \pi$. Then, for $y \geq 0$,

$$G(y) = \int_0^{\sqrt{y}} \int_0^{2\pi} \int_0^{\pi} \frac{1}{(2\pi)^{3/2}} e^{-\rho^2/2} \rho^2 \sin\varphi \, d\varphi \, d\theta \, d\rho$$

$$= \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{y}} \rho^2 e^{-\rho^2/2} \, d\rho.$$

If we change the variable of integration by setting $\rho = \sqrt{w}$, we have

$$G(y) = \sqrt{\frac{2}{\pi}} \int_0^y \frac{\sqrt{w}}{2} e^{-w/2} \, dw,$$

for $y \geq 0$. Since Y is a random variable of the continuous type, the p.d.f. of Y is $g(y) = G'(y)$. Thus

$$g(y) = \frac{1}{\sqrt{2\pi}} y^{3/2-1} e^{-y/2}, \qquad 0 < y < \infty,$$
$$= 0 \qquad \text{elsewhere.}$$

Because $\Gamma(\frac{3}{2}) = (\frac{1}{2})\Gamma(\frac{1}{2}) = (\frac{1}{2})\sqrt{\pi}$, and thus $\sqrt{2\pi} = \Gamma(\frac{3}{2})2^{3/2}$, we see that Y is $\chi^2(3)$.

The problem that we have just solved highlights the desirability of having, if possible, various methods of determining the distribution of a function of random variables. We shall find that other techniques are available and that often a particular technique is vastly superior to the others in a given situation. These techniques will be discussed in subsequent sections.

Example 1. Let the random variable Y be distributed uniformly over the unit interval $0 < y < 1$; that is, the distribution function of Y is

$$G(y) = 0, \qquad y \leq 0,$$
$$= y, \qquad 0 < y < 1,$$
$$= 1, \qquad 1 \leq y.$$

Suppose that $F(x)$ is a distribution function of the continuous type which is strictly increasing when $0 < F(x) < 1$. If we define the random variable X by the relationship $Y = F(X)$, we now show that X has a distribution which corresponds to $F(x)$. If $0 < F(x) < 1$, the inequalities $X \le x$ and $F(X) \le F(x)$ are equivalent. Thus, with $0 < F(x) < 1$, the distribution of X is

$$\Pr(X \le x) = \Pr[F(X) \le F(x)] = \Pr[Y \le F(x)]$$

because $Y = F(X)$. However, $\Pr(Y \le y) = G(y)$, so we have

$$\Pr(X \le x) = G[F(x)] = F(x), \qquad 0 < F(x) < 1.$$

That is, the distribution function of X is $F(x)$.

This result permits us to *simulate* random variables of different types. This is done by simply determining values of the uniform variable Y, usually with a computer. Then, after determining the observed value $Y = y$, solve the equation $y = F(x)$, either explicitly or by numerical methods. This yields the inverse function $x = F^{-1}(y)$. By the preceding result, this number x will be an observed value of X that has distribution function $F(x)$.

It is also interesting to note that the converse of this result is true. If X has distribution function $F(x)$ of the continuous type, then $Y = F(X)$ is uniformly distributed over $0 < y < 1$. The reason for this is, for $0 < y < 1$, that

$$\Pr(Y \le y) = \Pr[F(X) \le y] = \Pr[X \le F^{-1}(y)].$$

However, it is given that $\Pr(X \le x) = F(x)$, so

$$\Pr(Y \le y) = F[F^{-1}(y)] = y, \qquad 0 < y < 1.$$

This is the distribution function of a random variable that is distributed uniformly on the interval (0, 1).

EXERCISES

4.1. Show that

$$S^2 = \frac{1}{n}\sum_1^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_1^n X_i^2 - \bar{X}^2,$$

where $\bar{X} = \sum_1^n X_i/n$.

4.2. Find the probability that exactly four observations of a random sample of size 5 from the distribution having p.d.f. $f(x) = (x + 1)/2$, $-1 < x < 1$, zero elsewhere, exceed zero.

4.3. Let X_1, X_2, X_3 be a random sample of size 3 from a distribution that is $N(6, 4)$. Determine the probability that the largest sample observation exceeds 8.

4.4. What is the probability that at least one observation of a random sample of size $n = 5$ from a continuous-type distribution exceeds the 90th percentile?

4.5. Let X have the p.d.f. $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere. Show that $Y = -2 \ln X^4$ is $\chi^2(2)$.

4.6. Let X_1, X_2 be a random sample of size $n = 2$ from a distribution with p.d.f. $f(x) = 4x^3, 0 < x < 1$, zero elsewhere. Find the mean and the variance of the ratio $Y = X_1/X_2$.

Hint: First find the distribution function $\Pr(Y \le y)$ when $0 < y < 1$ and then when $1 \le y$.

4.7. Let X_1, X_2 be a random sample from the distribution having p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Find $\Pr(X_1/X_2 \le \frac{1}{2})$ and $\Pr(X_1 X_2 \ge \frac{1}{4})$.

4.8. If the sample size is $n = 2$, find the constant c so that $S^2 = c(X_1 - X_2)^2$.

4.9. If $x_i = i$, $i = 1, 2, \ldots, n$, compute the values of $\bar{x} = \Sigma\, x_i/n$ and $s^2 = \Sigma\,(x_i - \bar{x})^2/n$.

4.10. Let $y_i = a + bx_i$, $i = 1, 2, \ldots, n$, where a and b are constants. Find $\bar{y} = \Sigma\, y_i/n$ and $s_y^2 = \Sigma\,(y_i - \bar{y})^2/n$ in terms of a, b, $\bar{x} = \Sigma\, x_i/n$, and $s_x^2 = \Sigma\,(x_i - \bar{x})^2/n$.

4.11. Let X_1 and X_2 denote two i.i.d. random variables, each from a distribution that is $N(0, 1)$. Find the p.d.f. of $Y = X_1^2 + X_2^2$.

Hint: In the double integral representing $\Pr(Y \le y)$, use polar coordinates.

4.12. The four values $y_1 = 0.42$, $y_2 = 0.31$, $y_3 = 0.87$, and $y_4 = 0.65$ represent the observed values of a random sample of size $n = 4$ from the uniform distribution over $0 < y < 1$. Using these four values, find a corresponding observed random sample from a distribution that has p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere.

4.13. Let X_1, X_2 denote a random sample of size 2 from a distribution with p.d.f. $f(x) = \frac{1}{2}, 0 < x < 2$, zero elsewhere. Find the joint p.d.f. of X_1 and X_2. Let $Y = X_1 + X_2$. Find the distribution function and the p.d.f. of Y.

4.14. Let X_1, X_2 denote a random sample of size 2 from a distribution with p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere. Find the distribution function and the p.d.f. of $Y = X_1/X_2$.

4.15. Let X_1, X_2, X_3 be three i.i.d. random variables, each from a distribution having p.d.f. $f(x) = 5x^4$, $0 < x < 1$, zero elsewhere. Let Y be the largest observation in the sample. Find the distribution function and p.d.f. of Y.

4.16. Let X_1 and X_2 be observations of a random sample from a distribution with p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Evaluate the conditional probability $\Pr(X_1 < X_2 | X_1 < 2X_2)$.

4.2 Transformations of Variables of the Discrete Type

An alternative method of finding the distribution of a function of one or more random variables is called the *change-of-variable technique*. There are some delicate questions (with particular reference to random variables of the continuous type) involved in this technique, and these make it desirable for us first to consider special cases.

Let X have the Poisson p.d.f.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}, \qquad x = 0, 1, 2, \dots,$$
$$= 0 \qquad \text{elsewhere.}$$

As we have done before, let $\mathscr{A}$ denote the space $\mathscr{A} = \{x : x = 0, 1, 2, \dots\}$, so that $\mathscr{A}$ is the set where $f(x) > 0$. Define a new random variable Y by $Y = 4X$. We wish to find the p.d.f. of Y by the change-of-variable technique. Let $y = 4x$. We call $y = 4x$ a transformation from x to y, and we say that the transformation maps the space $\mathscr{A}$ onto the space $\mathscr{B} = \{y : y = 0, 4, 8, 12, \dots\}$. The space $\mathscr{B}$ is obtained by transforming each point in $\mathscr{A}$ in accordance with $y = 4x$. We note two things about this transformation. It is such that to each point in $\mathscr{A}$ there corresponds one, and only one, point in $\mathscr{B}$; and conversely, to each point in $\mathscr{B}$ there corresponds one, and only one, point in $\mathscr{A}$. That is, the transformation $y = 4x$ sets up a one-to-one correspondence between the points of $\mathscr{A}$ and those of $\mathscr{B}$. Any function $y = u(x)$ (not merely $y = 4x$) that maps a space $\mathscr{A}$ (not merely our $\mathscr{A}$) onto a space $\mathscr{B}$ (not merely our $\mathscr{B}$) such that there is a one-to-one correspondence between the points of $\mathscr{A}$ and those of $\mathscr{B}$ is called a *one-to-one transformation*. It is important to note that a one-to-one transformation, $y = u(x)$, implies that x is a single-valued function of y. In our case this is obviously true, since $y = 4x$ requires that $x = (\frac{1}{4})y$.

Our problem is that of finding the p.d.f. $g(y)$ of the discrete type of random variable $Y = 4X$. Now $g(y) = \Pr(Y = y)$. Because there is a one-to-one correspondence between the points of $\mathscr{A}$ and those of

$\mathscr{B}$, the event $Y = y$ or $4X = y$ can occur when, and only when, the event $X = (\frac{1}{4})y$ occurs. That is, the two events are equivalent and have the same probability. Hence

$$g(y) = \Pr(Y = y) = \Pr\left(X = \frac{y}{4}\right) = \frac{\mu^{y/4}e^{-\mu}}{(y/4)!}, \qquad y = 0, 4, 8, \ldots,$$

$$= 0 \qquad \text{elsewhere.}$$

The foregoing detailed discussion should make the subsequent text easier to read. Let X be a random variable of the discrete type, having p.d.f. $f(x)$. Let $\mathscr{A}$ denote the set of discrete points, at each of which $f(x) > 0$, and let $y = u(x)$ define a one-to-one transformation that maps $\mathscr{A}$ onto $\mathscr{B}$. If we solve $y = u(x)$ for x in terms of y, say, $x = w(y)$, then for each $y \in \mathscr{B}$, we have $x = w(y) \in \mathscr{A}$. Consider the random variable $Y = u(X)$. If $y \in \mathscr{B}$, then $x = w(y) \in \mathscr{A}$, and the events $Y = y$ [or $u(X) = y$] and $X = w(y)$ are equivalent. Accordingly, the p.d.f. of Y is

$$g(y) = \Pr(Y = y) = \Pr[X = w(y)] = f[w(y)], \qquad y \in \mathscr{B},$$

$$= 0 \qquad \text{elsewhere.}$$

Example 1. Let X have the binomial p.d.f.

$$f(x) = \frac{3!}{x!(3-x)!}\left(\frac{2}{3}\right)^x\left(\frac{1}{3}\right)^{3-x}, \qquad x = 0, 1, 2, 3,$$

$$= 0 \qquad \text{elsewhere.}$$

We seek the p.d.f. $g(y)$ of the random variable $Y = X^2$. The transformation $y = u(x) = x^2$ maps $\mathscr{A} = \{x : x = 0, 1, 2, 3\}$ onto $\mathscr{B} = \{y : y = 0, 1, 4, 9\}$. In general, $y = x^2$ does not define a one-to-one transformation; here, however, it does, for there are no negative values of x in $\mathscr{A} = \{x : x = 0, 1, 2, 3\}$. That is, we have the single-valued inverse function $x = w(y) = \sqrt{y}$ (not $-\sqrt{y}$), and so

$$g(y) = f(\sqrt{y}) = \frac{3!}{(\sqrt{y})!\,(3-\sqrt{y})!}\left(\frac{2}{3}\right)^{\sqrt{y}}\left(\frac{1}{3}\right)^{3-\sqrt{y}}, \qquad y = 0, 1, 4, 9,$$

$$= 0 \qquad \text{elsewhere.}$$

There are no essential difficulties involved in a problem like the following. Let $f(x_1, x_2)$ be the joint p.d.f. of two discrete-type random variables X_1 and X_2 with $\mathscr{A}$ the (two-dimensional) set of points at which $f(x_1, x_2) > 0$. Let $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define a one-to-one transformation that maps $\mathscr{A}$ onto $\mathscr{B}$. The joint p.d.f. of the two new random variables $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ is given by

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)], \qquad (y_1, y_2) \in \mathscr{B},$$
$$= 0 \qquad \text{elsewhere,}$$

where $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$ is the single-valued inverse of $y_1 = u_1(x_1, x_2)$, $y_2 = u_2(x_1, x_2)$. From this joint p.d.f. $g(y_1, y_2)$ we may obtain the marginal p.d.f. of Y_1 by summing on y_2 or the marginal p.d.f. of Y_2 by summing on y_1.

Perhaps it should be emphasized that the technique of change of variables involves the introduction of as many "new" variables as there were "old" variables. That is, suppose that $f(x_1, x_2, x_3)$ is the joint p.d.f. of X_1, X_2, and X_3, with $\mathscr{A}$ the set where $f(x_1, x_2, x_3) > 0$. Let us say we seek the p.d.f. of $Y_1 = u_1(X_1, X_2, X_3)$. We would then define (if possible) $Y_2 = u_2(X_1, X_2, X_3)$ and $Y_3 = u_3(X_1, X_2, X_3)$, so that $y_1 = u_1(x_1, x_2, x_3)$, $y_2 = u_2(x_1, x_2, x_3)$, $y_3 = u_3(x_1, x_2, x_3)$ define a one-to-one transformation of $\mathscr{A}$ onto $\mathscr{B}$. This would enable us to find the joint p.d.f. of Y_1, Y_2, and Y_3 from which we would get the marginal p.d.f. of Y_1 by summing on y_2 and y_3.

Example 2. Let X_1 and X_2 be two independent random variables that have Poisson distributions with means μ_1 and μ_2, respectively. The joint p.d.f. of X_1 and X_2 is

$$\frac{\mu_1^{x_1}\mu_2^{x_2}e^{-\mu_1-\mu_2}}{x_1!\,x_2!}, \qquad x_1 = 0, 1, 2, 3, \ldots, \qquad x_2 = 0, 1, 2, 3, \ldots,$$

and is zero elsewhere. Thus the space $\mathscr{A}$ is the set of points (x_1, x_2), where each of x_1 and x_2 is a nonnegative integer. We wish to find the p.d.f. of $Y_1 = X_1 + X_2$. If we use the change of variable technique, we need to define a second random variable Y_2. Because Y_2 is of no interest to us, let us choose it in such a way that we have a simple one-to-one transformation. For example, take $Y_2 = X_2$. Then $y_1 = x_1 + x_2$ and $y_2 = x_2$ represent a one-to-one transformation that maps $\mathscr{A}$ onto

$$\mathscr{B} = \{(y_1, y_2) : y_2 = 0, 1, \ldots, y_1 \quad \text{and} \quad y_1 = 0, 1, 2, \ldots\}.$$

Note that, if $(y_1, y_2) \in \mathscr{B}$, then $0 \le y_2 \le y_1$. The inverse functions are given by $x_1 = y_1 - y_2$ and $x_2 = y_2$. Thus the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = \frac{\mu_1^{y_1-y_2}\mu_2^{y_2}e^{-\mu_1-\mu_2}}{(y_1 - y_2)!\,y_2!}, \qquad (y_1, y_2) \in \mathscr{B},$$

and is zero elsewhere. Consequently, the marginal p.d.f. of Y_1 is given by

$$g_1(y_1) = \sum_{y_2=0}^{y_1} g(y_1, y_2)$$

$$= \frac{e^{-\mu_1-\mu_2}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{(y_1-y_2)!\,y_2!} \mu_1^{y_1-y_2}\mu_2^{y_2}$$

$$= \frac{(\mu_1+\mu_2)^{y_1} e^{-\mu_1-\mu_2}}{y_1!}, \qquad y_1 = 0, 1, 2, \ldots,$$

and is zero elsewhere. That is, $Y_1 = X_1 + X_2$ has a Poisson distribution with parameter $\mu_1 + \mu_2$.

Remark. It should be noted that Example 2 essentially illustrates the distribution function technique too. That is, without defining $Y_2 = X_2$, we have that the distribution function of $Y_1 = X_1 + X_2$ is

$$G_1(y_1) = \Pr(X_1 + X_2 \le y_1).$$

In this discrete case, with $y_1 = 0, 1, 2, \ldots$, the p.d.f. of Y_1 is equal to

$$g_1(y_1) = G_1(y_1) - G_1(y_1 - 1) = \Pr(X_1 + X_2 = y_1).$$

That is,

$$g_1(y_1) = \sum\sum_{x_1+x_2=y_1} \frac{\mu_1^{x_1}\mu_2^{x_2}e^{-\mu_1-\mu_2}}{x_1!\,x_2!}.$$

This summation is over all points of $\mathscr{A}$ such that $x_1 + x_2 = y_1$ and thus can be written as

$$g_1(y_1) = \sum_{x_2=0}^{y_1} \frac{\mu_1^{y_1-x_2}\mu_2^{x_2}e^{-\mu_1-\mu_2}}{(y_1-x_2)!\,x_2!},$$

which is exactly the summation given in Example 2.

Example 3. In Section 4.1, we found that we could simulate a continuous-type random variable X with distribution function $F(x)$ through $X = F^{-1}(Y)$, where Y has a uniform distribution on $0 < y < 1$. In a sense, we can simulate a discrete-type random variable X in much the same way, but we must understand what $X = F^{-1}(Y)$ means in this case. Here $F(x)$ is a step function with the height of the step at $x = x_0$ equal to $\Pr(X = x_0)$. For illustration, in Example 3 of Section 1.5, $\Pr(X = 3) = \frac{3}{6}$ is the height of the step at $x = 3$ in Figure 1.3 that depicts the distribution function. If we now think of selecting a random point Y, having the uniform distribution on $0 < y \le 1$, on the vertical axis of Figure 1.3, the probability of falling between $\frac{3}{6}$ and $\frac{6}{6}$ is $\frac{3}{6}$. However, if it falls between those two values, the horizontal line drawn from it would "hit" the step at $x = 3$. That is, for $\frac{3}{6} < y \le \frac{6}{6}$, then $F^{-1}(y) = 3$. Of course, if $\frac{1}{6} < y \le \frac{3}{6}$, then $F^{-1}(y) = 2$; and if $0 < y \le \frac{1}{6}$, we have $F^{-1}(y) = 1$. Thus, with this procedure, we can generate the numbers $x = 1$,

$x = 2$, and $x = 3$ with respective probabilities $\frac{1}{6}$, $\frac{2}{6}$, and $\frac{3}{6}$, as we desired. Clearly, this procedure can be generalized to simulate any random variable X of the discrete type.

EXERCISES

4.17. Let X have a p.d.f. $f(x) = \frac{1}{3}$, $x = 1, 2, 3$, zero elsewhere. Find the p.d.f. of $Y = 2X + 1$.

4.18. If $f(x_1, x_2) = (\frac{2}{3})^{x_1+x_2}(\frac{1}{3})^{2-x_1-x_2}$, $(x_1, x_2) = (0, 0), (0, 1), (1, 0), (1, 1)$, zero elsewhere, is the joint p.d.f. of X_1 and X_2, find the joint p.d.f. of $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$.

4.19. Let X have the p.d.f. $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere. Find the p.d.f. of $Y = X^3$.

4.20. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = x_1x_2/36$, $x_1 = 1, 2, 3$ and $x_2 = 1, 2, 3$, zero elsewhere. Find first the joint p.d.f. of $Y_1 = X_1X_2$ and $Y_2 = X_2$, and then find the marginal p.d.f. of Y_1.

4.21. Let the independent random variables X_1 and X_2 be $b(n_1, p)$ and $b(n_2, p)$, respectively. Find the joint p.d.f. of $Y_1 = X_1 + X_2$ and $Y_2 = X_2$, and then find the marginal p.d.f. of Y_1.

Hint: Use the fact that

$$\sum_{w=0}^{k} \binom{n_1}{w}\binom{n_2}{k-w} = \binom{n_1+n_2}{k}.$$

This can be proved by comparing the coefficients of x^k in each member of the identity $(1 + x)^{n_1}(1 + x)^{n_2} \equiv (1 + x)^{n_1+n_2}$.

4.22. Let X_1 and X_2 be independent random variables of the discrete type with joint p.d.f. $f_1(x_1)f_2(x_2)$, $(x_1, x_2) \in \mathscr{A}$. Let $y_1 = u_1(x_1)$ and $y_2 = u_2(x_2)$ denote a one-to-one transformation that maps $\mathscr{A}$ onto $\mathscr{B}$. Show that $Y_1 = u_1(X_1)$ and $Y_2 = u_2(X_2)$ are independent.

4.23. Consider the random variable X with p.d.f. $f(x) = x/15$, $x = 1, 2, 3, 4, 5$, and zero elsewhere.

(a) Graph the distribution function $F(x)$ of X.

(b) Using a computer or a table of random numbers, determine 30 values of Y, which has the (approximate) uniform distribution on $0 < y < 1$.

(c) From these 30 values of Y, find the corresponding 30 values of X and determine the relative frequencies of $x = 1$, $x = 2$, $x = 3$, $x = 4$, and $x = 5$. How do these compare to the respective probabilities of $\frac{1}{15}$, $\frac{2}{15}$, $\frac{3}{15}$, $\frac{4}{15}$, $\frac{5}{15}$?

4.24. Using the technique given in Example 3 and Exercise 4.23, generate 50 values having a Poisson distribution with $\mu = 1$.

Hint: Use Table I in Appendix B.

4.3 Transformations of Variables of the Continuous Type

In the preceding section we introduced the notion of a one-to-one transformation and the mapping of a set $\mathscr{A}$ onto a set $\mathscr{B}$ under that transformation. Those ideas were sufficient to enable us to find the distribution of a function of several random variables of the discrete type. In this section we examine the same problem when the random variables are of the continuous type. It is again helpful to begin with a special problem.

Example 1. Let X be a random variable of the continuous type, having p.d.f.

$$f(x) = 2x, \qquad 0 < x < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

Here $\mathscr{A}$ is the space $\{x : 0 < x < 1\}$, where $f(x) > 0$. Define the random variable Y by $Y = 8X^3$ and consider the transformation $y = 8x^3$. Under the transformation $y = 8x^3$, the set $\mathscr{A}$ is mapped onto the set $\mathscr{B} = \{y : 0 < y < 8\}$, and, moreover, the transformation is one-to-one. For every $0 < a < b < 8$, the event $a < Y < b$ will occur when, and only when, the event $\frac{1}{2}\sqrt[3]{a} < X < \frac{1}{2}\sqrt[3]{b}$ occurs because there is a one-to-one correspondence between the points of $\mathscr{A}$ and $\mathscr{B}$. Thus

$$\Pr(a < Y < b) = \Pr(\tfrac{1}{2}\sqrt[3]{a} < X < \tfrac{1}{2}\sqrt[3]{b})$$
$$= \int_{\sqrt[3]{a}/2}^{\sqrt[3]{b}/2} 2x\, dx.$$

Let us rewrite this integral by changing the variable of integration from x to y by writing $y = 8x^3$ or $x = \frac{1}{2}\sqrt[3]{y}$. Now

$$\frac{dx}{dy} = \frac{1}{6y^{2/3}},$$

and, accordingly, we have

$$\Pr(a < Y < b) = \int_a^b 2\left(\frac{\sqrt[3]{y}}{2}\right)\left(\frac{1}{6y^{2/3}}\right) dy$$
$$= \int_a^b \frac{1}{6y^{1/3}}\, dy.$$

Since this is true for every $0 < a < b < 8$, the p.d.f. $g(y)$ of Y is the integrand; that is,

$$g(y) = \frac{1}{6y^{1/3}}, \qquad 0 < y < 8,$$
$$= 0 \qquad \text{elsewhere.}$$

It is worth noting that we found the p.d.f. of the random variable $Y = 8X^3$ by using a theorem on the change of variable in a definite integral. However, to obtain $g(y)$ we actually need only two things: (1) the set $\mathscr{B}$ of points y where $g(y) > 0$ and (2) the integrand of the integral on y to which $\Pr(a < Y < b)$ is equal. These can be found by two simple rules:

1. Verify that the transformation $y = 8x^3$ maps $\mathscr{A} = \{x : 0 < x < 1\}$ onto $\mathscr{B} = \{y : 0 < y < 8\}$ and that the transformation is one-to-one.
2. Determine $g(y)$ on this set $\mathscr{B}$ by substituting $\frac{1}{2}\sqrt[3]{y}$ for x in $f(x)$ and then multiplying this result by the derivative of $\frac{1}{2}\sqrt[3]{y}$. That is,

$$g(y) = f\left(\frac{\sqrt[3]{y}}{2}\right)\frac{d[(\frac{1}{2})\sqrt[3]{y}]}{dy} = \frac{1}{6y^{1/3}}, \qquad 0 < y < 8,$$

$$= 0 \qquad \text{elsewhere.}$$

We shall accept a theorem in analysis on the change of variable in a definite integral to enable us to state a more general result. Let X be a random variable of the continuous type having p.d.f. $f(x)$. Let $\mathscr{A}$ be the one-dimensional space where $f(x) > 0$. Consider the random variable $Y = u(X)$, where $y = u(x)$ defines a one-to-one transformation that maps the set $\mathscr{A}$ onto the set $\mathscr{B}$. Let the inverse of $y = u(x)$ be denoted by $x = w(y)$, and let the derivative $dx/dy = w'(y)$ be continuous and not equal zero for all points y in $\mathscr{B}$. Then the p.d.f. of the random variable $Y = u(X)$ is given by

$$g(y) = f[w(y)]|w'(y)|, \qquad y \in \mathscr{B},$$

$$= 0 \qquad \text{elsewhere,}$$

where $|w'(y)|$ represents the absolute value of $w'(y)$. This is precisely what we did in Example 1 of this section, except there we deliberately chose $y = 8x^3$ to be an increasing function so that

$$\frac{dx}{dy} = w'(y) = \frac{1}{6y^{2/3}}, \qquad 0 < y < 8,$$

is positive, and hence

$$\left|\frac{1}{6y^{2/3}}\right| = \frac{1}{6y^{2/3}}, \qquad 0 < y < 8.$$

Henceforth, we shall refer to $dx/dy = w'(y)$ as the Jacobian (denoted by J) of the transformation. In most mathematical areas, $J = w'(y)$ is referred to as the Jacobian of the inverse transformation $x = w(y)$, but in this book it will be called the Jacobian of the transformation, simply for convenience.

Example 2. Let X have the p.d.f.

$$f(x) = 1, \qquad 0 < x < 1,$$
$$= 0 \qquad \text{elsewhere}.$$

We are to show that the random variable $Y = -2 \ln X$ has a chi-square distribution with 2 degrees of freedom. Here the transformation is $y = u(x) = -2 \ln x$, so that $x = w(y) = e^{-y/2}$. The space $\mathscr{A}$ is $\mathscr{A} = \{x : 0 < x < 1\}$, which the one-to-one transformation $y = -2 \ln x$ maps onto $\mathscr{B} = \{y : 0 < y < \infty\}$. The Jacobian of the transformation is

$$J = \frac{dx}{dy} = w'(y) = -\frac{1}{2} e^{-y/2}.$$

Accordingly, the p.d.f. $g(y)$ of $Y = -2 \ln X$ is

$$g(y) = f(e^{-y/2})|J| = \tfrac{1}{2}e^{-y/2}, \qquad 0 < y < \infty,$$
$$= 0 \qquad \text{elsewhere},$$

a p.d.f. that is chi-square with 2 degrees of freedom. Note that this problem was first proposed in Exercise 3.46.

This method of finding the p.d.f. of a function of one random variable of the continuous type will now be extended to functions of two random variables of this type. Again, only functions that define a one-to-one transformation will be considered at this time. Let $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define a one-to-one transformation that maps a (two-dimensional) set $\mathscr{A}$ in the x_1x_2-plane onto a (two-dimensional) set $\mathscr{B}$ in the y_1y_2-plane. If we express each of x_1 and x_2 in terms of y_1 and y_2, we can write $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$. The determinant of order 2,

$$\begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix},$$

is called the *Jacobian* of the transformation and will be denoted by the symbol J. It will be assumed that these first-order partial derivatives are continuous and that the Jacobian J is not identically equal to zero in $\mathcal{B}$. An illustrative example may be desirable before we proceed with the extension of the change of variable technique to two random variables of the continuous type.

Example 3. Let $\mathcal{A}$ be the set $\mathcal{A} = \{(x_1, x_2) : 0 < x_1 < 1, 0 < x_2 < 1\}$ depicted in Figure 4.1. We wish to determine the set $\mathcal{B}$ in the y_1y_2-plane that is the mapping of $\mathcal{A}$ under the one-to-one transformation

$$y_1 = u_1(x_1, x_2) = x_1 + x_2,$$

$$y_2 = u_2(x_1, x_2) = x_1 - x_2,$$

and we wish to compute the Jacobian of the transformation. Now

$$x_1 = w_1(y_1, y_2) = \tfrac{1}{2}(y_1 + y_2),$$

$$x_2 = w_2(y_1, y_2) = \tfrac{1}{2}(y_1 - y_2).$$

To determine the set $\mathcal{B}$ in the y_1y_2-plane onto which $\mathcal{A}$ is mapped under the transformation, note that the boundaries of $\mathcal{A}$ are transformed as follows into the boundaries of $\mathcal{B}$;

$$x_1 = 0 \quad \text{into} \quad 0 = \tfrac{1}{2}(y_1 + y_2),$$

$$x_1 = 1 \quad \text{into} \quad 1 = \tfrac{1}{2}(y_1 + y_2),$$

$$x_2 = 0 \quad \text{into} \quad 0 = \tfrac{1}{2}(y_1 - y_2),$$

$$x_2 = 1 \quad \text{into} \quad 1 = \tfrac{1}{2}(y_1 - y_2).$$

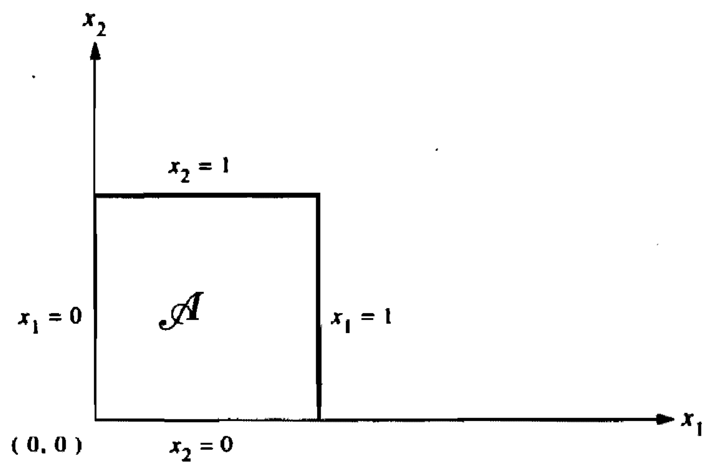

FIGURE 4.1

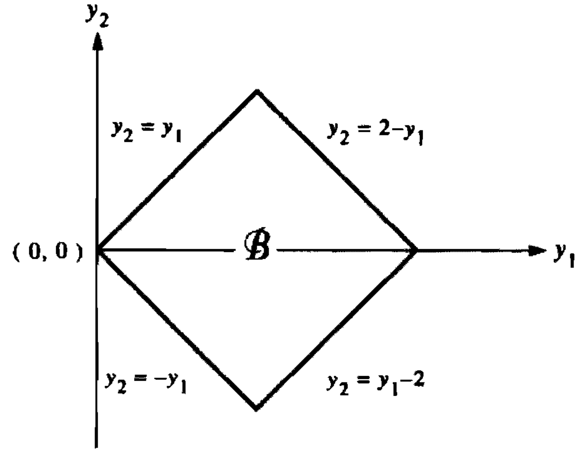

FIGURE 4.2

Accordingly, $\mathscr{B}$ is shown in Figure 4.2. Finally,

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{1}{2} & -\dfrac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Remark. Although, in Example 3, we suggest transforming the boundaries of $\mathscr{A}$, others might want to use the inequalities

$$0 < x_1 < 1 \quad \text{and} \quad 0 < x_2 < 1$$

directly. These four inequalities become

$$0 < \tfrac{1}{2}(y_1 + y_2) < 1 \quad \text{and} \quad 0 < \tfrac{1}{2}(y_1 - y_2) < 1.$$

It is easy to see that these are equivalent to

$$-y_1 < y_2, \qquad y_2 < 2 - y_1, \qquad y_2 < y_1, \qquad y_1 - 2 < y_2;$$

and they define the set $\mathscr{B}$. In this example, these methods were rather simple and essentially the same. Other examples could present more complicated transformations, and only experience can help one decide which is the best method in each case.

We now proceed with the problem of finding the joint p.d.f. of two functions of two continuous-type random variables. Let X_1 and X_2 be random variables of the continuous type, having joint p.d.f. $h(x_1, x_2)$. Let $\mathscr{A}$ be the two-dimensional set in the x_1x_2-plane where $h(x_1, x_2) > 0$. Let $Y_1 = u_1(X_1, X_2)$ be a random variable whose p.d.f. is to be found. If $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define a one-to-one transformation of $\mathscr{A}$ onto a set $\mathscr{B}$ in the y_1y_2-plane (with nonidentically zero Jacobian), we can find, by use of a theorem in analysis, the joint p.d.f. of $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$. Let A be a subset of $\mathscr{A}$, and let B denote the mapping of A under the one-to-one transformation (see Figure 4.3). The events $(X_1, X_2) \in A$ and $(Y_1, Y_2) \in B$ are equivalent. Hence

$$\begin{aligned}\Pr[(Y_1, Y_2) \in B] &= \Pr[(X_1, X_2) \in A] \\ &= \iint_A h(x_1, x_2)\, dx_1\, dx_2.\end{aligned}$$

We wish now to change variables of integration by writing $y_1 = u_1(x_1, x_2), y_2 = u_2(x_1, x_2)$, or $x_1 = w_1(y_1, y_2), x_2 = w_2(y_1, y_2)$. It has been proved in analysis that this change of variables requires

$$\iint_A h(x_1, x_2)\, dx_1\, dx_2 = \iint_B h[w_1(y_1, y_2), w_2(y_1, y_2)]|J|\, dy_1\, dy_2.$$

Thus, for every set B in $\mathscr{B}$,

$$\Pr[(Y_1, Y_2) \in B] = \iint_B h[w_1(y_1, y_2), w_2(y_1, y_2)]|J|\, dy_1\, dy_2,$$

which implies that the joint p.d.f. $g(y_1, y_2)$ of Y_1 and Y_2 is

$$\begin{aligned} g(y_1, y_2) &= h[w_1(y_1, y_2), w_2(y_1, y_2)]|J|, \qquad (y_1, y_2) \in \mathscr{B}, \\ &= 0 \qquad \text{elsewhere.}\end{aligned}$$

Accordingly, the marginal p.d.f. $g_1(y_1)$ of Y_1 can be obtained from the joint p.d.f. $g(y_1, y_2)$ in the usual manner by integrating on y_2. Several examples of this result will be given.

Example 4. Let the random variable X have the p.d.f.

$$\begin{aligned} f(x) &= 1, \qquad 0 < x < 1, \\ &= 0 \qquad \text{elsewhere,}\end{aligned}$$

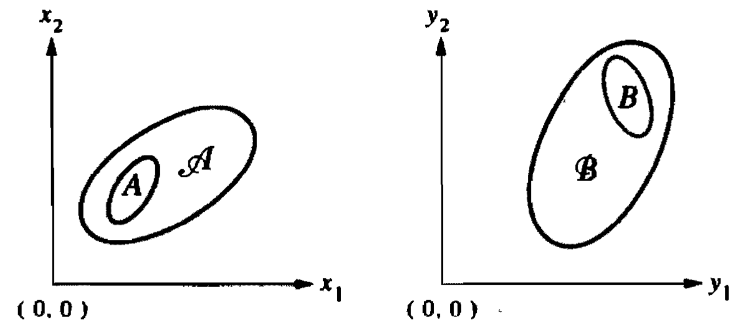

FIGURE 4.3

and let X_1, X_2 denote a random sample from this distribution. The joint p.d.f. of X_1 and X_2 is then

$$\begin{aligned} h(x_1, x_2) &= f(x_1)f(x_2) = 1, \qquad 0 < x_1 < 1, \quad 0 < x_2 < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Consider the two random variables $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. We wish to find the joint p.d.f. of Y_1 and Y_2. Here the two-dimensional space $\mathscr{A}$ in the x_1x_2-plane is that of Example 3 of this section. The one-to-one transformation $y_1 = x_1 + x_2, y_2 = x_1 - x_2$ maps $\mathscr{A}$ onto the space $\mathscr{B}$ of that example. Moreover, the Jacobian of that transformation has been shown to be $J = -\frac{1}{2}$. Thus

$$\begin{aligned} g(y_1, y_2) &= h[\tfrac{1}{2}(y_1 + y_2), \tfrac{1}{2}(y_1 - y_2)]|J| \\ &= f[\tfrac{1}{2}(y_1 + y_2)]f[\tfrac{1}{2}(y_1 - y_2)]|J| = \tfrac{1}{2}, \qquad (y_1, y_2) \in \mathscr{B}, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Because $\mathscr{B}$ is not a product space, the random variables Y_1 and Y_2 are dependent. The marginal p.d.f. of Y_1 is given by

$$g_1(y_1) = \int_{-\infty}^{\infty} g(y_1, y_2)\, dy_2.$$

If we refer to Figure 4.2, it is seen that

$$\begin{aligned} g_1(y_1) &= \int_{-y_1}^{y_1} \tfrac{1}{2}\, dy_2 = y_1, \qquad 0 < y_1 \le 1, \\ &= \int_{y_1 - 2}^{2 - y_1} \tfrac{1}{2}\, dy_2 = 2 - y_1, \qquad 1 < y_1 < 2, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

In a similar manner, the marginal p.d.f. $g_2(y_2)$ is given by

$$\begin{aligned} g_2(y_2) &= \int_{-y_2}^{y_2 + 2} \tfrac{1}{2}\, dy_1 = y_2 + 1, \qquad -1 < y_2 \le 0, \\ &= \int_{y_2}^{2 - y_2} \tfrac{1}{2}\, dy_1 = 1 - y_2, \qquad 0 < y_2 < 1, \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

Example 5. Let X_1, X_2 be a random sample of size $n = 2$ from a standard normal distribution. Say that we are interested in the distribution of $Y_1 = X_1/X_2$. Often in selecting the second random variable, we use the denominator of the ratio or a function of that denominator. So let $Y_2 = X_2$. With the set $\{(x_1, x_2) : -\infty < x_1 < \infty, -\infty < x_2 < \infty\}$, we note that the ratio is not defined at $x_2 = 0$. However, $\Pr(X_2 = 0) = 0$; so we take the p.d.f. of X_2 to be zero at $x_2 = 0$. This results in the set

$$\mathcal{A} = \{(x_1, x_2) : -\infty < x_1 < \infty, \quad -\infty < x_2 < 0 \quad \text{or} \quad 0 < x_2 < \infty\}.$$

With $y_1 = x_1/x_2$, $y_2 = x_2$ or, equivalently, $x_1 = y_1 y_2$, $x_2 = y_2$, $\mathcal{A}$ maps onto

$$\mathcal{B} = \{(y_1, y_2) : -\infty < y_1 < \infty, \quad -\infty < y_2 < 0 \quad \text{or} \quad 0 < y_2 < \infty\}.$$

Also,

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2 \not\equiv 0.$$

Since

$$h(x_1, x_2) = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2) \right], \qquad (x_1, x_2) \in \mathcal{A},$$

we have that the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = \frac{1}{2\pi} \exp\left[-\frac{1}{2} y_2^2(1 + y_1^2) \right] |y_2|, \qquad (y_1, y_2) \in \mathcal{B}.$$

Thus

$$g_1(y_1) = \int_{-\infty}^{0} g(y_1, y_2)\, dy_2 + \int_0^{\infty} g(y_1, y_2)\, dy_2.$$

Since $g(y_1, y_2)$ is an even function of y_2, we can write

$$\begin{aligned} g_1(y_1) &= 2 \int_0^{\infty} \frac{1}{2\pi} \exp\left[-\frac{1}{2} y_2^2(1 + y_1^2) \right] (y_2)\, dy_2 \\ &= \frac{1}{\pi} \left\{ \frac{-\exp[-\frac{1}{2} y_2^2(1 + y_1^2)]}{1 + y_1^2} \right\}_0^{\infty} = \frac{1}{\pi(1 + y_1^2)}, \qquad -\infty < y_1 < \infty. \end{aligned}$$

This marginal p.d.f. of $Y_1 = X_1/X_2$ is that of a *Cauchy distribution*. Although the Cauchy p.d.f. is symmetric about $y_1 = 0$, the mean does not exist because the integral

$$\int_{-\infty}^{\infty} |y_1| g_1(y_1)\, dy_1$$

does not exist. The median and the mode, however, are both equal to zero.

Example 6. Let $Y_1 = \frac{1}{2}(X_1 - X_2)$, where X_1 and X_2 are i.i.d. random variables, each being $\chi^2(2)$. The joint p.d.f. of X_1 and X_2 is

$$\begin{aligned} f(x_1)f(x_2) &= \frac{1}{4} \exp\left(-\frac{x_1 + x_2}{2} \right), \qquad 0 < x_1 < \infty, \quad 0 < x_2 < \infty, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Let $Y_2 = X_2$ so that $y_1 = \frac{1}{2}(x_1 - x_2)$, $y_2 = x_2$ or $x_1 = 2y_1 + y_2$, $x_2 = y_2$ define a one-to-one transformation from $\mathscr{A} = \{(x_1, x_2) : 0 < x_1 < \infty, 0 < x_2 < \infty\}$ onto $\mathscr{B} = \{(y_1, y_2) : -2y_1 < y_2 \text{ and } 0 < y_2, -\infty < y_1 < \infty\}$. The Jacobian of the transformation is

$$J = \begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = 2;$$

hence the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = \frac{|2|}{4} e^{-y_1 - y_2}, \qquad (y_1, y_2) \in \mathscr{B},$$

$$= 0 \qquad \text{elsewhere.}$$

Thus the p.d.f. of Y_1 is given by

$$g_1(y_1) = \int_{-2y_1}^{\infty} \tfrac{1}{2} e^{-y_1 - y_2}\, dy_2 = \tfrac{1}{2} e^{y_1}, \qquad -\infty < y_1 < 0,$$

$$= \int_{0}^{\infty} \tfrac{1}{2} e^{-y_1 - y_2}\, dy_2 = \tfrac{1}{2} e^{-y_1}, \qquad 0 \le y_1 < \infty,$$

or

$$g_1(y_1) = \tfrac{1}{2} e^{-|y_1|}, \qquad -\infty < y_1 < \infty.$$

This p.d.f. is now frequently called the *double exponential p.d.f.*

***Example* 7.** In this example a rather important result is established. Let X_1 and X_2 be independent random variables of the continuous type with joint p.d.f. $f_1(x_1)f_2(x_2)$ that is positive on the two-dimensional space $\mathscr{A}$. Let $Y_1 = u_1(X_1)$, a function of X_1 alone, and $Y_2 = u_2(X_2)$, a function of X_2 alone. We assume for the present that $y_1 = u_1(x_1)$, $y_2 = u_2(x_2)$ define a one-to-one transformation from $\mathscr{A}$ onto a two-dimensional set $\mathscr{B}$ in the y_1y_2-plane. Solving for x_1 and x_2 in terms of y_1 and y_2, we have $x_1 = w_1(y_1)$ and $x_2 = w_2(y_2)$, so

$$J = \begin{vmatrix} w_1'(y_1) & 0 \\ 0 & w_2'(y_2) \end{vmatrix} = w_1'(y_1)w_2'(y_2) \not\equiv 0.$$

Hence the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = f_1[w_1(y_1)]f_2[w_2(y_2)]|w_1'(y_1)w_2'(y_2)|, \qquad (y_1, y_2) \in \mathscr{B},$$

$$= 0 \qquad \text{elsewhere.}$$

However, from the procedure for changing variables in the case of one random variable, we see that the marginal probability density functions of Y_1 and Y_2 are, respectively, $g_1(y_1) = f_1[w_1(y_1)]|w_1'(y_1)|$ and

$g_2(y_2) = f_2[w_2(y_2)]|w_2'(y_2)|$ for y_1 and y_2 in some appropriate sets. Consequently,

$$g(y_1, y_2) \equiv g_1(y_1)g_2(y_2).$$

Thus, summarizing, we note that if X_1 and X_2 are independent random variables, then $Y_1 = u_1(X_1)$ and $Y_2 = u_2(X_2)$ are also independent random variables. It has been seen that the result holds if X_1 and X_2 are of the discrete type; see Exercise 4.22.

In the *simulation* of random variables using uniform random variables, it is frequently difficult to solve $y = F(x)$ for x. Thus other methods are necessary. For instance, consider the important normal case in which we desire to determine X so that it is $N(0, 1)$. Of course, once X is determined, other normal variables can then be obtained through X by the transformation $Z = \sigma X + \mu$.

To simulate normal variables, Box and Muller suggested the following procedure. Let Y_1, Y_2 be a random sample from the uniform distribution over $0 < y < 1$. Define X_1 and X_2 by

$$X_1 = (-2 \ln Y_1)^{1/2} \cos (2\pi Y_2),$$

$$X_2 = (-2 \ln Y_1)^{1/2} \sin (2\pi Y_2).$$

The corresponding transformation is one-to-one and maps $\{(y_1, y_2): 0 < y_1 < 1,\ 0 < y_2 < 1\}$ onto $\{(x_1, x_2): -\infty < x_1 < \infty, -\infty < x_2 < \infty\}$ except for sets involving $x_1 = 0$ and $x_2 = 0$, which have probability zero. The inverse transformation is given by

$$y_1 = \exp\left(-\frac{x_1^2 + x_2^2}{2}\right),$$

$$y_2 = \frac{1}{2\pi} \arctan \frac{x_2}{x_1}.$$

This has the Jacobian

$$J = \begin{vmatrix} (-x_1)\exp\left(-\dfrac{x_1^2 + x_2^2}{2}\right) & (-x_2)\exp\left(-\dfrac{x_1^2 + x_2^2}{2}\right) \\ \dfrac{-x_2/x_1^2}{(2\pi)(1 + x_2^2/x_1^2)} & \dfrac{1/x_1}{(2\pi)(1 + x_2^2/x_1^2)} \end{vmatrix}$$

$$= \frac{-(1 + x_2^2/x_1^2)\exp\left(-\dfrac{x_1^2 + x_2^2}{2}\right)}{(2\pi)(1 + x_2^2/x_1^2)} = \frac{-\exp\left(-\dfrac{x_1^2 + x_2^2}{2}\right)}{2\pi}.$$

Since the joint p.d.f. of Y_1 and Y_2 is 1 on $0 < y_1 < 1$, $0 < y_2 < 1$, and zero elsewhere, the joint p.d.f. of X_1 and X_2 is

$$\frac{\exp\left(-\frac{x_1^2 + x_2^2}{2}\right)}{2\pi}, \qquad -\infty < x_1 < \infty, \quad -\infty < x_2 < \infty.$$

That is, X_1 and X_2 are independent standard normal random variables.

We close this section by observing a way of finding the p.d.f. of a sum of two independent random variables. Let X_1 and X_2 be independent with respective probability density functions $f_1(x_1)$ and $f_2(x_2)$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. Thus we have the one-to-one transformation $x_1 = y_1 - y_2$ and $x_2 = y_2$ with Jacobian $J = 1$. Here we say that $\mathscr{A} = \{(x_1, x_2): -\infty < x_1 < \infty, -\infty < x_2 < \infty\}$ maps onto $\mathscr{B} = \{(y_1, y_2): -\infty < y_1 < \infty, \ -\infty < y_2 < \infty\}$, but we recognize that in a particular problem the joint p.d.f. might equal zero on some part of these sets. Thus the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = f_1(y_1 - y_2)f_2(y_2), \qquad (y_1, y_2) \in \mathscr{B},$$

and the marginal p.d.f. of $Y_1 = X_1 + X_2$ is given by

$$g_1(y_1) = \int_{-\infty}^{\infty} f_1(y_1 - y_2)f_2(y_2)\, dy_2,$$

which is the well-known *convolution formula*.

EXERCISES

4.25. Let X have the p.d.f. $f(x) = x^2/9$, $0 < x < 3$, zero elsewhere. Find the p.d.f. of $Y = X^3$.

4.26. If the p.d.f. of X is $f(x) = 2xe^{-x^2}$, $0 < x < \infty$, zero elsewhere, determine the p.d.f. of $Y = X^2$.

4.27. Let X have the *logistic p.d.f.* $f(x) = e^{-x}/(1 + e^{-x})^2$, $-\infty < x < \infty$.

(a) Show that the graph of $f(x)$ is symmetric about the vertical axis through $x = 0$.

(b) Find the distribution function of X.

(c) Find the p.d.f. of $Y = e^{-X}$.

(d) Show that the m.g.f. $M(t)$ of X is $\Gamma(1 - t)\Gamma(1 + t)$, $-1 < t < 1$. *Hint:* In the integral representing $M(t)$, let $y = (1 + e^{-x})^{-1}$.

4.28. Let X have the uniform distribution over the interval $(-\pi/2, \pi/2)$. Show that $Y = \tan X$ has a Cauchy distribution.

4.29. Let X_1 and X_2 be two independent normal random variables, each with mean zero and variance one (possibly resulting from a Box–Muller transformation). Show that

$$Z_1 = \mu_1 + \sigma_1 X_1,$$

$$Z_2 = \mu_2 + \rho\sigma_2 X_1 + \sigma_2\sqrt{1-\rho^2}X_2,$$

where $0 < \sigma_1$, $0 < \sigma_2$, and $0 < \rho < 1$, have a bivariate normal distribution with respective parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ.

4.30. Let X_1 and X_2 denote a random sample of size 2 from a distribution that is $N(\mu, \sigma^2)$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Find the joint p.d.f. of Y_1 and Y_2 and show that these random variables are independent.

4.31. Let X_1 and X_2 denote a random sample of size 2 from a distribution that is $N(\mu, \sigma^2)$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 + 2X_2$. Show that the joint p.d.f. of Y_1 and Y_2 is bivariate normal with correlation coefficient $3/\sqrt{10}$.

4.32. Use the convolution formula to determine the p.d.f. of $Y_1 = X_1 + X_2$, where X_1 and X_2 are i.i.d. random variables, each with p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere.

Hint: Note that the integral on y_2 has limits of 0 and y_1, where $0 < y_1 < \infty$. Why?

4.33. Let X_1 and X_2 have the joint p.d.f. $h(x_1, x_2) = 2e^{-x_1-x_2}$, $0 < x_1 < x_2 < \infty$, zero elsewhere. Find the joint p.d.f. of $Y_1 = 2X_1$ and $Y_2 = X_2 - X_1$ and argue that Y_1 and Y_2 are independent.

4.34. Let X_1 and X_2 have the joint p.d.f. $h(x_1, x_2) = 8x_1x_2$, $0 < x_1 < x_2 < 1$, zero elsewhere. Find the joint p.d.f. of $Y_1 = X_1/X_2$ and $Y_2 = X_2$ and argue that Y_1 and Y_2 are independent.

Hint: Use the inequalities $0 < y_1y_2 < y_2 < 1$ in considering the mapping from $\mathscr{A}$ onto $\mathscr{B}$.

4.4 The Beta, *t*, and *F* Distributions

It is the purpose of this section to define three additional distributions quite useful in certain problems of statistical inference. These are called, respectively, the beta distribution, the (Student's) t-distribution, and the F-distribution.

The beta distribution. Let X_1 and X_2 be two independent random variables that have gamma distributions and joint p.d.f.

$$h(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1-x_2}, \qquad 0 < x_1 < \infty, \quad 0 < x_2 < \infty,$$

zero elsewhere, where $\alpha > 0$, $\beta > 0$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$. We shall show that Y_1 and Y_2 are independent.

The space $\mathscr{A}$ is, exclusive of the points on the coordinate axes, the first quadrant of the x_1x_2-plane. Now

$$y_1 = u_1(x_1, x_2) = x_1 + x_2,$$

$$y_2 = u_2(x_1, x_2) = \frac{x_1}{x_1 + x_2}$$

may be written $x_1 = y_1y_2$, $x_2 = y_1(1 - y_2)$, so

$$J = \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} = -y_1 \not\equiv 0.$$

The transformation is one-to-one, and it maps $\mathscr{A}$ onto $\mathscr{B} = \{(y_1, y_2) : 0 < y_1 < \infty, 0 < y_2 < 1\}$ in the y_1y_2-plane. The joint p.d.f. of Y_1 and Y_2 is then

$$g(y_1, y_2) = (y_1) \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (y_1y_2)^{\alpha-1} [y_1(1 - y_2)]^{\beta-1} e^{-y_1}$$

$$= \frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha+\beta-1} e^{-y_1}, \qquad 0 < y_1 < \infty, \quad 0 < y_2 < 1,$$

$$= 0 \qquad \text{elsewhere.}$$

In accordance with Theorem 1, Section 2.4, the random variables are independent. The marginal p.d.f. of Y_2 is

$$g_2(y_2) = \frac{y_2^{\alpha-1}(1 - y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty y_1^{\alpha+\beta-1} e^{-y_1}\, dy_1,$$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1}(1 - y_2)^{\beta-1}, \qquad 0 < y_2 < 1,$$

$$= 0 \qquad \text{elsewhere.}$$

This p.d.f. is that of the *beta distribution* with parameters α and β. Since $g(y_1, y_2) \equiv g_1(y_1)g_2(y_2)$, it must be that the p.d.f. of Y_1 is

$$g_1(y_1) = \frac{1}{\Gamma(\alpha + \beta)} y_1^{\alpha+\beta-1} e^{-y_1}, \qquad 0 < y_1 < \infty,$$

$$= 0 \qquad \text{elsewhere,}$$

which is that of a gamma distribution with parameter values of $\alpha + \beta$ and 1.

It is an easy exercise to show that the mean and the variance of Y_2, which has a beta distribution with parameters α and β, are, respectively,

$$\mu = \frac{\alpha}{\alpha + \beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

The *t*-distribution. Let W denote a random variable that is $N(0, 1)$; let V denote a random variable that is $\chi^2(r)$; and let W and V be independent. Then the joint p.d.f. of W and V, say $h(w, v)$, is the product of the p.d.f. of W and that of V or

$$h(w,v) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \frac{1}{\Gamma(r/2)2^{r/2}} v^{r/2-1} e^{-v/2}, \qquad -\infty < w < \infty, \quad 0 < v < \infty,$$

$$= 0 \qquad \text{elsewhere.}$$

Define a new random variable T by writing

$$T = \frac{W}{\sqrt{V/r}}.$$

The change-of-variable technique will be used to obtain the p.d.f. $g_1(t)$ of T. The equations

$$t = \frac{w}{\sqrt{v/r}} \qquad \text{and} \qquad u = v$$

define a one-to-one transformation that maps $\mathcal{A} = \{(w, v) : -\infty < w < \infty,\ 0 < v < \infty\}$ onto $\mathcal{B} = \{(t, u) : -\infty < t < \infty,\ 0 < u < \infty\}$. Since $w = t\sqrt{u}/\sqrt{r}$, $v = u$, the absolute value of the Jacobian of the transformation is $|J| = \sqrt{u}/\sqrt{r}$. Accordingly, the joint p.d.f. of T and $U = V$ is given by

$$g(t, u) = h\left(\frac{t\sqrt{u}}{\sqrt{r}}, u\right)|J|$$

$$= \frac{1}{\sqrt{2\pi}\,\Gamma(r/2)2^{r/2}} u^{r/2-1} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{r}\right)\right]\frac{\sqrt{u}}{\sqrt{r}}, \qquad -\infty < t < \infty, \quad 0 < u < \infty,$$

$$= 0 \qquad \text{elsewhere.}$$

The marginal p.d.f. of T is then

$$g_1(t) = \int_{-\infty}^{\infty} g(t, u)\, du$$

$$= \int_0^{\infty} \frac{1}{\sqrt{2\pi r}\,\Gamma(r/2)2^{r/2}} u^{(r+1)/2-1} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{r}\right)\right] du.$$

In this integral let $z = u[1 + (t^2/r)]/2$, and it is seen that

$$g_1(t) = \int_0^{\infty} \frac{1}{\sqrt{2\pi r}\,\Gamma(r/2)2^{r/2}} \left(\frac{2z}{1 + t^2/r}\right)^{(r+1)/2-1} e^{-z}\left(\frac{2}{1 + t^2/r}\right) dz$$

$$= \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)} \frac{1}{(1 + t^2/r)^{(r+1)/2}}, \qquad -\infty < t < \infty.$$

Thus, if W is $N(0, 1)$, if V is $\chi^2(r)$, and if W and V are independent, then

$$T = \frac{W}{\sqrt{V/r}}$$

has the immediately preceding p.d.f. $g_1(t)$. The distribution of the random variable T is usually called a *t-distribution*. It should be observed that a t-distribution is completely determined by the parameter r, the number of degrees of freedom of the random variable that has the chi-square distribution. Some approximate values of

$$\Pr(T \le t) = \int_{-\infty}^{t} g_1(w)\, dw$$

for selected values of r and t can be found in Table IV in Appendix B.

Remark. This distribution was first discovered by W. S. Gosset when he was working for an Irish brewery. Because that brewery did not want other breweries to know that statistical methods were being used, Gosset published under the pseudonym Student. Thus this distribution is often known as Student's t-distribution.

The *F*-distribution. Next consider two independent chi-square random variables U and V having r_1 and r_2 degrees of freedom, respectively. The joint p.d.f. $h(u, v)$ of U and V is then

$$h(u, v) = \frac{1}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} u^{r_1/2-1} v^{r_2/2-1} e^{-(u+v)/2},$$

$$0 < u < \infty, \quad 0 < v < \infty,$$

$$= 0 \qquad \text{elsewhere.}$$

We define the new random variable

$$W = \frac{U/r_1}{V/r_2}$$

and we propose finding the p.d.f. $g_1(w)$ of W. The equations

$$w = \frac{u/r_1}{v/r_2}, \qquad z = v,$$

define a one-to-one transformation that maps the set $\mathscr{A} = \{(u, v) : 0 < u < \infty, 0 < v < \infty\}$ onto the set $\mathscr{B} = \{(w, z) : 0 < w < \infty, 0 < z < \infty\}$. Since $u = (r_1/r_2)zw$, $v = z$, the absolute value of the Jacobian of the transformation is $|J| = (r_1/r_2)z$. The joint p.d.f. $g(w, z)$ of the random variables W and $Z = V$ is then

$$g(w, z) = \frac{1}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} \left(\frac{r_1 zw}{r_2}\right)^{r_1/2-1} z^{r_2/2-1}$$

$$\times \exp\left[-\frac{z}{2}\left(\frac{r_1 w}{r_2} + 1\right)\right]\frac{r_1 z}{r_2},$$

provided that $(w, z) \in \mathscr{B}$, and zero elsewhere. The marginal p.d.f. $g_1(w)$ of W is then

$$g_1(w) = \int_{-\infty}^{\infty} g(w, z)\, dz$$

$$= \int_0^{\infty} \frac{(r_1/r_2)^{r_1/2}(w)^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} z^{(r_1+r_2)/2-1}$$

$$\times \exp\left[-\frac{z}{2}\left(\frac{r_1 w}{r_2} + 1\right)\right] dz.$$

If we change the variable of integration by writing

$$y = \frac{z}{2}\left(\frac{r_1 w}{r_2} + 1\right),$$

it can be seen that

$$g_1(w) = \int_0^\infty \frac{(r_1/r_2)^{r_1/2}(w)^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}}\left(\frac{2y}{r_1w/r_2+1}\right)^{(r_1+r_2)/2-1} e^{-y} \times \left(\frac{2}{r_1w/r_2+1}\right) dy$$

$$= \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \frac{(w)^{r_1/2-1}}{(1+r_1w/r_2)^{(r_1+r_2)/2}}, \qquad 0 < w < \infty,$$

$$= 0 \qquad \text{elsewhere.}$$

Accordingly, if U and V are independent chi-square variables with r_1 and r_2 degrees of freedom, respectively, then

$$W = \frac{U/r_1}{V/r_2}$$

has the immediately preceding p.d.f. $g_1(w)$. The distribution of this random variable is usually called an *F-distribution*; and we often call the ratio, which we have denoted by W, F. That is,

$$F = \frac{U/r_1}{V/r_2}.$$

It should be observed that an F-distribution is completely determined by the two parameters r_1 and r_2. Table V in Appendix B gives some approximate values of

$$\Pr(F \le b) = \int_0^b g_1(w)\, dw$$

for selected values of r_1, r_2, and b.

EXERCISES

4.35. Find the mean and variance of the beta distribution.

Hint: From that p.d.f., we know that

$$\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\, dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

for all $\alpha > 0$, $\beta > 0$.

4.36. Determine the constant c in each of the following so that each $f(x)$ is a *beta p.d.f.*

(a) $f(x) = cx(1-x)^3$, $0 < x < 1$, zero elsewhere.

(b) $f(x) = cx^4(1-x)^5$, $0 < x < 1$, zero elsewhere.
(c) $f(x) = cx^2(1-x)^8$, $0 < x < 1$, zero elsewhere.

4.37. Determine the constant c so that $f(x) = cx(3-x)^4$, $0 < x < 3$, zero elsewhere, is a p.d.f.

4.38. Show that the graph of the beta p.d.f. is symmetric about the vertical line through $x = \frac{1}{2}$ if $\alpha = \beta$.

4.39. Show, for $k = 1, 2, \ldots, n$, that

$$\int_p^1 \frac{n!}{(k-1)!\,(n-k)!} z^{k-1}(1-z)^{n-k}\,dz = \sum_{x=0}^{k-1} \binom{n}{x} p^x(1-p)^{n-x}.$$

This demonstrates the relationship between the distribution functions of the beta and binomial distributions.

4.40. Let T have a t-distribution with 10 degrees of freedom. Find $\Pr(|T| > 2.228)$ from Table IV.

4.41. Let T have a t-distribution with 14 degrees of freedom. Determine b so that $\Pr(-b < T < b) = 0.90$.

4.42. Let F have an F-distribution with parameters r_1 and r_2. Prove that $1/F$ has an F-distribution with parameters r_2 and r_1.

4.43. If F has an F-distribution with parameters $r_1 = 5$ and $r_2 = 10$, find a and b so that $\Pr(F \le a) = 0.05$ and $\Pr(F \le b) = 0.95$, and, accordingly, $\Pr(a < F < b) = 0.90$.

Hint: Write $\Pr(F \le a) = \Pr(1/F \ge 1/a) = 1 - \Pr(1/F \le 1/a)$, and use the result of Exercise 4.42 and Table V.

4.44. Let $T = W/\sqrt{V/r}$, where the independent variables W and V are, respectively, normal with mean zero and variance 1 and chi-square with r degrees of freedom. Show that T^2 has an F-distribution with parameters $r_1 = 1$ and $r_2 = r$.

Hint: What is the distribution of the numerator of T^2?

4.45. Show that the t-distribution with $r = 1$ degree of freedom and the Cauchy distribution are the same.

4.46. Show that

$$Y = \frac{1}{1 + (r_1/r_2)W},$$

where W has an F-distribution with parameters r_1 and r_2, has a beta distribution.

4.47. Let X_1, X_2 be a random sample from a distribution having the p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that $Z = X_1/X_2$ has an F-distribution.

4.5 Extensions of the Change-of-Variable Technique

In Section 4.3 it was seen that the determination of the joint p.d.f. of two functions of two random variables of the continuous type was essentially a corollary to a theorem in analysis having to do with the change of variables in a twofold integral. This theorem has a natural extension to n-fold integrals. This extension is as follows. Consider an integral of the form

$$\int_A \cdots \int h(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

taken over a subset A of an n-dimensional space $\mathscr{A}$. Let

$$y_1 = u_1(x_1, x_2, \ldots, x_n), \qquad y_2 = u_2(x_1, x_2, \ldots, x_n), \ldots,$$
$$y_n = u_n(x_1, \ldots, x_n),$$

together with the inverse functions

$$x_1 = w_1(y_1, y_2, \ldots, y_n), \qquad x_2 = w_2(y_1, y_2, \ldots, y_n), \ldots,$$
$$x_n = w_n(y_1, y_2, \ldots, y_n)$$

define a one-to-one transformation that maps $\mathscr{A}$ onto $\mathscr{B}$ in the $y_1, y_2, \ldots, y_n$ space (and hence maps the subset A of $\mathscr{A}$ onto a subset B of $\mathscr{B}$). Let the first partial derivatives of the inverse functions be continuous and let the n by n determinant (called the Jacobian)

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} & \cdots & \dfrac{\partial x_1}{\partial y_n} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} & \cdots & \dfrac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial x_n}{\partial y_1} & \dfrac{\partial x_n}{\partial y_2} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{vmatrix}$$

not be identically zero in $\mathscr{B}$. Then

$$\int_A \cdots \int h(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$
$$= \int_B \cdots \int h[w_1(y_1, \ldots, y_n), w_2(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n)]$$
$$\times |J|\, dy_1\, dy_2 \cdots dy_n.$$

Whenever the conditions of this theorem are satisfied, we can determine the joint p.d.f. of n functions of n random variables. Appropriate changes of notation in Section 4.3 (to indicate n-space as opposed to 2-space) are all that is needed to show that the joint p.d.f. of the random variables $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, $Y_2 = u_2(X_1, X_2, \ldots, X_n)$, $\ldots, Y_n = u_n(X_1, X_2, \ldots, X_n)$—where the joint p.d.f. of $X_1, X_2, \ldots, X_n$ is $h(x_1, \ldots, x_n)$—is given by

$$g(y_1, y_2, \ldots, y_n) = |J|h[w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n)],$$

when $(y_1, y_2, \ldots, y_n) \in \mathscr{B}$, and is zero elsewhere.

Example 1. Let $X_1, X_2, \ldots, X_{k+1}$ be independent random variables, each having a gamma distribution with $\beta = 1$. The joint p.d.f. of these variables may be written as

$$h(x_1, x_2, \ldots, x_{k+1}) = \prod_{i=1}^{k+1} \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{-x_i}, \qquad 0 < x_i < \infty,$$

$$= 0 \qquad \text{elsewhere.}$$

Let

$$Y_i = \frac{X_i}{X_1 + X_2 + \cdots + X_{k+1}}, \qquad i = 1, 2, \ldots, k,$$

and $Y_{k+1} = X_1 + X_2 + \cdots + X_{k+1}$ denote $k + 1$ new random variables. The associated transformation maps $\mathscr{A} = \{(x_1, \ldots, x_{k+1}) : 0 < x_i < \infty,\ i = 1, \ldots, k + 1\}$ onto the space

$$\mathscr{B} = \{(y_1, \ldots, y_k, y_{k+1}) : 0 < y_i,\ i = 1, \ldots, k,$$
$$y_1 + \cdots + y_k < 1,\ 0 < y_{k+1} < \infty\}.$$

The single-valued inverse functions are $x_1 = y_1 y_{k+1}, \ldots, x_k = y_k y_{k+1}$, $x_{k+1} = y_{k+1}(1 - y_1 - \cdots - y_k)$, so that the Jacobian is

$$J = \begin{vmatrix} y_{k+1} & 0 & \cdots & 0 & y_1 \\ 0 & y_{k+1} & \cdots & 0 & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & y_{k+1} & y_k \\ -y_{k+1} & -y_{k+1} & \cdots & -y_{k+1} & (1 - y_1 - \cdots - y_k) \end{vmatrix} = y_{k+1}^k.$$

Hence the joint p.d.f. of $Y_1, \ldots, Y_k, Y_{k+1}$ is given by

$$\frac{y_{k+1}^{\alpha_1 + \cdots + \alpha_{k+1} - 1} y_1^{\alpha_1 - 1} \cdots y_k^{\alpha_k - 1}(1 - y_1 - \cdots - y_k)^{\alpha_{k+1} - 1} e^{-y_{k+1}}}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)\Gamma(\alpha_{k+1})},$$

provided that $(y_1, \ldots, y_k, y_{k+1}) \in \mathscr{B}$ and is equal to zero elsewhere. The joint p.d.f. of $Y_1, \ldots, Y_k$ is seen by inspection to be given by

$$g(y_1, \ldots, y_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_{k+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{k+1})} y_1^{\alpha_1 - 1} \cdots y_k^{\alpha_k - 1}(1 - y_1 - \cdots - y_k)^{\alpha_{k+1} - 1},$$

when $0 < y_i$, $i = 1, \ldots, k$, $y_1 + \cdots + y_k < 1$, while the function g is equal to zero elsewhere. Random variables $Y_1, \ldots, Y_k$ that have a joint p.d.f. of this form are said to have a *Dirichlet distribution* with parameters $\alpha_1, \ldots, \alpha_k, \alpha_{k+1}$, and any such $g(y_1, \ldots, y_k)$ is called a *Dirichlet p.d.f.* It is seen, in the special case of $k = 1$, that the Dirichlet p.d.f. becomes a beta p.d.f. Moreover, it is also clear from the joint p.d.f. of $Y_1, \ldots, Y_k, Y_{k+1}$ that Y_{k+1} has a gamma distribution with parameters $\alpha_1 + \cdots + \alpha_k + \alpha_{k+1}$ and $\beta = 1$ and that Y_{k+1} is independent of $Y_1, Y_2, \ldots, Y_k$.

We now consider some other problems that are encountered when transforming variables. Let X have the Cauchy p.d.f.

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

and let $Y = X^2$. We seek the p.d.f. $g(y)$ of Y. Consider the transformation $y = x^2$. This transformation maps the space of X, $\mathscr{A} = \{x : -\infty < x < \infty\}$, onto $\mathscr{B} = \{y : 0 \le y < \infty\}$. However, the transformation is not one-to-one. To each $y \in \mathscr{B}$, with the exception of $y = 0$, there correspond two points $x \in \mathscr{A}$. For example, if $y = 4$, we may have either $x = 2$ or $x = -2$. In such an instance, we represent $\mathscr{A}$ as the union of two disjoint sets A_1 and A_2 such that $y = x^2$ defines a one-to-one transformation that maps each of A_1 and A_2 onto $\mathscr{B}$. If we take A_1 to be $\{x : -\infty < x < 0\}$ and A_2 to be $\{x : 0 \le x < \infty\}$, we see that A_1 is mapped onto $\{y : 0 < y < \infty\}$, whereas A_2 is mapped onto $\{y : 0 \le y < \infty\}$, and these sets are not the same. Our difficulty is caused by the fact that $x = 0$ is an element of $\mathscr{A}$. Why, then, do we not return to the Cauchy p.d.f. and take $f(0) = 0$? Then our new $\mathscr{A}$ is $\mathscr{A} = \{-\infty < x < \infty$ but $x \ne 0\}$. We then take $A_1 = \{x : -\infty < x < 0\}$ and $A_2 = \{x : 0 < x < \infty\}$. Thus $y = x^2$, with the inverse $x = -\sqrt{y}$, maps A_1 onto $\mathscr{B} = \{y : 0 < y < \infty\}$ and the transformation is one-to-one. Moreover, the transformation $y = x^2$, with inverse $x = \sqrt{y}$, maps A_2 onto $\mathscr{B} = \{y : 0 < y < \infty\}$ and the transformation is one-to-one. Consider the probability $\Pr(Y \in B)$, where $B \subset \mathscr{B}$. Let $A_3 = \{x : x = -\sqrt{y}, y \in B\} \subset A_1$ and let $A_4 = \{x : x = \sqrt{y}, y \in B\} \subset A_2$. Then $Y \in B$ when and only when

$X \in A_3$ or $X \in A_4$. Thus we have

$$\Pr(Y \in B) = \Pr(X \in A_3) + \Pr(X \in A_4)$$
$$= \int_{A_3} f(x)\, dx + \int_{A_4} f(x)\, dx.$$

In the first of these integrals, let $x = -\sqrt{y}$. Thus the Jacobian, say J_1, is $-1/2\sqrt{y}$; moreover, the set A_3 is mapped onto B. In the second integral let $x = \sqrt{y}$. Thus the Jacobian, say J_2, is $1/2\sqrt{y}$; moreover, the set A_4 is also mapped onto B. Finally,

$$\Pr(Y \in B) = \int_B f(-\sqrt{y}) \left| -\frac{1}{2\sqrt{y}} \right| dy + \int_B f(\sqrt{y}) \frac{1}{2\sqrt{y}}\, dy$$
$$= \int_B [f(-\sqrt{y}) + f(\sqrt{y})] \frac{1}{2\sqrt{y}}\, dy.$$

Hence the p.d.f. of Y is given by

$$g(y) = \frac{1}{2\sqrt{y}} [f(-\sqrt{y}) + f(\sqrt{y})], \qquad y \in B.$$

With $f(x)$ the Cauchy p.d.f. we have

$$g(y) = \frac{1}{\pi(1 + y)\sqrt{y}}, \qquad 0 < y < \infty,$$
$$= 0 \qquad \text{elsewhere.}$$

In the preceding discussion of a random variable of the continuous type, we had two inverse functions, $x = -\sqrt{y}$ and $x = \sqrt{y}$. That is why we sought to partition $\mathscr{A}$ (or a modification of $\mathscr{A}$) into two disjoint subsets such that the transformation $y = x^2$ maps each onto the same $\mathscr{B}$. Had there been three inverse functions, we would have sought to partition $\mathscr{A}$ (or a modified form of $\mathscr{A}$) into three disjoint subsets, and so on. It is hoped that this detailed discussion will make the following paragraph easier to read.

Let $h(x_1, x_2, \ldots, x_n)$ be the joint p.d.f. of $X_1, X_2, \ldots, X_n$, which are random variables of the continuous type. Let $\mathscr{A}$ be the n-dimensional space where $h(x_1, x_2, \ldots, x_n) > 0$, and consider the transformation $y_1 = u_1(x_1, x_2, \ldots, x_n)$, $y_2 = u_2(x_1, x_2, \ldots, x_n)$, $\ldots$, $y_n = u_n(x_1, x_2, \ldots, x_n)$, which maps $\mathscr{A}$ onto $\mathscr{B}$ in the $y_1, y_2, \ldots, y_n$ space. To each point of $\mathscr{A}$ there will correspond, of course, but one point in $\mathscr{B}$; but to a point in $\mathscr{B}$ there may correspond more than one point in $\mathscr{A}$. That is, the transformation may not be one-to-one.

Suppose, however, that we can represent $\mathscr{A}$ as the union of a finite number, say k, of mutually disjoint sets $A_1, A_2, \ldots, A_k$ so that

$$y_1 = u_1(x_1, x_2, \ldots, x_n), \ldots, \qquad y_n = u_n(x_1, x_2, \ldots, x_n)$$

define a one-to-one transformation of each A_i onto $\mathscr{B}$. Thus, to each point in $\mathscr{B}$ there will correspond exactly one point in each of $A_1, A_2, \ldots, A_k$. Let

$$\begin{aligned} x_1 &= w_{1i}(y_1, y_2, \ldots, y_n), \\ x_2 &= w_{2i}(y_1, y_2, \ldots, y_n), \\ &\vdots \\ x_n &= w_{ni}(y_1, y_2, \ldots, y_n), \end{aligned} \qquad i = 1, 2, \ldots, k,$$

denote the k groups of n inverse functions, one group for each of these k transformations. Let the first partial derivatives be continuous and let each

$$J_i = \begin{vmatrix} \frac{\partial w_{1i}}{\partial y_1} & \frac{\partial w_{1i}}{\partial y_2} & \cdots & \frac{\partial w_{1i}}{\partial y_n} \\ \frac{\partial w_{2i}}{\partial y_1} & \frac{\partial w_{2i}}{\partial y_2} & \cdots & \frac{\partial w_{2i}}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial w_{ni}}{\partial y_1} & \frac{\partial w_{ni}}{\partial y_2} & \cdots & \frac{\partial w_{ni}}{\partial y_n} \end{vmatrix}, \quad i = 1, 2, \ldots, k,$$

be not identically equal to zero in $\mathscr{B}$. From a consideration of the probability of the union of k mutually exclusive events and by applying the change of variable technique to the probability of each of these events, it can be seen that the joint p.d.f. of $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, $Y_2 = u_2(X_1, X_2, \ldots, X_n), \ldots, Y_n = u_n(X_1, X_2, \ldots, X_n)$, is given by

$$g(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{k} |J_i| h[w_{1i}(y_1, \ldots, y_n), \ldots, w_{ni}(y_1, \ldots, y_n)],$$

provided that $(y_1, y_2, \ldots, y_n) \in \mathscr{B}$, and equals zero elsewhere. The p.d.f. of any Y_i, say Y_1, is then

$$g_1(y_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_n)\, dy_2 \cdots dy_n.$$

An illustrative example follows.

Example 2. To illustrate the result just obtained, take $n = 2$ and let X_1, X_2 denote a random sample of size 2 from a standard normal distribution. The joint p.d.f. of X_1 and X_2 is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right), \qquad -\infty < x_1 < \infty, \quad -\infty < x_2 < \infty.$$

Let Y_1 denote the mean and let Y_2 denote twice the variance of the random sample. The associated transformation is

$$y_1 = \frac{x_1 + x_2}{2},$$

$$y_2 = \frac{(x_1 - x_2)^2}{2}.$$

This transformation maps $\mathscr{A} = \{(x_1, x_2): -\infty < x_1 < \infty, -\infty < x_2 < \infty\}$ onto $\mathscr{B} = \{(y_1, y_2): -\infty < y_1 < \infty,\ 0 \le y_2 < \infty\}$. But the transformation is not one-to-one because, to each point in $\mathscr{B}$, exclusive of points where $y_2 = 0$, there correspond two points in $\mathscr{A}$. In fact, the two groups of inverse functions are

$$x_1 = y_1 - \sqrt{\frac{y_2}{2}}, \qquad x_2 = y_1 + \sqrt{\frac{y_2}{2}}$$

and

$$x_1 = y_1 + \sqrt{\frac{y_2}{2}}, \qquad x_2 = y_1 - \sqrt{\frac{y_2}{2}}.$$

Moreover, the set $\mathscr{A}$ cannot be represented as the union of two disjoint sets, each of which under our transformation maps onto $\mathscr{B}$. Our difficulty is caused by those points of $\mathscr{A}$ that lie on the line whose equation is $x_2 = x_1$. At each of these points, we have $y_2 = 0$. However, we can define $f(x_1, x_2)$ to be zero at each point where $x_1 = x_2$. We can do this without altering the distribution of probability, because the probability measure of this set is zero. Thus we have a new $\mathscr{A} = \{(x_1, x_2): -\infty < x_1 < \infty, -\infty < x_2 < \infty, \text{ but } x_1 \ne x_2\}$. This space is the union of the two disjoint sets $A_1 = \{(x_1, x_2): x_2 > x_1\}$ and $A_2 = \{(x_1, x_2): x_2 < x_1\}$. Moreover, our transformation now defines a one-to-one transformation of each A_i, $i = 1, 2$, onto the new $\mathscr{B} = \{(y_1, y_2): -\infty < y_1 < \infty, 0 < y_2 < \infty\}$. We can now find the joint p.d.f., say $g(y_1, y_2)$, of the mean Y_1 and twice the variance Y_2 of our random sample. An easy computation shows that $|J_1| = |J_2| = 1/\sqrt{2y_2}$. Thus

$$\begin{aligned} g(y_1, y_2) &= \frac{1}{2\pi} \exp\left[-\frac{(y_1 - \sqrt{y_2/2})^2}{2} - \frac{(y_1 + \sqrt{y_2/2})^2}{2}\right] \frac{1}{\sqrt{2y_2}} \\ &\quad + \frac{1}{2\pi} \exp\left[-\frac{(y_1 + \sqrt{y_2/2})^2}{2} - \frac{(y_1 - \sqrt{y_2/2})^2}{2}\right] \frac{1}{\sqrt{2y_2}} \\ &= \sqrt{\frac{2}{2\pi}}\, e^{-y_1^2} \frac{1}{\sqrt{2}\,\Gamma(\frac{1}{2})} y_2^{1/2-1} e^{-y_2/2}, \qquad -\infty < y_1 < \infty, \quad 0 < y_2 < \infty. \end{aligned}$$

We can make three interesting observations. The mean Y_1 of our random sample is $N(0, \frac{1}{2})$; Y_2, which is twice the variance of our sample, is $\chi^2(1)$; and the two are independent. Thus the mean and the variance of our sample are independent.

EXERCISES

4.48. Let X_1, X_2, X_3 denote a random sample from a standard normal distribution. Let the random variables Y_1, Y_2, Y_3 be defined by

$$X_1 = Y_1 \cos Y_2 \sin Y_3, \qquad X_2 = Y_1 \sin Y_2 \sin Y_3, \qquad X_3 = Y_1 \cos Y_3,$$

where $0 \le Y_1 < \infty$, $0 \le Y_2 < 2\pi$, $0 \le Y_3 \le \pi$. Show that Y_1, Y_2, Y_3 are mutually independent.

4.49. Let X_1, X_2, X_3 be i.i.d., each with the distribution having p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that

$$Y_1 = \frac{X_1}{X_1 + X_2}, \qquad Y_2 = \frac{X_1 + X_2}{X_1 + X_2 + X_3}, \qquad Y_3 = X_1 + X_2 + X_3$$

are mutually independent.

4.50. Let $X_1, X_2, \ldots, X_r$ be r independent gamma variables with parameters $\alpha = \alpha_i$ and $\beta = 1$, $i = 1, 2, \ldots, r$, respectively. Show that $Y_1 = X_1 + X_2 + \cdots + X_r$ has a gamma distribution with parameters $\alpha = \alpha_1 + \cdots + \alpha_r$ and $\beta = 1$.

Hint: Let $Y_2 = X_2 + \cdots + X_r$, $Y_3 = X_3 + \cdots + X_r, \ldots, Y_r = X_r$.

4.51. Let $Y_1, \ldots, Y_k$ have a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k, \alpha_{k+1}$.

(a) Show that Y_1 has a beta distribution with parameters $\alpha = \alpha_1$ and $\beta = \alpha_2 + \cdots + \alpha_{k+1}$.

(b) Show that $Y_1 + \cdots + Y_r$, $r \le k$, has a beta distribution with parameters $\alpha = \alpha_1 + \cdots + \alpha_r$ and $\beta = \alpha_{r+1} + \cdots + \alpha_{k+1}$.

(c) Show that $Y_1 + Y_2$, $Y_3 + Y_4$, $Y_5, \ldots, Y_k$, $k \ge 5$, have a Dirichlet distribution with parameters $\alpha_1 + \alpha_2$, $\alpha_3 + \alpha_4$, $\alpha_5, \ldots, \alpha_k, \alpha_{k+1}$.

Hint: Recall the definition of Y_i in Example 1 and use the fact that the sum of several independent gamma variables with $\beta = 1$ is a gamma variable (Exercise 4.50).

4.52. Let X_1, X_2, and X_3 be three independent chi-square variables with r_1, r_2, and r_3 degrees of freedom, respectively.

(a) Show that $Y_1 = X_1/X_2$ and $Y_2 = X_1 + X_2$ are independent and that Y_2 is $\chi^2(r_1 + r_2)$.

(b) Deduce that

$$\frac{X_1/r_1}{X_2/r_2} \quad \text{and} \quad \frac{X_3/r_3}{(X_1 + X_2)/(r_1 + r_2)}$$

are independent F-variables.

4.53. If $f(x) = \frac{1}{2}$, $-1 < x < 1$, zero elsewhere, is the p.d.f. of the random variable X, find the p.d.f. of $Y = X^2$.

4.54. If X_1, X_2 is a random sample from a standard normal distribution, find the joint p.d.f. of $Y_1 = X_1^2 + X_2^2$ and $Y_2 = X_2$ and the marginal p.d.f. of Y_1.

Hint: Note that the space of Y_1 and Y_2 is given by $-\sqrt{y_1} < y_2 < \sqrt{y_1}$, $0 < y_1 < \infty$.

4.55. If X has the p.d.f. $f(x) = \frac{1}{4}$, $-1 < x < 3$, zero elsewhere, find the p.d.f. of $Y = X^2$.

Hint: Here $\mathscr{B} = \{y : 0 \le y < 9\}$ and the event $Y \in B$ is the union of two mutually exclusive events if $B = \{y : 0 < y < 1\}$.

4.6 Distributions of Order Statistics

In this section the notion of an order statistic will be defined and we shall investigate some of the simpler properties of such a statistic. These statistics have in recent times come to play an important role in statistical inference partly because some of their properties do not depend upon the distribution from which the random sample is obtained.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of the *continuous type* having a p.d.f. $f(x)$ that is positive, provided that $a < x < b$. Let Y_1 be the smallest of these X_i, Y_2 the next X_i in order of magnitude, . . . , and Y_n the largest X_i. That is, $Y_1 < Y_2 < \cdots < Y_n$ represent $X_1, X_2, \ldots, X_n$ when the latter are arranged in ascending order of magnitude. Then Y_i, $i = 1, 2, \ldots, n$, is called the ith order statistic of the random sample $X_1, X_2, \ldots, X_n$. It will be shown that the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is given by

$$\begin{aligned} g(y_1, y_2, \ldots, y_n) &= (n!)f(y_1)f(y_2)\cdots f(y_n), \\ &\qquad\qquad a < y_1 < y_2 < \cdots < y_n < b, \\ &= 0 \qquad \text{elsewhere.} \end{aligned} \tag{1}$$

We shall prove this only for the case $n = 3$, but the argument is seen to be entirely general. With $n = 3$, the joint p.d.f. of X_1, X_2, X_3 is

$f(x_1)f(x_2)f(x_3)$. Consider a probability such as $\Pr(a < X_1 = X_2 < b, a < X_3 < b)$. This probability is given by

$$\int_a^b \int_a^b \int_{x_2}^{x_2} f(x_1)f(x_2)f(x_3)\,dx_1\,dx_2\,dx_3 = 0,$$

since

$$\int_{x_2}^{x_2} f(x_1)\,dx_1$$

is defined in calculus to be zero. As has been pointed out, we may, without altering the distribution of X_1, X_2, X_3, define the joint p.d.f. $f(x_1)f(x_2)f(x_3)$ to be zero at all points (x_1, x_2, x_3) that have at least two of their coordinates equal. Then the set $\mathscr{A}$, where $f(x_1)f(x_2)f(x_3) > 0$, is the union of the six mutually disjoint sets:

$$A_1 = \{(x_1, x_2, x_3) : a < x_1 < x_2 < x_3 < b\},$$

$$A_2 = \{(x_1, x_2, x_3) : a < x_2 < x_1 < x_3 < b\},$$

$$A_3 = \{(x_1, x_2, x_3) : a < x_1 < x_3 < x_2 < b\},$$

$$A_4 = \{(x_1, x_2, x_3) : a < x_2 < x_3 < x_1 < b\},$$

$$A_5 = \{(x_1, x_2, x_3) : a < x_3 < x_1 < x_2 < b\},$$

$$A_6 = \{(x_1, x_2, x_3) : a < x_3 < x_2 < x_1 < b\}.$$

There are six of these sets because we can arrange x_1, x_2, x_3 in precisely $3! = 6$ ways. Consider the functions y_1 = minimum of x_1, x_2, x_3; y_2 = middle in magnitude of x_1, x_2, x_3; and y_3 = maximum of x_1, x_2, x_3. These functions define one-to-one transformations that map each of $A_1, A_2, \ldots, A_6$ onto the same set $\mathscr{B} = \{(y_1, y_2, y_3) : a < y_1 < y_2 < y_3 < b\}$. The inverse functions are, for points in A_1, $x_1 = y_1$, $x_2 = y_2$, $x_3 = y_3$; for points in A_2, they are $x_1 = y_2$, $x_2 = y_1$, $x_3 = y_3$; and so on, for each of the remaining four sets. Then we have that

$$J_1 = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1$$

and

$$J_2 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1.$$

It is easily verified that the absolute value of each of the $3! = 6$ Jacobians is $+1$. Thus the joint p.d.f. of the three order statistics Y_1 = minimum of X_1, X_2, X_3; Y_2 = middle in magnitude of X_1, X_2, X_3; Y_3 = maximum of X_1, X_2, X_3 is

$$\begin{aligned} g(y_1, y_2, y_3) &= |J_1| f(y_1)f(y_2)f(y_3) + |J_2| f(y_2)f(y_1)f(y_3) + \cdots \\ &\quad + |J_6| f(y_3)f(y_2)f(y_1), \qquad a < y_1 < y_2 < y_3 < b, \\ &= (3!)f(y_1)f(y_2)f(y_3), \qquad a < y_1 < y_2 < y_3 < b, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

This is Equation (1) with $n = 3$.

In accordance with the natural extension of Theorem 1, Section 2.4, to distributions of more than two random variables, it is seen that the order statistics, unlike the items of the random sample, are dependent.

Example 1. Let X denote a random variable of the continuous type with a p.d.f. $f(x)$ that is positive and continuous, provided that $a < x < b$ and is zero elsewhere. The distribution function $F(x)$ of X may be written

$$F(x) = \int_a^x f(w)\, dw, \qquad a < x < b.$$

If $x \le a$, $F(x) = 0$; and if $b \le x$, $F(x) = 1$. Thus there is a unique median m of the distribution with $F(m) = \frac{1}{2}$. Let X_1, X_2, X_3 denote a random sample from this distribution and let $Y_1 < Y_2 < Y_3$ denote the order statistics of the sample. We shall compute the probability that $Y_2 \le m$. The joint p.d.f. of the three order statistics is

$$\begin{aligned} g(y_1, y_2, y_3) &= 6f(y_1)f(y_2)f(y_3), \qquad a < y_1 < y_2 < y_3 < b, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

The p.d.f. of Y_2 is then

$$\begin{aligned} h(y_2) &= 6f(y_2) \int_{y_2}^b \int_a^{y_2} f(y_1)f(y_3)\, dy_1\, dy_3, \\ &= 6f(y_2)F(y_2)[1 - F(y_2)], \qquad a < y_2 < b, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Accordingly,

$$\begin{aligned} \Pr(Y_2 \le m) &= 6 \int_a^m \{F(y_2)f(y_2) - [F(y_2)]^2 f(y_2)\}\, dy_2 \\ &= 6\left\{\frac{[F(y_2)]^2}{2} - \frac{[F(y_2)]^3}{3}\right\}_a^m = \frac{1}{2}. \end{aligned}$$

The procedure used in Example 1 can be used to obtain general formulas for the marginal probability density functions of the order statistics. We shall do this now. Let X denote a random variable of the continuous type having a p.d.f. $f(x)$ that is positive and continuous, provided that $a < x < b$, and is zero elsewhere. Then the distribution function $F(x)$ may be written

$$
\begin{aligned}
F(x) &= 0, \qquad x \le a, \\
&= \int_a^x f(w)\,dw, \qquad a < x < b, \\
&= 1, \qquad b \le x.
\end{aligned}
$$

Accordingly, $F'(x) = f(x)$, $a < x < b$. Moreover, if $a < x < b$,

$$
\begin{aligned}
1 - F(x) &= F(b) - F(x) \\
&= \int_a^b f(w)\,dw - \int_a^x f(w)\,dw \\
&= \int_x^b f(w)\,dw.
\end{aligned}
$$

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from this distribution, and let $Y_1, Y_2, \ldots, Y_n$ denote the order statistics of this random sample. Then the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is

$$
\begin{aligned}
g(y_1, y_2, \ldots, y_n) &= n!\,f(y_1)f(y_2)\cdots f(y_n), \quad a<y_1<y_2<\cdots<y_n<b, \\
&= 0 \qquad \text{elsewhere.}
\end{aligned}
$$

It will first be shown how the marginal p.d.f. of Y_n may be expressed in terms of the distribution function $F(x)$ and the p.d.f. $f(x)$ of the random variable X. If $a < y_n < b$, the marginal p.d.f. of y_n is given by

$$
\begin{aligned}
g_n(y_n) &= \int_a^{y_n} \cdots \int_a^{y_4}\int_a^{y_3}\int_a^{y_2} n!\,f(y_1)f(y_2)\cdots f(y_n)\,dy_1\,dy_2\,dy_3\cdots dy_{n-1} \\
&= \int_a^{y_n} \cdots \int_a^{y_4}\int_a^{y_3} n!\left(\int_a^{y_2} f(y_1)\,dy_1\right)f(y_2)\cdots f(y_n)\,dy_2\cdots dy_{n-1} \\
&= \int_a^{y_n} \cdots \int_a^{y_4}\int_a^{y_3} n!\,F(y_2)f(y_2)\cdots f(y_n)\,dy_2\cdots dy_{n-1},
\end{aligned}
$$

since $F(x) = \int_a^x f(w)\,dw$. Now

$$\int_a^{y_3} F(y_2)f(y_2)\,dy_2 = \frac{[F(y_2)]^2}{2}\Bigg|_a^{y_3}$$

$$= \frac{[F(y_3)]^2}{2},$$

since $F(a) = 0$. Thus

$$g_n(y_n) = \int_a^{y_n} \cdots \int_a^{y_4} n!\,\frac{[F(y_3)]^2}{2} f(y_3)\cdots f(y_n)\,dy_3 \cdots dy_{n-1}.$$

But

$$\int_a^{y_4} \frac{[F(y_3)]^2}{2} f(y_3)\,dy_3 = \frac{[F(y_3)]^3}{2\cdot 3}\Bigg|_a^{y_4} = \frac{[F(y_4)]^3}{2\cdot 3},$$

so

$$g_n(y_n) = \int_a^{y_n} \cdots \int_a^{y_5} n!\,\frac{[F(y_4)]^3}{3!} f(y_4)\cdots f(y_n)\,dy_4 \cdots dy_{n-1}.$$

If the successive integrations on $y_4, \ldots, y_{n-1}$ are carried out, it is seen that

$$\begin{aligned} g_n(y_n) &= n!\,\frac{[F(y_n)]^{n-1}}{(n-1)!} f(y_n) \\ &= n[F(y_n)]^{n-1} f(y_n), \qquad a < y_n < b, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

It will next be shown how to express the marginal p.d.f. of Y_1 in terms of $F(x)$ and $f(x)$. We have, for $a < y_1 < b$,

$$g_1(y_1) = \int_{y_1}^{b} \cdots \int_{y_{n-3}}^{b} \int_{y_{n-2}}^{b} \int_{y_{n-1}}^{b} n!\,f(y_1)f(y_2)\cdots f(y_n)\,dy_n\,dy_{n-1}\cdots dy_2$$

$$= \int_{y_1}^{b} \cdots \int_{y_{n-3}}^{b} \int_{y_{n-2}}^{b} n!\,f(y_1)f(y_2)\cdots$$

$$f(y_{n-1})[1 - F(y_{n-1})]\,dy_{n-1}\cdots dy_2.$$

But

$$\int_{y_{n-2}}^{b} [1 - F(y_{n-1})]f(y_{n-1})\, dy_{n-1} = -\frac{[1 - F(y_{n-1})]^2}{2}\Bigg|_{y_{n-2}}^{b}$$

$$= \frac{[1 - F(y_{n-2})]^2}{2},$$

so that

$$g_1(y_1) = \int_{y_1}^{b} \cdots \int_{y_{n-3}}^{b} n!\, f(y_1) \cdots f(y_{n-2}) \frac{[1 - F(y_{n-2})]^2}{2}\, dy_{n-2} \cdots dy_2.$$

Upon completing the integrations, it is found that

$$g_1(y_1) = n[1 - F(y_1)]^{n-1} f(y_1), \qquad a < y_1 < b,$$
$$= 0 \qquad \text{elsewhere.}$$

Once it is observed that

$$\int_a^x [F(w)]^{\alpha - 1} f(w)\, dw = \frac{[F(x)]^\alpha}{\alpha}, \qquad \alpha > 0$$

and that

$$\int_y^b [1 - F(w)]^{\beta - 1} f(w)\, dw = \frac{[1 - F(y)]^\beta}{\beta}, \qquad \beta > 0,$$

it is easy to express the marginal p.d.f. of any order statistic, say Y_k, in terms of $F(x)$ and $f(x)$. This is done by evaluating the integral

$$g_k(y_k) = \int_a^{y_k} \cdots \int_a^{y_2} \int_{y_k}^{b} \cdots \int_{y_{n-1}}^{b} n!\, f(y_1) f(y_2) \cdots f(y_n)\, dy_n \cdots$$
$$dy_{k+1}\, dy_1 \cdots dy_{k-1}.$$

The result is

$$g_k(y_k) = \frac{n!}{(k-1)!\,(n-k)!} [F(y_k)]^{k-1} [1 - F(y_k)]^{n-k} f(y_k),$$
$$a < y_k < b,$$
$$= 0 \qquad \text{elsewhere.} \tag{2}$$

Example 2. Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of a random sample of size 4 from a distribution having p.d.f.

$$f(x) = 2x, \qquad 0 < x < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

We shall express the p.d.f. of Y_3 in terms of $f(x)$ and $F(x)$ and then compute $\Pr(\frac{1}{2} < Y_3)$. Here $F(x) = x^2$, provided that $0 < x < 1$, so that

$$g_3(y_3) = \frac{4!}{2!\,1!}(y_3^2)^2(1 - y_3^2)(2y_3), \qquad 0 < y_3 < 1,$$
$$= 0 \qquad \text{elsewhere.}$$

Thus

$$\Pr(\tfrac{1}{2} < Y_3) = \int_{1/2}^{\infty} g_3(y_3)\,dy_3$$
$$= \int_{1/2}^{1} 24(y_3^5 - y_3^7)\,dy_3 = \tfrac{243}{256}.$$

Finally, the joint p.d.f. of any two order statistics, say $Y_i < Y_j$, is as easily expressed in terms of $F(x)$ and $f(x)$. We have

$$g_{ij}(y_i, y_j) = \int_a^{y_i} \cdots \int_a^{y_2} \int_{y_i}^{y_j} \cdots \int_{y_{j-2}}^{y_j} \int_{y_j}^{b} \cdots \int_{y_{n-1}}^{b} n!\, f(y_1) \cdots$$
$$f(y_n)\,dy_n \cdots dy_{j+1}\,dy_{j-1} \cdots dy_{i+1}\,dy_1 \cdots dy_{i-1}.$$

Since, for $\gamma > 0$,

$$\int_x^y [F(y) - F(w)]^{\gamma - 1} f(w)\,dw = -\frac{[F(y) - F(w)]^\gamma}{\gamma}\Bigg|_x^y$$
$$= \frac{[F(y) - F(x)]^\gamma}{\gamma},$$

it is found that

$$g_{ij}(y_i, y_j) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}$$
$$\times [F(y_i)]^{i-1}[F(y_j) - F(y_i)]^{j-i-1}[1 - F(y_j)]^{n-j} f(y_i) f(y_j) \qquad (3)$$

for $a < y_i < y_j < b$, and zero elsewhere.

Remark. There is an easy method of remembering a p.d.f. like that given in Formula (3). The probability $\Pr(y_i < Y_i < y_i + \Delta_i, y_j < Y_j < y_j + \Delta_j)$, where Δ_i and Δ_j are small, can be approximated by the following multinomial probability. In n independent trials, $i - 1$ outcomes must be less than y_i (an event that has probability $p_1 = F(y_i)$ on each trial); $j - i - 1$ outcomes must be between $y_i + \Delta_i$ and y_j [an event with approximate probability $p_2 = F(y_j) - F(y_i)$ on each trial]; $n - j$ outcomes must be greater than $y_j + \Delta_j$ (an event with approximate probability $p_3 = 1 - F(y_j)$ on each trial); one outcome must be between y_i and $y_i + \Delta_i$ (an event with approximate probability $p_4 = f(y_i)\,\Delta_i$ on each trial); and finally one outcome must be between y_j and $y_j + \Delta_j$ [an event with approximate probability $p_5 = f(y_j)\Delta_j$ on each trial]. This multinomial probability is

$$\frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!\,1!\,1!}\,p_1^{i-1}p_2^{j-i-1}p_3^{n-j}p_4p_5,$$

which is $g_{i,j}(y_i, y_j)\Delta_i\Delta_j$.

Certain functions of the order statistics $Y_1, Y_2, \ldots, Y_n$ are important statistics themselves. A few of these are: (a) $Y_n - Y_1$, which is called the range of the random sample; (b) $(Y_1 + Y_n)/2$, which is called the midrange of the random sample; and (c) if n odd, $Y_{(n+1)/2}$, which is called the median of the random sample.

Example 3. Let Y_1, Y_2, Y_3 be the order statistics of a random sample of size 3 from a distribution having p.d.f.

$$\begin{aligned} f(x) &= 1, \qquad 0 < x < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

We seek the p.d.f. of the sample range $Z_1 = Y_3 - Y_1$. Since $F(x) = x$, $0 < x < 1$, the joint p.d.f. of Y_1 and Y_3 is

$$\begin{aligned} g_{13}(y_1, y_3) &= 6(y_3 - y_1), \qquad 0 < y_1 < y_3 < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

In addition to $Z_1 = Y_3 - Y_1$, let $Z_2 = Y_3$. Consider the functions $z_1 = y_3 - y_1$, $z_2 = y_3$, and their inverses $y_1 = z_2 - z_1$, $y_3 = z_2$, so that the corresponding Jacobian of the one-to-one transformation is

$$J = \begin{vmatrix} \dfrac{\partial y_1}{\partial z_1} & \dfrac{\partial y_1}{\partial z_2} \\ \dfrac{\partial y_3}{\partial z_1} & \dfrac{\partial y_3}{\partial z_2} \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

Thus the joint p.d.f. of Z_1 and Z_2 is

$$\begin{aligned} h(z_1, z_2) &= |-1|6z_1 = 6z_1, \qquad 0 < z_1 < z_2 < 1. \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Accordingly, the p.d.f. of the range $Z_1 = Y_3 - Y_1$ of the random sample of size 3 is

$$\begin{aligned} h_1(z_1) &= \int_{z_1}^{1} 6z_1\,dz_2 = 6z_1(1 - z_1), \qquad 0 < z_1 < 1, \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

EXERCISES

4.56. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size 4 from the distribution having p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Find $\Pr(3 \le Y_4)$.

4.57. Let X_1, X_2, X_3 be a random sample from a distribution of the continuous type having p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere.
(a) Compute the probability that the smallest of these X_i exceeds the median of the distribution.
(b) If $Y_1 < Y_2 < Y_3$ are the order statistics, find the correlation between Y_2 and Y_3.

4.58. Let $f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere, be the p.d.f. of a distribution of the discrete type. Show that the p.d.f. of the smallest observation of a random sample of size 5 from this distribution is

$$g_1(y_1) = \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{6 - y_1}{6}\right)^5, \qquad y_1 = 1, 2, \ldots, 6,$$

zero elsewhere. Note that in this exercise the random sample is from a distribution of the discrete type. All formulas in the text were derived under the assumption that the random sample is from a distribution of the continuous type and are not applicable. Why?

4.59. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution having p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that $Z_1 = Y_2$ and $Z_2 = Y_4 - Y_2$ are independent.
Hint: First find the joint p.d.f. of Y_2 and Y_4.

4.60. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution with p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere. Show that the kth order statistic Y_k has a beta p.d.f. with parameters $\alpha = k$ and $\beta = n - k + 1$.

4.61. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics from a Weibull distribution, Exercise 3.44, Section 3.3. Find the distribution function and p.d.f. of Y_1.

4.62. Find the probability that the range of a random sample of size 4 from the uniform distribution having the p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere, is less than $\frac{1}{2}$.

4.63. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from a distribution having the p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Show that $Z_1 = Y_1/Y_2$, $Z_2 = Y_2/Y_3$, and $Z_3 = Y_3$ are mutually independent.

4.64. If a random sample of size 2 is taken from a distribution having p.d.f. $f(x) = 2(1 - x)$, $0 < x < 1$, zero elsewhere, compute the probability that one sample observation is at least twice as large as the other.

4.65. Let $Y_1 < Y_2 < Y_3$ denote the order statistics of a random sample of size 3 from a distribution with p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere. Let $Z = (Y_1 + Y_3)/2$ be the midrange of the sample. Find the p.d.f. of Z.

4.66. Let $Y_1 < Y_2$ denote the order statistics of a random sample of size 2 from $N(0, \sigma^2)$.

(a) Show that $E(Y_1) = -\sigma/\sqrt{\pi}$.
Hint: Evaluate $E(Y_1)$ by using the joint p.d.f. of Y_1 and Y_2, and first integrating on y_1.

(b) Find the covariance of Y_1 and Y_2.

4.67. Let $Y_1 < Y_2$ be the order statistics of a random sample of size 2 from a distribution of the continuous type which has p.d.f. $f(x)$ such that $f(x) > 0$, provided that $x \geq 0$, and $f(x) = 0$ elsewhere. Show that the independence of $Z_1 = Y_1$ and $Z_2 = Y_2 - Y_1$ characterizes the gamma p.d.f. $f(x)$, which has parameters $\alpha = 1$ and $\beta > 0$.

Hint: Use the change-of-variable technique to find the joint p.d.f. of Z_1 and Z_2 from that of Y_1 and Y_2. Accept the fact that the functional equation $h(0)h(x + y) \equiv h(x)h(y)$ has the solution $h(x) = c_1e^{c_2x}$, where c_1 and c_2 are constants.

4.68. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with p.d.f. $f(x) = 2x$, $0 < x < 1$.

(a) Find the joint p.d.f. of Y_3 and Y_4.

(b) Find the conditional p.d.f. of Y_3, given $Y_4 = y_4$.

(c) Evaluate $E(Y_3|y_4)$.

4.69. Two numbers are selected at random from the interval (0, 1). If these values are uniformly and independently distributed, compute the probability that the three resulting line segments, by cutting the interval at the numbers, can form a triangle.

4.70. Let X and Y denote independent random variables with respective probability density functions $f(x) = 2x$, $0 < x < 1$, zero elsewhere, and $g(y) = 3y^2$, $0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint p.d.f. of U and V.

Hint: Here the two inverse transformations are given by $x = u$, $y = v$ and $x = v$, $y = u$.

4.71. Let the joint p.d.f. of X and Y be $f(x, y) = \frac{12}{7}x(x + y)$, $0 < x < 1$, $0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint p.d.f. of U and V.

4.72. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of either type. A measure of spread is *Gini's mean difference*

$$G = \sum_{j=2}^{n} \sum_{i=1}^{j-1} |X_i - X_j| \Big/ \binom{n}{2}.$$

(a) If $n = 10$, find $a_1, a_2, \ldots, a_{10}$ so that $G = \sum_{i=1}^{10} a_i Y_i$, where $Y_1, Y_2, \ldots, Y_{10}$ are the order statistics of the sample.

(b) Show that $E(G) = 2\sigma/\sqrt{\pi}$ if the sample arises from the normal distribution $N(\mu, \sigma^2)$.

4.73. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from the exponential distribution with p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere.

(a) Show that $Z_1 = nY_1$, $Z_2 = (n-1)(Y_2 - Y_1)$, $Z_3 = (n-2)(Y_3 - Y_2)$, $\ldots, Z_n = Y_n - Y_{n-1}$ are independent and that each Z_i has the exponential distribution.

(b) Demonstrate that all linear functions of $Y_1, Y_2, \ldots, Y_n$, such as $\sum_1^n a_i Y_i$, can be expressed as linear functions of independent random variables.

4.74. In the Program Evaluation and Review Technique (PERT), we are interested in the total time to complete a project that is comprised of a large number of subprojects. For illustration, let X_1, X_2, X_3 be three independent random times for three subprojects. If these subprojects are in series (the first one must be completed before the second starts, etc.), then we are interested in the sum $Y = X_1 + X_2 + X_3$. If these are in parallel (can be worked on simultaneously), then we are interested in $Z = \max(X_1, X_2, X_3)$. In the case each of these random variables has the uniform distribution with p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere, find (a) the p.d.f. of Y and (b) the p.d.f. of Z.

4.7 The Moment-Generating-Function Technique

The change-of-variable procedure has been seen, in certain cases, to be an effective method of finding the distribution of a function of several random variables. An alternative procedure, built around the concept of the m.g.f. of a distribution, will be presented in this section. This procedure is particularly effective in certain instances. We should recall that an m.g.f., when it exists, is unique and that it uniquely determines the distribution of probability.

Let $h(x_1, x_2, \ldots, x_n)$ denote the joint p.d.f. of the n random variables $X_1, X_2, \ldots, X_n$. These random variables may or may not be the observations of a random sample from some distribution that has a given p.d.f. $f(x)$. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$. We seek $g(y_1)$, the p.d.f. of the random variable Y_1. Consider the m.g.f. of Y_1. If it exists, it is given by

$$M(t) = E(e^{tY_1}) = \int_{-\infty}^{\infty} e^{ty_1}g(y_1)\,dy_1$$

in the continuous case. It would seem that we need to know $g(y_1)$ before we can compute $M(t)$. That this is not the case is a fundamental fact. To see this consider

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\,[tu_1(x_1, \ldots, x_n)]h(x_1, \ldots, x_n)\,dx_1 \cdots dx_n, \qquad (1)$$

which we assume to exist for $-h < t < h$. We shall introduce n new variables of integration. They are $y_1 = u_1(x_1, x_2, \ldots, x_n), \ldots, y_n = u_n(x_1, x_2, \ldots, x_n)$. Momentarily, we assume that these functions define a one-to-one transformation. Let $x_i = w_i(y_1, y_2, \ldots, y_n)$, $i = 1, 2, \ldots, n$, denote the inverse functions and let J denote the Jacobian. Under this transformation, display (1) becomes

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{ty_1}|J|h(w_1, \ldots, w_n)\,dy_2 \cdots dy_n\,dy_1. \qquad (2)$$

In accordance with Section 4.5,

$$|J|h[w_1(y_1, y_2, \ldots, y_n), \ldots, w_n(y_1, y_2, \ldots, y_n)]$$

is the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$. The marginal p.d.f. $g(y_1)$ of Y_1 is obtained by integrating this joint p.d.f. on $y_2, \ldots, y_n$. Since the factor e^{ty_1} does not involve the variables $y_2, \ldots, y_n$, display (2) may be written as

$$\int_{-\infty}^{\infty} e^{ty_1}g(y_1)\,dy_1. \qquad (3)$$

But this is by definition the m.g.f. $M(t)$ of the distribution of Y_1. That is, we can compute $E\{\exp\,[tu_1(X_1, \ldots, X_n)]\}$ and have the value of $E(e^{tY_1})$, where $Y_1 = u_1(X_1, \ldots, X_n)$. This fact provides another technique to help us find the p.d.f. of a function of several random variables. For if the m.g.f. of Y_1 is seen to be that of a certain kind of distribution, the uniqueness property makes it certain that Y_1 has that kind of distribution. When the p.d.f. of Y_1 is obtained in this manner, we say that we use the *moment-generating-function technique*.

The reader will observe that we have assumed the transformation to be one-to-one. We did this for simplicity of presentation. If the transformation is not one-to-one, let

$$x_j = w_{ji}(y_1, \dots, y_n), \qquad j = 1, 2, \dots, n, \quad i = 1, 2, \dots, k,$$

denote the k groups of n inverse functions each. Let J_i, $i = 1, 2, \dots, k$, denote the k Jacobians. Then

$$\sum_{i=1}^{k} |J_i| h[w_{1i}(y_1, \dots, y_n), \dots, w_{ni}(y_1, \dots, y_n)] \tag{4}$$

is the joint p.d.f. of $Y_1, \dots, Y_n$. Then display (1) becomes display (2) with $|J|h(w_1, \dots, w_n)$ replaced by display (4). Hence our result is valid if the transformation is not one-to-one. It seems evident that we can treat the discrete case in an analogous manner with the same result.

It should be noted that the expectation of Y_1 can be computed in like manner. That is,

$$\begin{aligned} E(Y_1) &= \int_{-\infty}^{\infty} y_1 g(y_1)\, dy_1 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_1(x_1, \dots, x_n) h(x_1, \dots, x_n)\, dx_1 \cdots dx_n, \end{aligned}$$

and this fact has been mentioned earlier in the book. Moreover, this holds for the expectation of any function of Y_1, say $w(Y_1)$; that is,

$$\begin{aligned} E[w(Y_1)] &= \int_{-\infty}^{\infty} w(y_1) g(y_1)\, dy_1 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} w[u_1(x_1, \dots, x_n)] h(x_1, \dots, x_n)\, dx_1 \cdots dx_n. \end{aligned}$$

We shall now give some examples and prove some theorems where we use the moment-generating-function technique. In the first example, to emphasize the nature of the problem, we find the distribution of a rather simple statistic both by a direct probabilistic argument and by the moment-generating-function technique.

Example 1. Let the independent random variables X_1 and X_2 have the same p.d.f.

$$\begin{aligned} f(x) &= \frac{x}{6}, \qquad x = 1, 2, 3, \\ &= 0 \qquad \text{elsewhere;} \end{aligned}$$

so the joint p.d.f. of X_1 and X_2 is

$$f(x_1)f(x_2) = \frac{x_1 x_2}{36}, \qquad x_1 = 1, 2, 3, \quad x_2 = 1, 2, 3,$$

$$= 0 \qquad \text{elsewhere.}$$

A probability, such as $\Pr(X_1 = 2, X_2 = 3)$, can be seen immediately to be $(2)(3)/36 = \frac{1}{6}$. However, consider a probability such as $\Pr(X_1 + X_2 = 3)$. The computation can be made by first observing that the event $X_1 + X_2 = 3$ is the union, exclusive of the events with probability zero, of the two mutually exclusive events $(X_1 = 1, X_2 = 2)$ and $(X_1 = 2, X_2 = 1)$. Thus

$$\Pr(X_1 + X_2 = 3) = \Pr(X_1 = 1, X_2 = 2) + \Pr(X_1 = 2, X_2 = 1)$$

$$= \frac{(1)(2)}{36} + \frac{(2)(1)}{36} = \frac{4}{36}.$$

More generally, let y represent any of the numbers 2, 3, 4, 5, 6. The probability of each of the events $X_1 + X_2 = y$, $y = 2, 3, 4, 5, 6$, can be computed as in the case $y = 3$. Let $g(y) = \Pr(X_1 + X_2 = y)$. Then the table

y	2	3	4	5	6
$g(y)$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{10}{36}$	$\frac{12}{36}$	$\frac{9}{36}$

gives the values of $g(y)$ for $y = 2, 3, 4, 5, 6$. For all other values of y, $g(y) = 0$. What we have actually done is to define a new random variable Y by $Y = X_1 + X_2$, and we have found the p.d.f. $g(y)$ of this random variable Y. We shall now solve the same problem, and by the moment-generating-function technique.

Now the m.g.f. of Y is

$$M(t) = E(e^{t(X_1 + X_2)})$$

$$= E(e^{tX_1}e^{tX_2})$$

$$= E(e^{tX_1})E(e^{tX_2}),$$

since X_1 and X_2 are independent. In this example X_1 and X_2 have the same distribution, so they have the same m.g.f.; that is,

$$E(e^{tX_1}) = E(e^{tX_2}) = \tfrac{1}{6}e^t + \tfrac{2}{6}e^{2t} + \tfrac{3}{6}e^{3t}.$$

Thus

$$M(t) = (\tfrac{1}{6}e^t + \tfrac{2}{6}e^{2t} + \tfrac{3}{6}e^{3t})^2$$

$$= \tfrac{1}{36}e^{2t} + \tfrac{4}{36}e^{3t} + \tfrac{10}{36}e^{4t} + \tfrac{12}{36}e^{5t} + \tfrac{9}{36}e^{6t}.$$

This form of $M(t)$ tells us immediately that the p.d.f. $g(y)$ of Y is zero except at $y = 2, 3, 4, 5, 6$, and that $g(y)$ assumes the values $\frac{1}{36}$, $\frac{4}{36}$, $\frac{10}{36}$, $\frac{12}{36}$, $\frac{9}{36}$, respectively, at these points where $g(y) > 0$. This is, of course, the same result that was obtained in the first solution. There appears here to be little, if any, preference for one solution over the other. But in more complicated situations, and particularly with random variables of the continuous type, the moment-generating-function technique can prove very powerful.

Example 2. Let X_1 and X_2 be independent with normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Define the random variable Y by $Y = X_1 - X_2$. The problem is to find $g(y)$, the p.d.f. of Y. This will be done by first finding the m.g.f. of Y. It is

$$\begin{aligned} M(t) &= E(e^{t(X_1 - X_2)}) \\ &= E(e^{tX_1}e^{-tX_2}) \\ &= E(e^{tX_1})E(e^{-tX_2}), \end{aligned}$$

since X_1 and X_2 are independent. It is known that

$$E(e^{tX_1}) = \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right)$$

and that

$$E(e^{tX_2}) = \exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right)$$

for all real t. Then $E(e^{-tX_2})$ can be obtained from $E(e^{tX_2})$ by replacing t by $-t$. That is,

$$E(e^{-tX_2}) = \exp\left(-\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right).$$

Finally, then,

$$\begin{aligned} M(t) &= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right)\exp\left(-\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left((\mu_1 - \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right). \end{aligned}$$

The distribution of Y is completely determined by its m.g.f. $M(t)$, and it is seen that Y has the p.d.f. $g(y)$, which is $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. That is, the difference between two independent, normally distributed, random variables is itself a random variable which is normally distributed with mean equal to the difference of the means (in the order indicated) and the variance equal to the sum of the variances.

The following theorem, which is a generalization of Example 2, is very important in distribution theory.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be independent random variables having, respectively, the normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, ..., and $N(\mu_n, \sigma_n^2)$. The random variable $Y = k_1X_1 + k_2X_2 + \cdots + k_nX_n$, where $k_1, k_2, \ldots, k_n$ are real constants, is normally distributed with mean $k_1\mu_1 + \cdots + k_n\mu_n$ and variance $k_1^2\sigma_1^2 + \cdots + k_n^2\sigma_n^2$. That is, Y is $N\left(\sum_1^n k_i\mu_i, \sum_1^n k_i^2\sigma_i^2\right)$.*

Proof. Because $X_1, X_2, \ldots, X_n$ are independent, the m.g.f. of Y is given by

$$\begin{aligned} M(t) &= E\{\exp[t(k_1X_1 + k_2X_2 + \cdots + k_nX_n)]\} \\ &= E(e^{tk_1X_1})E(e^{tk_2X_2}) \cdots E(e^{tk_nX_n}). \end{aligned}$$

Now

$$E(e^{tX_i}) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right),$$

for all real t, $i = 1, 2, \ldots, n$. Hence we have

$$E(e^{tk_iX_i}) = \exp\left[\mu_i(k_it) + \frac{\sigma_i^2(k_it)^2}{2}\right].$$

That is, the m.g.f. of Y is

$$\begin{aligned} M(t) &= \prod_{i=1}^n \exp\left[(k_i\mu_i)t + \frac{(k_i^2\sigma_i^2)t^2}{2}\right] \\ &= \exp\left[\left(\sum_1^n k_i\mu_i\right)t + \frac{\left(\sum_1^n k_i^2\sigma_i^2\right)t^2}{2}\right]. \end{aligned}$$

But this is the m.g.f. of a distribution that is $N\left(\sum_1^n k_i\mu_i, \sum_1^n k_i^2\sigma_i^2\right)$. This is the desired result.

The next theorem is a generalization of Theorem 1.

Theorem 2. If *$X_1, X_2, \ldots, X_n$ are independent random variables with respective moment-generating functions $M_i(t)$, $i = 1, 2, 3, \ldots, n$,* *then the moment-generating function of*

$$Y = \sum_{i=1}^{n} a_i X_i,$$

where $a_1, a_2, \dots, a_k$ *are real constants, is*

$$M_Y(t) = \prod_{i=1}^{n} M_i(a_i t).$$

Proof. The m.g.f. of Y is given by

$$\begin{aligned} M_Y(t) &= E[e^{tY}] = E[e^{t(a_1X_1 + a_2X_2 + \cdots + a_nX_n)}] \\ &= E[e^{a_1tX_1}e^{a_2tX_2} \cdots e^{a_ntX_n}] \\ &= E[e^{a_1tX_1}]E[e^{a_2tX_2}] \cdots E[e^{a_ntX_n}] \end{aligned}$$

because $X_1, X_2, \dots, X_n$ are independent. However, since

$$E(\ ^{tX_i}) = M_i(t),$$

then

$$E(e^{a_itX_i}) = M_i(a_it).$$

Thus we have that

$$\begin{aligned} M_Y(t) &= M_1(a_1t)M_2(a_2t) \cdots M_n(a_nt) \\ &= \prod_{i=1}^{n} M_i(a_it). \end{aligned}$$

A corollary follows immediately, and it will be used in some important examples.

Corollary. *If* $X_1, X_2, \dots, X_n$ *are observations of a random sample from a distribution with moment-generating function* $M(t)$, *then*

(a) *The moment-generating function of* $Y = \sum_{i=1}^{n} X_i$ *is*

$$M_Y(t) = \prod_{i=1}^{n} M(t) = [M(t)]^n;$$

(b) *The moment-generating function of* $\bar{X} = \sum_{i=1}^{n} (1/n)X_i$ *is*

$$M_{\bar{X}}(t) = \prod_{i=1}^{n} M\left(\frac{t}{n}\right) = \left[M\left(\frac{t}{n}\right)\right]^n.$$

Proof. For (a), let $a_i = 1, i = 1, 2, \dots, n$, in Theorem 2. For (b), take $a_i = 1/n, i = 1, 2, \dots, n$.

The following examples and the exercises give some important applications of Theorem 2 and its corollary.

Example 3. Let $X_1, X_2, \ldots, X_n$ denote the outcomes on n Bernoulli trials. The m.g.f. of X_i, $i = 1, 2, \ldots, n$, is

$$M(t) = 1 - p + pe^t.$$

If $Y = \sum_{i=1}^{n} X_i$, then

$$M_Y(t) = \prod_{i=1}^{n} (1 - p + pe^t) = (1 - p + pe^t)^n.$$

Thus we again see that Y is $b(n, p)$.

Example 4. Let X_1, X_2, X_3 be the observations of a random sample of size $n = 3$ from the exponential distribution having mean β and, of course, m.g.f. $M(t) = 1/(1 - \beta t)$, $t < 1/\beta$. The m.g.f. of $Y = X_1 + X_2 + X_3$ is

$$M_Y(t) = [(1 - \beta t)^{-1}]^3 = (1 - \beta t)^{-3}, \qquad t < 1/\beta,$$

which is that of a gamma distribution with parameters $\alpha = 3$ and β. Thus Y has this distribution. On the other hand, the m.g.f. of $\bar{X}$ is

$$M_{\bar{X}}(t) = \left[\left(1 - \frac{\beta t}{3}\right)^{-1}\right]^3 = \left(1 - \frac{\beta t}{3}\right)^{-3}, \qquad t < 3/\beta;$$

and hence the distribution of $\bar{X}$ is gamma with parameters $\alpha = 3$ and $\beta/3$, respectively.

The next example is so important that we state it as a theorem.

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ be independent variables that have, respectively, the chi-square distributions $\chi^2(r_1), \chi^2(r_2), \ldots,$ and $\chi^2(r_n)$. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has a chi-square distribution with $r_1 + \cdots + r_n$ degrees of freedom; that is, Y is*

$$\chi^2(r_1 + \cdots + r_n).$$

Proof. Since

$$M_i(t) = E(e^{tX_i}) = (1 - 2t)^{-r_i/2}, \qquad t < \tfrac{1}{2}, \quad i = 1, 2, \ldots, n,$$

we have, using Theorem 2 with $a_1 = \cdots = a_n = 1$,

$$M(t) = (1 - 2t)^{-(r_1 + r_2 + \cdots + r_n)/2}, \qquad t < \tfrac{1}{2}.$$

But this is the m.g.f. of a distribution that is $\chi^2(r_1 + r_2 + \cdots + r_n)$. Accordingly, Y has this chi-square distribution.

Next, let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution that is $N(\mu, \sigma^2)$. In accordance with Theorem 2 of

Section 3.4, each of the random variables $(X_i - \mu)^2/\sigma^2$, $i = 1, 2, \ldots, n$, is $\chi^2(1)$. Moreover, these n random variables are independent. Accordingly, by Theorem 3, the random variable $Y = \sum_1^n [(X_i - \mu)/\sigma]^2$ is $\chi^2(n)$. This proves the following theorem.

Theorem 4. *Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that is $N(\mu, \sigma^2)$. The random variable*

$$Y = \sum_1^n \left(\frac{X_i - \mu}{\sigma}\right)^2$$

has a chi-square distribution with n degrees of freedom.

Not always do we sample from a distribution of one random variable. Let the random variables X and Y have the joint p.d.f. $f(x, y)$ and let the $2n$ random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ have the joint p.d.f.

$$f(x_1, y_1)f(x_2, y_2) \cdots f(x_n, y_n).$$

The n random pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are then independent and are said to constitute a *random sample* of size n from the distribution of X and Y. In the next paragraph we shall take $f(x, y)$ to be the normal bivariate p.d.f., and we shall solve a problem in sampling theory when we are sampling from this two-variable distribution.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample of size n from a bivariate normal distribution with p.d.f. $f(x, y)$ and parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. We wish to find the joint p.d.f. of the two statistics $\bar{X} = \sum_1^n X_i/n$ and $\bar{Y} = \sum_1^n Y_i/n$. We call $\bar{X}$ the mean of $X_1, \ldots, X_n$ and $\bar{Y}$ the mean of $Y_1, \ldots, Y_n$. Since the joint p.d.f. of the $2n$ random variables (X_i, Y_i), $i = 1, 2, \ldots, n$, is given by

$$h = f(x_1, y_1)f(x_2, y_2) \cdots f(x_n, y_n),$$

the m.g.f. of the two means $\bar{X}$ and $\bar{Y}$ is given by

$$M(t_1, t_2) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\frac{t_1 \sum_1^n x_i}{n} + \frac{t_2 \sum_1^n y_i}{n}\right) h \, dx_1 \cdots dy_n$$

$$= \prod_{i=1}^n \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(\frac{t_1 x_i}{n} + \frac{t_2 y_i}{n}\right) f(x_i, y_i) \, dx_i \, dy_i\right].$$

The justification of the form of the right-hand member of the second equality is that each pair (X_i, Y_i) has the same p.d.f. and that these n pairs are independent. The twofold integral in the brackets in the last equality is the joint m.g.f. of X_i and Y_i (see Section 3.5) with t_1 replaced by t_1/n and t_2 replaced by t_2/n. Accordingly,

$$M(t_1, t_2) = \prod_{i=1}^{n} \exp\left[\frac{t_1\mu_1}{n} + \frac{t_2\mu_2}{n} + \frac{\sigma_1^2(t_1/n)^2 + 2\rho\sigma_1\sigma_2(t_1/n)(t_2/n) + \sigma_2^2(t_2/n)^2}{2}\right]$$

$$= \exp\left[t_1\mu_1 + t_2\mu_2 + \frac{(\sigma_1^2/n)t_1^2 + 2\rho(\sigma_1\sigma_2/n)t_1t_2 + (\sigma_2^2/n)t_2^2}{2}\right].$$

But this is the m.g.f. of a bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2/n and σ_2^2/n, and correlation coefficient ρ; therefore, $\bar{X}$ and $\bar{Y}$ have this joint distribution.

EXERCISES

4.75. Let the i.i.d. random variables X_1 and X_2 have the same p.d.f. $f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere. Find the p.d.f. of $Y = X_1 + X_2$. Note, under appropriate assumptions, that Y may be interpreted as the sum of the spots that appear when two dice are cast.

4.76. Let X_1 and X_2 be independent with normal distributions $N(6, 1)$ and $N(7, 1)$, respectively. Find $\Pr(X_1 > X_2)$.

Hint: Write $\Pr(X_1 > X_2) = \Pr(X_1 - X_2 > 0)$ and determine the distribution of $X_1 - X_2$.

4.77. Let X_1 and X_2 be independent random variables. Let X_1 and $Y = X_1 + X_2$ have chi-square distributions with r_1 and r degrees of freedom, respectively. Here $r_1 < r$. Show that X_2 has a chi-square distribution with $r - r_1$ degrees of freedom.

Hint: Write $M(t) = E(e^{t(X_1 + X_2)})$ and make use of the independence of X_1 and X_2.

4.78. Let the independent random variables X_1 and X_2 have binomial distributions with parameters $n_1, p_1 = \frac{1}{2}$ and $n_2, p_2 = \frac{1}{2}$, respectively. Show that $Y = X_1 - X_2 + n_2$ has a binomial distribution with parameters $n = n_1 + n_2$, $p = \frac{1}{2}$.

4.79. Let X_1, X_2, X_3 be a random sample of size $n = 3$ from $N(1, 4)$. Compute $P(X_1 + 2X_2 - 2X_3 > 7)$.

4.80. Let X_1 and X_2 be two independent random variables. Let X_1 and $Y = X_1 + X_2$ have Poisson distributions with means μ_1 and $\mu > \mu_1$, respectively. Find the distribution of X_2.

4.81. Let X_1, X_2 be two independent gamma random variables with parameters $\alpha_1 = 3$, $\beta_1 = 3$ and $\alpha_2 = 5$, $\beta_2 = 1$, respectively.
(a) Find the m.g.f. of $Y = 2X_1 + 6X_2$.
(b) What is the distribution of Y?

4.82. A certain job is completed in three steps in series. The means and standard deviations for the steps are (in minutes):

Step	Mean	Standard Deviation
1	17	2
2	13	1
3	13	2

Assuming independent steps and normal distributions, compute the probability that the job will take less than 40 minutes to complete.

4.83. Let X be $N(0, 1)$. Use the moment-generating-function technique to show that $Y = X^2$ is $\chi^2(1)$.

Hint: Evaluate the integral that represents $E(e^{tX^2})$ by writing $w = x\sqrt{1 - 2t}$, $t < \frac{1}{2}$.

4.84. Let $X_1, X_2, \ldots, X_n$ denote n mutually independent random variables with the moment-generating functions $M_1(t), M_2(t), \ldots, M_n(t)$, respectively.
(a) Show that $Y = k_1X_1 + k_2X_2 + \cdots + k_nX_n$, where $k_1, k_2, \ldots, k_n$ are real constants, has the m.g.f. $M(t) = \prod_1^n M_i(k_i t)$.
(b) If each $k_i = 1$ and if X_i is Poisson with mean μ_i, $i = 1, 2, \ldots, n$, prove that Y is Poisson with mean $\mu_1 + \cdots + \mu_n$.

4.85. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with m.g.f. $M(t)$, show that the moment-generating functions of $\sum_1^n X_i$ and $\sum_1^n X_i/n$ are, respectively, $[M(t)]^n$ and $[M(t/n)]^n$.

4.86. In Exercise 4.74 concerning PERT, assume that each of the three independent variables has the p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Find:
(a) The p.d.f. of Y.
(b) The p.d.f. of Z.

4.87. If X and Y have a bivariate normal distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ, show that $Z = aX + bY + c$ is

$$N(a\mu_1 + b\mu_2 + c,\ a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2),$$

where a, b, and c are constants.

Hint: Use the m.g.f. $M(t_1, t_2)$ of X and Y to find the m.g.f. of Z.

4.88. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 25$, $\mu_2 = 35$, $\sigma_1^2 = 4$, $\sigma_2^2 = 16$, and $\rho = \frac{17}{32}$. If $Z = 3X - 2Y$, find $\Pr(-2 < Z < 19)$.

4.89. Let U and V be independent random variables, each having a standard normal distribution. Show that the m.g.f. $E(e^{t(UV)})$ of the product UV is $(1 - t^2)^{-1/2}$, $-1 < t < 1$.

Hint: Compare $E(e^{tUV})$ with the integral of a bivariate normal p.d.f. that has means equal to zero.

4.90. Let X and Y have a bivariate normal distribution with the parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ. Show that

$$W = X - \mu_1 \quad \text{and} \quad Z = (Y - \mu_2) - \rho(\sigma_2/\sigma_1)(X - \mu_1)$$

are independent normal variables.

4.91. Let X_1, X_2, X_3 be a random sample of size $n = 3$ from the standard normal distribution.

(a) Show that $Y_1 = X_1 + \delta X_3$, $Y_2 = X_2 + \delta X_3$ has a bivariate normal distribution.

(b) Find the value of δ so that the correlation coefficient $\rho = \frac{1}{2}$.

(c) What additional transformation involving Y_1 and Y_2 would produce a bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and the same correlation coefficient ρ?

4.92. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. Find the joint distribution of $Y = \sum_1^n a_i X_i$ and $Z = \sum_1^n b_i X_i$, where the a_i and b_i are real constants. When, and only when, are Y and Z independent?

Hint: Note that the joint m.g.f. $E\left[\exp\left(t_1 \sum_1^n a_i X_i + t_2 \sum_1^n b_i X_i\right)\right]$ is that of a bivariate normal distribution.

4.93. Let X_1, X_2 be a random sample of size 2 from a distribution with positive variance and m.g.f. $M(t)$. If $Y = X_1 + X_2$ and $Z = X_1 - X_2$ are independent, prove that the distribution from which the sample is taken is a normal distribution.

Hint: Show that

$$m(t_1, t_2) = E\{\exp[t_1(X_1 + X_2) + t_2(X_1 - X_2)]\} = M(t_1 + t_2)M(t_1 - t_2).$$

Express each member of $m(t_1, t_2) = m(t_1, 0)m(0, t_2)$ in terms of M; differentiate twice with respect to t_2; set $t_2 = 0$; and solve the resulting differential equation in M.

4.8 The Distributions of $\bar{X}$ and nS^2/σ^2

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size $n \geq 2$ from a distribution that is $N(\mu, \sigma^2)$. In this section we shall investigate the distributions of the mean and the variance of this random sample, that is, the distributions of the two statistics $\bar{X} = \sum_1^n X_i/n$ and $S^2 = \sum_1^n (X_i - \bar{X})^2/n$.

The problem of the distribution of $\bar{X}$, the mean of the sample, is solved by the use of Theorem 1 of Section 4.7. We have here, in the notation of the statement of that theorem, $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$, and $k_1 = k_2 = \cdots = k_n = 1/n$. Accordingly, $Y = \bar{X}$ has a normal distribution with mean and variance given by

$$\sum_1^n \left(\frac{1}{n}\mu\right) = \mu, \qquad \sum_1^n \left[\left(\frac{1}{n}\right)^2 \sigma^2\right] = \frac{\sigma^2}{n},$$

respectively. That is, $\bar{X}$ is $N(\mu, \sigma^2/n)$.

Example 1. Let $\bar{X}$ be the mean of a random sample of size 25 from a distribution that is $N(75, 100)$. Thus $\bar{X}$ is $N(75, 4)$. Then, for instance,

$$\begin{aligned} \Pr(71 < \bar{X} < 79) &= \Phi\left(\frac{79-75}{2}\right) - \Phi\left(\frac{71-75}{2}\right) \\ &= \Phi(2) - \Phi(-2) = 0.954. \end{aligned}$$

We now take up the problem of the distribution of S^2, the variance of a random sample $X_1, \ldots, X_n$ from a distribution that is $N(\mu, \sigma^2)$. To do this, let us first consider the joint distribution of $Y_1 = \bar{X}$, $Y_2 = X_2 - \bar{X}$, $Y_3 = X_3 - \bar{X}, \ldots, Y_n = X_n - \bar{X}$. The corresponding inverse transformation

$$\begin{aligned} x_1 &= y_1 - y_2 - y_3 - \cdots - y_n \\ x_2 &= y_1 + y_2 \\ x_3 &= y_1 + y_3 \\ &\vdots \qquad \vdots \\ x_n &= y_1 + y_n \end{aligned}$$

has Jacobian n. Since

$$\begin{aligned} \sum_1^n (x_i - \mu)^2 &= \sum_1^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_1^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \end{aligned}$$

because $2(\bar{x}-\mu)\sum_1^n (x_i - \bar{x}) = 0$, the joint p.d.f. of $X_1, X_2, \ldots, X_n$ can be written

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{\sum (x_i-\bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right],$$

where $\bar{x}$ represents $(x_1 + x_2 + \cdots + x_n)/n$ and $-\infty < x_i < \infty$, $i = 1, 2, \ldots, n$. Accordingly, with $y_1 = \bar{x}$ and $x_1 - \bar{x} = -y_2 - y_3 - \cdots - y_n$, we find that the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is

$$(n)\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{(-y_2 - \cdots - y_n)^2}{2\sigma^2} - \frac{\sum_2^n y_i^2}{2\sigma^2} - \frac{n(y_1-\mu)^2}{2\sigma^2}\right],$$

$-\infty < y_i < \infty$, $i = 1, 2, \ldots, n$. Note that this is the product of the p.d.f. of Y_1, namely,

$$\frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\frac{(y_1-\mu)^2}{2\sigma^2/n}\right], \qquad -\infty < y_1 < \infty,$$

and a function of $y_2, \ldots, y_n$. Thus Y_1 must be independent of the $n-1$ random variables $Y_2, Y_3, \ldots, Y_n$ and that function of $y_2, \ldots, y_n$ is the joint p.d.f. of $Y_2, Y_3, \ldots, Y_n$. Moreover, this means that $Y_1 = \bar{X}$ and thus

$$\frac{n(Y_1-\mu)^2}{\sigma^2} = \frac{n(\bar{X}-\mu)^2}{\sigma^2} = W_1$$

are independent of

$$\frac{(-Y_2 - \cdots - Y_n)^2 + \sum_2^n Y_i^2}{\sigma^2} = \frac{\sum_1^n (X_i - \bar{X})^2}{\sigma^2} = W_2.$$

Since W_1 is the square of a standard normal variable, it is distributed as $\chi^2(1)$. Also, we know that

$$W = \sum_1^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = W_1 + W_2$$

is $\chi^2(n)$. From the independence of W_1 and W_2, we have

$$E(e^{tW}) = E(e^{tW_1})E(e^{tW_2})$$

or, equivalently,

$$(1-2t)^{-n/2} = (1-2t)^{-1/2}E(e^{tW_2}), \qquad t < \tfrac{1}{2}.$$

Thus

$$E(e^{tW_2}) = (1 - 2t)^{-(n-1)/2}, \qquad t < \tfrac{1}{2},$$

and hence $W_2 = nS^2/\sigma^2$ is $\chi^2(n-1)$. The determination of the p.d.f. of S^2 is an easy exercise from this result (see Exercise 4.99).

To summarize, we have established, in this section, three important properties of $\bar{X}$ and S^2 when the sample arises from a distribution which is $N(\mu, \sigma^2)$:

1. $\bar{X}$ is $N(\mu, \sigma^2/n)$.
2. nS^2/σ^2 is $\chi^2(n-1)$.
3. $\bar{X}$ and S^2 are independent.

For illustration, as the result of properties (1), (2), and (3), we have that $\sqrt{n}(\bar{X} - \mu)/\sigma$ is $N(0, 1)$. Thus, from the definition of Student's t,

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{nS^2/\sigma^2(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n-1}}$$

has a t-distribution with $n - 1$ degrees of freedom. It was a random variable like this one that motivated Gosset's search for the distribution of T. This t-statistic will play an important role in statistical applications.

EXERCISES

4.94. Let $\bar{X}$ be the mean of a random sample of size 5 from a normal distribution with $\mu = 0$ and $\sigma^2 = 125$. Determine c so that $\Pr(\bar{X} < c) = 0.90$.

4.95. If $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ and variance 100, find n so that $\Pr(\mu - 5 < \bar{X} < \mu + 5) = 0.954$.

4.96. Let $X_1, X_2, \ldots, X_{25}$ and $Y_1, Y_2, \ldots, Y_{25}$ be two independent random samples from two normal distributions $N(0, 16)$ and $N(1, 9)$, respectively. Let $\bar{X}$ and $\bar{Y}$ denote the corresponding sample means. Compute $\Pr(\bar{X} > \bar{Y})$.

4.97. Find the mean and variance of $S^2 = \sum_{1}^{n} (X_i - \bar{X})^2/n$, where $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$.

Hint: Find the mean and variance of nS^2/σ^2.

4.98. Let S^2 be the variance of a random sample of size 6 from the normal distribution $N(\mu, 12)$. Find $\Pr(2.30 < S^2 < 22.2)$.

4.99. Find the p.d.f. of the sample variance $V = S^2$, provided that the distribution from which the sample arises is $N(\mu, \sigma^2)$.

4.100. Let $\bar{X}$ and $\bar{Y}$ be the respective means of two independent random samples, each of size 4, from the two respective normal distributions $N(10, 9)$ and $N(3, 4)$. Compute $\Pr(\bar{X} > 2\bar{Y})$.

4.101. Let $X_1, X_2, \ldots, X_5$ be a random sample of size $n = 5$ from $N(0, \sigma^2)$. (a) Find the constant c so that $c(X_1 - X_2)/\sqrt{X_3^2 + X_4^2 + X_5^2}$ has a t-distribution. (b) How many degrees of freedom are associated with this T?

4.102. If a random sample of size 2 is taken from a normal distribution with mean 7 and variance 8, find the probability that the absolute value of the difference of these two observations exceeds 2.

4.103. Let $\bar{X}$ and S^2 be the mean and the variance of a random sample of size 25 from a distribution that is $N(3, 100)$. Then evaluate $\Pr(0 < \bar{X} < 6, 55.2 < S^2 < 145.6)$.

4.9 Expectations of Functions of Random Variables

Let $X_1, X_2, \ldots, X_n$ denote random variables that have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$. Let the random variable Y be defined by $Y = u(X_1, X_2, \ldots, X_n)$. In Section 4.7, we found that we could compute expectations of functions of Y without first finding the p.d.f. of Y. Indeed, this fact was the basis of the moment-generating-function procedure for finding the p.d.f. of Y. We can take advantage of this fact in a number of other instances. Some illustrative examples will be given.

Example 1. Say that W is $N(0, 1)$, that V is $\chi^2(r)$ with $r \geq 2$, and that W and V are independent. The mean of the random variable $T = W\sqrt{r/V}$ exists and is zero because the graph of the p.d.f. of T (see Section 4.4) is symmetric about the vertical axis through $t = 0$. The variance of T, when it exists, could be computed by integrating the product of t^2 and the p.d.f. of T. But it seems much simpler to compute

$$\sigma_T^2 = E(T^2) = E\left(W^2 \frac{r}{V}\right) = E(W^2)E\left(\frac{r}{V}\right).$$

Now W^2 is $\chi^2(1)$, so $E(W^2) = 1$. Furthermore,

$$E\left(\frac{r}{V}\right) = \int_0^\infty \frac{r}{v} \frac{1}{2^{r/2}\Gamma(r/2)} v^{r/2-1}e^{-v/2}\,dv$$

exists if $r > 2$ and is given by

$$\frac{r\Gamma[(r-2)/2]}{2\Gamma(r/2)} = \frac{r\Gamma[(r-2)/2]}{2[(r-2)/2]\Gamma[(r-2)/2]} = \frac{r}{r-2}.$$

Thus $\sigma_T^2 = r/(r-2)$, $r > 2$.

Example 2. Let X_i denote a random variable with mean μ_i and variance σ_i^2, $i = 1, 2, \ldots, n$. Let $X_1, X_2, \ldots, X_n$ be independent and let $k_1, k_2, \ldots, k_n$ denote real constants. We shall compute the mean and variance of a linear function $Y = k_1X_1 + k_2X_2 + \cdots + k_nX_n$. Because E is a linear operator, the mean of Y is given by

$$\begin{aligned}\mu_Y &= E(k_1X_1 + k_2X_2 + \cdots + k_nX_n)\\ &= k_1E(X_1) + k_2E(X_2) + \cdots + k_nE(X_n)\\ &= k_1\mu_1 + k_2\mu_2 + \cdots + k_n\mu_n = \sum_1^n k_i\mu_i.\end{aligned}$$

The variance of Y is given by

$$\begin{aligned}\sigma_Y^2 &= E\{[(k_1X_1 + \cdots + k_nX_n) - (k_1\mu_1 + \cdots + k_n\mu_n)]^2\}\\ &= E\{[k_1(X_1 - \mu_1) + \cdots + k_n(X_n - \mu_n)]^2\}\\ &= E\left\{\sum_{i=1}^n k_i^2(X_i - \mu_i)^2 + 2\sum_{i<j}\sum k_ik_j(X_i - \mu_i)(X_j - \mu_j)\right\}\\ &= \sum_{i=1}^n k_i^2E[(X_i - \mu_i)^2] + 2\sum_{i<j}\sum k_ik_jE[(X_i - \mu_i)(X_j - \mu_j)].\end{aligned}$$

Consider $E[(X_i - \mu_i)(X_j - \mu_j)]$, $i < j$. Because X_i and X_j are independent, we have

$$E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i - \mu_i)E(X_j - \mu_j) = 0.$$

Finally, then,

$$\sigma_Y^2 = \sum_{i=1}^n k_i^2E[(X_i - \mu_i)^2] = \sum_{i=1}^n k_i^2\sigma_i^2.$$

We can obtain a more general result if, in Example 2, we remove the hypothesis of independence of $X_1, X_2, \ldots, X_n$. We shall do this and we shall let ρ_{ij} denote the correlation coefficient of X_i and X_j. Thus for easy reference to Example 2, we write

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \rho_{ij}\sigma_i\sigma_j, \qquad i < j.$$

If we refer to Example 2, we see that again $\mu_Y = \sum_1^n k_i\mu_i$. But now

$$\sigma_Y^2 = \sum_1^n k_i^2\sigma_i^2 + 2\sum_{i<j}\sum k_ik_j\rho_{ij}\sigma_i\sigma_j.$$

Thus we have the following theorem.

Theorem 5. *Let $X_1, \ldots, X_n$ denote random variables that have means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$. Let ρ_{ij}, $i \neq j$, denote the correlation coefficient of X_i and X_j and let $k_1, \ldots, k_n$ denote real constants. The mean and the variance of the linear function*

$$Y = \sum_1^n k_iX_i$$

are, respectively,

$$\mu_Y = \sum_1^n k_i\mu_i$$

and

$$\sigma_Y^2 = \sum_1^n k_i^2\sigma_i^2 + 2\sum_{i<j}\sum k_ik_j\rho_{ij}\sigma_i\sigma_j.$$

The following corollary of this theorem is quite useful.

Corollary. *Let $X_1, \ldots, X_n$ denote the observations of a random sample of size n from a distribution that has mean μ and variance σ^2. The mean and the variance of $Y = \sum_1^n k_iX_i$ are, respectively, $\mu_Y = \left(\sum_1^n k_i\right)\mu$ and $\sigma_Y^2 = \left(\sum_1^n k_i^2\right)\sigma^2$.*

Example 3. Let $\bar{X} = \sum_1^n X_i/n$ denote the mean of a random sample of size n from a distribution that has mean μ and variance σ^2. In accordance with the corollary, we have $\mu_{\bar{X}} = \mu \sum_1^n (1/n) = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2 \sum_1^n (1/n)^2 = \sigma^2/n$. We have seen, in Section 4.8, that if our sample is from a distribution that is $N(\mu, \sigma^2)$, then $\bar{X}$ is $N(\mu, \sigma^2/n)$. It is interesting that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2/n$ whether the sample is or is not from a normal distribution.

EXERCISES

4.104. Let X_1, X_2, X_3, X_4 be four i.i.d. random variables having the same p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Find the mean and variance of the sum Y of these four random variables.

4.105. Let X_1 and X_2 be two independent random variables so that the variances of X_1 and X_2 are $\sigma_1^2 = k$ and $\sigma_2^2 = 2$, respectively. Given that the variance of $Y = 3X_2 - X_1$ is 25, find k.

4.106. If the independent variables X_1 and X_2 have means μ_1, μ_2 and variances σ_1^2, σ_2^2, respectively, show that the mean and variance of the product $Y = X_1X_2$ are $\mu_1\mu_2$ and $\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2$, respectively.

4.107. Find the mean and variance of the sum Y of the observations of a random sample of size 5 from the distribution having p.d.f. $f(x) = 6x(1 - x)$, $0 < x < 1$, zero elsewhere.

4.108. Determine the mean and variance of the mean $\bar{X}$ of a random sample of size 9 from a distribution having p.d.f. $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

4.109. Let X and Y be random variables with $\mu_1 = 1$, $\mu_2 = 4$, $\sigma_1^2 = 4$, $\sigma_2^2 = 6$, $\rho = \frac{1}{2}$. Find the mean and variance of $Z = 3X - 2Y$.

4.110. Let X and Y be independent random variables with means μ_1, μ_2 and variances σ_1^2, σ_2^2. Determine the correlation coefficient of X and $Z = X - Y$ in terms of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$.

4.111. Let μ and σ^2 denote the mean and variance of the random variable X. Let $Y = c + bX$, where b and c are real constants. Show that the mean and the variance of Y are, respectively, $c + b\mu$ and $b^2\sigma^2$.

4.112. Find the mean and the variance of $Y = X_1 - 2X_2 + 3X_3$, where X_1, X_2, X_3 are observations of a random sample from a chi-square distribution with 6 degrees of freedom.

4.113. Let X and Y be random variables such that var $(X) = 4$, var $(Y) = 2$, and var $(X + 2Y) = 15$. Determine the correlation coefficient of X and Y.

4.114. Let X and Y be random variables with means μ_1, μ_2; variances σ_1^2, σ_2^2; and correlation coefficient ρ. Show that the correlation coefficient of $W = aX + b$, $a > 0$, and $Z = cY + d$, $c > 0$, is ρ.

4.115. A person rolls a die, tosses a coin, and draws a card from an ordinary deck. He receives \$3 for each point up on the die, \$10 for a head, \$0 for a tail, and \$1 for each spot on the card (jack = 11, queen = 12, king = 13). If we assume that the three random variables involved are independent and uniformly distributed, compute the mean and variance of the amount to be received.

4.116. Let U and V be two independent chi-square variables with r_1 and r_2 degrees of freedom, respectively. Find the mean and variance of $F = (r_2U)/(r_1V)$. What restriction is needed on the parameters r_1 and r_2 in order to ensure the existence of both the mean and the variance of F?

4.117. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with mean μ and variance σ^2. Show that $E(S^2) = (n-1)\sigma^2/n$, where S^2 is the variance of the random sample.

Hint: Write $S^2 = (1/n)\sum_1^n (X_i - \mu)^2 - (\bar{X} - \mu)^2$.

4.118. Let X_1 and X_2 be independent random variables with nonzero variances. Find the correlation coefficient of $Y = X_1X_2$ and X_1 in terms of the means and variances of X_1 and X_2.

4.119. Let X_1 and X_2 have a joint distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. Find the correlation coefficient of the linear functions $Y = a_1X_1 + a_2X_2$ and $Z = b_1X_1 + b_2X_2$ in terms of the real constants a_1, a_2, b_1, b_2, and the parameters of the distribution.

4.120. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution which has mean μ and variance σ^2. Use Chebyshev's inequality to show, for every $\epsilon > 0$, that $\lim_{n\to\infty} \Pr(|\bar{X} - \mu| < \epsilon) = 1$; this is another form of the law of large numbers.

4.121. Let X_1, X_2, and X_3 be random variables with equal variances but with correlation coefficients $\rho_{12} = 0.3$, $\rho_{13} = 0.5$, and $\rho_{23} = 0.2$. Find the correlation coefficient of the linear functions $Y = X_1 + X_2$ and $Z = X_2 + X_3$.

4.122. Find the variance of the sum of 10 random variables if each has variance 5 and if each pair has correlation coefficient 0.5.

4.123. Let X and Y have the parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. Show that the correlation coefficient of X and $[Y - \rho(\sigma_2/\sigma_1)X]$ is zero.

4.124. Let X_1 and X_2 have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. Compute the means, the variances, and the correlation coefficient of $Y_1 = \exp(X_1)$ and $Y_2 = \exp(X_2)$.

Hint: Various moments of Y_1 and Y_2 can be found by assigning appropriate values to t_1 and t_2 in $E[\exp(t_1X_1 + t_2X_2)]$.

4.125. Let X be $N(\mu, \sigma^2)$ and consider the transformation $X = \ln Y$ or, equivalently, $Y = e^X$.

(a) Find the mean and the variance of Y by first determining $E(e^X)$ and $E[(e^X)^2]$.
Hint: Use the m.g.f. of X.

(b) Find the p.d.f. of Y. This is the p.d.f. of the *lognormal distribution*.

4.126. Let X_1 and X_2 have a trinomial distribution with parameters n, p_1, p_2.

(a) What is the distribution of $Y = X_1 + X_2$?

(b) From the equality $\sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$, once again determine the correlation coefficient ρ of X_1 and X_2.

4.127. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_2 + X_3$, where X_1, X_2, and X_3 are three independent random variables. Find the joint m.g.f. and the correlation coefficient of Y_1 and Y_2 provided that:
(a) X_i has a Poisson distribution with mean μ_i, $i = 1, 2, 3$.
(b) X_i is $N(\mu_i, \sigma_i^2)$, $i = 1, 2, 3$.

4.128. Let $X_1, \ldots, X_n$ be random variables that have means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$. Let ρ_{ij}, $i \neq j$, denote the correlation coefficient of X_i and X_j. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be real constants. Show that the covariance of $Y = \sum_{i=1}^{n} a_i X_i$ and $Z = \sum_{j=1}^{n} b_j X_j$ is $\sum_{j=1}^{n} \sum_{i=1}^{n} a_i b_j \sigma_i \sigma_j \rho_{ij}$, where $\rho_{ii} = 1$, $i = 1, 2, \ldots, n$.

*4.10 The Multivariate Normal Distribution

We have studied in some detail normal distributions of one random variable. In this section we investigate a joint distribution of n random variables that will be called a *multivariate normal* distribution. This investigation assumes that the student is familiar with elementary matrix algebra, with real symmetric quadratic forms, and with orthogonal transformations. Henceforth, the expression *quadratic form* means a quadratic form in a prescribed number of variables whose matrix is real and symmetric. All symbols that represent matrices will be set in boldface type.

Let $\mathbf{A}$ denote an $n \times n$ real symmetric matrix which is positive definite. Let $\boldsymbol{\mu}$ denote the $n \times 1$ matrix such that $\boldsymbol{\mu}'$, the transpose of $\boldsymbol{\mu}$, is $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_n]$, where each μ_i is a real constant. Finally, let $\mathbf{x}$ denote the $n \times 1$ matrix such that $\mathbf{x}' = [x_1, x_2, \ldots, x_n]$. We shall show that if C is an appropriately chosen positive constant, the nonnegative function

$$f(x_1, x_2, \ldots, x_n) = C \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})}{2} \right],$$

$$-\infty < x_i < \infty, \quad i = 1, 2, \ldots, n,$$

is a joint p.d.f. of n random variables $X_1, X_2, \ldots, X_n$ that are of the continuous type. Thus we need to show that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n = 1. \tag{1}$$

Let $\mathbf{t}$ denote the $n \times 1$ matrix such that $\mathbf{t}' = [t_1, t_2, \dots, t_n]$, where $t_1, t_2, \dots, t_n$ are arbitrary real numbers. We shall evaluate the integral

$$C \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}'\mathbf{x} - \frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})}{2}\right] dx_1 \cdots dx_n, \qquad (2)$$

and then we shall subsequently set $t_1 = t_2 = \cdots = t_n = 0$, and thus establish Equation (1). First, we change the variables of integration in integral (2) from $x_1, x_2, \dots, x_n$ to $y_1, y_2, \dots, y_n$ by writing $\mathbf{x} - \boldsymbol{\mu} = \mathbf{y}$, where $\mathbf{y}' = [y_1, y_2, \dots, y_n]$. The Jacobian of the transformation is one and the n-dimensional x-space is mapped onto an n-dimensional y-space, so that integral (2) may be written as

$$C \exp(\mathbf{t}'\boldsymbol{\mu}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\mathbf{t}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{2}\right) dy_1 \cdots dy_n. \qquad (3)$$

Because the real symmetric matrix $\mathbf{A}$ is positive definite, the n characteristic numbers (proper values, latent roots, or eigenvalues) $a_1, a_2, \dots, a_n$ of $\mathbf{A}$ are positive. There exists an appropriately chosen $n \times n$ real orthogonal matrix $\mathbf{L}$ ($\mathbf{L}' = \mathbf{L}^{-1}$, where $\mathbf{L}^{-1}$ is the inverse of $\mathbf{L}$) such that

$$\mathbf{L}'\mathbf{A}\mathbf{L} = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_n \end{bmatrix},$$

for a suitable ordering of $a_1, a_2, \dots, a_n$. We shall sometimes write $\mathbf{L}'\mathbf{A}\mathbf{L} = \text{diag}[a_1, a_2, \dots, a_n]$. In integral (3), we shall change the variables of integration from $y_1, y_2, \dots, y_n$ to $z_1, z_2, \dots, z_n$ by writing $\mathbf{y} = \mathbf{L}\mathbf{z}$, where $\mathbf{z}' = [z_1, z_2, \dots, z_n]$. The Jacobian of the transformation is the determinant of the orthogonal matrix $\mathbf{L}$. Since $\mathbf{L}'\mathbf{L} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the unit matrix of order n, we have the determinant $|\mathbf{L}'\mathbf{L}| = 1$ and $|\mathbf{L}|^2 = 1$. Thus the absolute value of the Jacobian is one. Moreover, the n-dimensional y-space is mapped onto an n-dimensional z-space. The integral (3) becomes

$$C \exp(\mathbf{t}'\boldsymbol{\mu}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}'\mathbf{L}\mathbf{z} - \frac{\mathbf{z}'(\mathbf{L}'\mathbf{A}\mathbf{L})\mathbf{z}}{2}\right] dz_1 \cdots dz_n. \qquad (4)$$

It is computationally convenient to write, momentarily, $\mathbf{t}'\mathbf{L} = \mathbf{w}'$, where $\mathbf{w}' = [w_1, w_2, \dots, w_n]$. Then

$$\exp[\mathbf{t}'\mathbf{L}\mathbf{z}] = \exp[\mathbf{w}'\mathbf{z}] = \exp\left(\sum_1^n w_i z_i\right).$$

Moreover,

$$\exp\left[-\frac{\mathbf{z}'(\mathbf{L}'\mathbf{A}\mathbf{L})\mathbf{z}}{2}\right] = \exp\left[-\frac{\sum_1^n a_i z_i^2}{2}\right].$$

Then integral (4) may be written as the product of n integrals in the following manner:

$$C \exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu}) \prod_{i=1}^n \left[\int_{-\infty}^{\infty} \exp\left(w_i z_i - \frac{a_i z_i^2}{2}\right) dz_i\right]$$

$$= C \exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu}) \prod_{i=1}^n \left[\sqrt{\frac{2\pi}{a_i}} \int_{-\infty}^{\infty} \frac{\exp\left(w_i z_i - \frac{a_i z_i^2}{2}\right)}{\sqrt{2\pi/a_i}} dz_i\right]. \quad (5)$$

The integral that involves z_i can be treated as the m.g.f., with the more familiar symbol t replaced by w_i, of a distribution which is $N(0, 1/a_i)$. Thus the right-hand member of Equation (5) is equal to

$$C \exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu}) \prod_{i=1}^n \left[\sqrt{\frac{2\pi}{a_i}} \exp\left(\frac{w_i^2}{2a_i}\right)\right]$$

$$= C \exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu}) \sqrt{\frac{(2\pi)^n}{a_1 a_2 \cdots a_n}} \exp\left(\sum_1^n \frac{w_i^2}{2a_i}\right). \quad (6)$$

Now, because $\mathbf{L}^{-1} = \mathbf{L}'$, we have

$$(\mathbf{L}'\mathbf{A}\mathbf{L})^{-1} = \mathbf{L}'\mathbf{A}^{-1}\mathbf{L} = \operatorname{diag}\left[\frac{1}{a_1}, \frac{1}{a_2}, \ldots, \frac{1}{a_n}\right].$$

Thus

$$\sum_1^n \frac{w_i^2}{a_i} = \mathbf{w}'(\mathbf{L}'\mathbf{A}^{-1}\mathbf{L})\mathbf{w} = (\mathbf{L}\mathbf{w})'\mathbf{A}^{-1}(\mathbf{L}\mathbf{w}) = \mathbf{t}'\mathbf{A}^{-1}\mathbf{t}.$$

Moreover, the determinant $|\mathbf{A}^{-1}|$ of $\mathbf{A}^{-1}$ is

$$|\mathbf{A}^{-1}| = |\mathbf{L}'\mathbf{A}^{-1}\mathbf{L}| = \frac{1}{a_1 a_2 \cdots a_n}.$$

Accordingly, the right-hand member of Equation (6), which is equal to integral (2), may be written as

$$Ce^{\mathbf{t}'\boldsymbol{\mu}}\sqrt{(2\pi)^n|\mathbf{A}^{-1}|} \exp\left(\frac{\mathbf{t}'\mathbf{A}^{-1}\mathbf{t}}{2}\right). \quad (7)$$

If, in this function, we set $t_1 = t_2 = \cdots = t_n = 0$, we have the value of the left-hand member of Equation (1). Thus we have

$$C\sqrt{(2\pi)^n|\mathbf{A}^{-1}|} = 1.$$

Accordingly, the function

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{A}^{-1}|}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})}{2}\right],$$

$-\infty < x_i < \infty$, $i = 1, 2, \ldots, n$, is a joint p.d.f. of n random variables $X_1, X_2, \ldots, X_n$ that are of the continuous type. Such a p.d.f. is called a *nonsingular multivariate normal* p.d.f.

We have now proved that $f(x_1, x_2, \ldots, x_n)$ is a p.d.f. However, we have proved more than that. Because $f(x_1, x_2, \ldots, x_n)$ is a p.d.f., integral (2) is the m.g.f. $M(t_1, t_2, \ldots, t_n)$ of this joint distribution of probability. Since integral (2) is equal to function (7), the m.g.f. of the multivariate normal distribution is given by

$$M(t_1, t_2, \ldots, t_n) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\mathbf{A}^{-1}\mathbf{t}}{2}\right).$$

Let the elements of the real, symmetric, and positive definite matrix $\mathbf{A}^{-1}$ be denoted by σ_{ij}, $i, j = 1, 2, \ldots, n$. Then

$$M(0, \ldots, 0, t_i, 0, \ldots, 0) = \exp\left(t_i\mu_i + \frac{\sigma_{ii}t_i^2}{2}\right)$$

is the m.g.f. of X_i, $i = 1, 2, \ldots, n$. Thus X_i is $N(\mu_i, \sigma_{ii})$, $i = 1, 2, \ldots, n$. Moreover, with $i \neq j$, we see that $M(0, \ldots, 0, t_i, 0, \ldots, t_j, 0, \ldots, 0)$, the m.g.f. of X_i and X_j, is equal to

$$\exp\left(t_i\mu_i + t_j\mu_j + \frac{\sigma_{ii}t_i^2 + 2\sigma_{ij}t_it_j + \sigma_{jj}t_j^2}{2}\right),$$

which is the m.g.f. of a *bivariate normal distribution*. In Exercise 4.131 the reader is asked to show that σ_{ij} is the covariance of the random variables X_i and X_j. Thus the matrix $\boldsymbol{\mu}$, where $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_n]$, is the matrix of the means of the random variables $X_1, \ldots, X_n$. Moreover, the elements on the principal diagonal of $\mathbf{A}^{-1}$ are, respectively, the variances $\sigma_{ii} = \sigma_i^2$, $i = 1, 2, \ldots, n$, and the elements not on the principal diagonal of $\mathbf{A}^{-1}$ are, respectively, the covariances

$\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, $i \neq j$, of the random variables $X_1, X_2, \ldots, X_n$. We call the matrix $\mathbf{A}^{-1}$, which is given by

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{nn} \end{bmatrix},$$

the *covariance matrix* of the multivariate normal distribution and henceforth we shall denote this matrix by the symbol $\mathbf{V}$. In terms of the positive definite covariance matrix $\mathbf{V}$, the multivariate normal p.d.f. is written

$$\frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{V}|}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right], \qquad -\infty < x_i < \infty,$$

$i = 1, 2, \ldots, n$, and the m.g.f. of this distribution is given by

$$\exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right)$$

for all real values of $\mathbf{t}$.

Note that this m.g.f. equals the product of n functions, where the first is a function of t_1 alone, the second is a function of t_2 alone, and so on, if and only if $\mathbf{V}$ is a diagonal matrix. This condition, $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j = 0$, means $\rho_{ij} = 0$, $i \neq j$. That is, the multivariate normal random variables are independent if and only if $\rho_{ij} = 0$ for all $i \neq j$.

Example 1. Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with matrix $\boldsymbol{\mu}$ of means and positive definite covariance matrix $\mathbf{V}$. If we let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$, then the m.g.f. $M(t_1, t_2, \ldots, t_n)$ of this joint distribution of probability is

$$E(e^{\mathbf{t}'\mathbf{X}}) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right). \tag{8}$$

Consider a linear function Y of $X_1, X_2, \ldots, X_n$ which is defined by $Y = \mathbf{c}'\mathbf{X} = \sum_1^n c_iX_i$, where $\mathbf{c}' = [c_1, c_2, \ldots, c_n]$ and the several c_i are real and not all zero. We wish to find the p.d.f. of Y. The m.g.f. $m(t)$ of the distribution of Y is given by

$$m(t) = E(e^{tY}) = E(e^{t\mathbf{c}'\mathbf{X}}).$$

Now the expectation (8) exists for all real values of $\mathbf{t}$. Thus we can replace $\mathbf{t}'$ in expectation (8) by $t\mathbf{c}'$ and obtain

$$m(t) = \exp\left(t\mathbf{c}'\boldsymbol{\mu} + \frac{\mathbf{c}'\mathbf{V}\mathbf{c}t^2}{2}\right).$$

Thus the random variable Y is $N(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{V}\mathbf{c})$.

EXERCISES

4.129. Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with positive definite covariance matrix $\mathbf{V}$. Prove that these random variables are mutually independent if and only if $\mathbf{V}$ is a diagonal matrix.

4.130. Let $n = 2$ and take

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Determine $|\mathbf{V}|$, $\mathbf{V}^{-1}$, and $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})$. Compare the bivariate normal p.d.f. of Section 3.5 with this multivariate normal p.d.f. when $n = 2$.

4.131. Let $m(t_i, t_j)$ represent the m.g.f. of X_i and X_j as given in the text. Show that

$$\frac{\partial^2 m(0, 0)}{\partial t_i \partial t_j} - \left[\frac{\partial m(0, 0)}{\partial t_i}\right]\left[\frac{\partial m(0, 0)}{\partial t_j}\right] = \sigma_{ij};$$

that is, prove that the covariance of X_i and X_j is σ_{ij}, which appears in that formula for $m(t_i, t_j)$.

4.132. Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution, where $\boldsymbol{\mu}$ is the matrix of the means and $\mathbf{V}$ is the positive definite covariance matrix. Let $Y = \mathbf{c}'\mathbf{X}$ and $Z = \mathbf{d}'\mathbf{X}$, where $\mathbf{X}' = [X_1, \ldots, X_n]$, $\mathbf{c}' = [c_1, \ldots, c_n]$, and $\mathbf{d}' = [d_1, \ldots, d_n]$ are real matrices.

(a) Find $m(t_1, t_2) = E(e^{t_1 Y + t_2 Z})$ to see that Y and Z have a bivariate normal distribution.

(b) Prove that Y and Z are independent if and only if $\mathbf{c}'\mathbf{V}\mathbf{d} = 0$.

(c) If $X_1, X_2, \ldots, X_n$ are independent random variables which have the same variance σ^2, show that the necessary and sufficient condition of part (b) becomes $\mathbf{c}'\mathbf{d} = 0$.

4.133. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$ have the multivariate normal distribution of Exercise 4.132. Consider the p linear functions of $X_1, \ldots, X_n$ defined by $\mathbf{W} = \mathbf{B}\mathbf{X}$, where $\mathbf{W}' = [W_1, \ldots, W_p]$, $p \le n$, and $\mathbf{B}$ is a $p \times n$ real matrix of rank p. Find $m(v_1, \ldots, v_p) = E(e^{\mathbf{v}'\mathbf{W}})$, where $\mathbf{v}'$ is a real matrix $[v_1, \ldots, v_p]$, to see that $W_1, \ldots, W_p$ have a p-variate normal distribution which has $\mathbf{B}\boldsymbol{\mu}$ for the matrix of the means and $\mathbf{B}\mathbf{V}\mathbf{B}'$ for the covariance matrix.

4.134. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$ have the n-variate normal distribution of Exercise 4.132. Show that $X_1, X_2, \ldots, X_p$, $p < n$, have a p-variate normal distribution. What submatrix of $\mathbf{V}$ is the covariance matrix of $X_1, X_2, \ldots, X_p$?

Hint: In the m.g.f. $M(t_1, t_2, \ldots, t_n)$ of $X_1, X_2, \ldots, X_n$, let $t_{p+1} = \cdots = t_n = 0$.

ADDITIONAL EXERCISES

4.135. If X has the p.d.f. $f(x) = \frac{1}{3}$, $-1 < x < 2$, zero elsewhere, find the p.d.f. of $Y = X^4$.

4.136. The continuous random variable X has a p.d.f. given by $f(x) = 1$, $0 < x < 1$, zero elsewhere. The random variable Y is such that $Y = -2 \ln X$. What is the distribution of Y? What are the mean and the variance of Y?

4.137. Let X_1, X_2 be a random sample of size $n = 2$ from a Poisson distribution with mean μ. If $\Pr(X_1 + X_2 = 3) = (\frac{32}{3})e^{-4}$, compute $\Pr(X_1 = 2, X_2 = 4)$.

4.138. Let $X_1, X_2, \ldots, X_{25}$ be a random sample of size $n = 25$ from a distribution with p.d.f. $f(x) = 3/x^4$, $1 < x < \infty$, zero elsewhere. Let Y equal the number of these X values less than or equal to 2. What is the distribution of Y?

4.139. Find the probability that the range of a random sample of size 3 from the uniform distribution over the interval $(-5, 5)$ is less than 7.

4.140. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a sample of size 3 from a distribution having p.d.f. $f(x) = \frac{1}{3}$, $-1 < x < 2$, zero elsewhere. Determine $\Pr[-\frac{1}{2} < Y_2 < \frac{1}{2}]$.

4.141. Let X and Y be random variables so that $Z = X - 2Y$ has variance equal to 28. If $\sigma_X^2 = 4$ and $\rho_{XY} = \frac{1}{2}$, find the variance σ_Y^2 of Y.

4.142. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with p.d.f. $f(x) = 2(1 - x)$, $0 < x < 1$, zero elsewhere. Compute $\Pr(Y_1 < 0.1)$.

4.143. A certain job is completed in three steps in series. The means and standard deviations for the steps are (in hours):

Step	Mean	Standard Deviation
1	3	0.2
2	1	0.1
3	4	0.2

Assuming normal distributions and independent steps, compute the probability that the job will take less than 7.6 hours to complete.

4.144. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution having mean μ and variance 25. Use Chebyshev's inequality to determine the smallest value of n so that 0.75 is a lower bound for $\Pr[|\bar{X} - \mu| \le 1]$.

4.145. Let X_1 and X_2 be independent random variables with joint p.d.f.

$$f(x_1, x_2) = \frac{x_1(4 - x_2)}{36}, \qquad x_1 = 1, 2, 3, \quad x_2 = 1, 2, 3,$$

and zero elsewhere. Find the p.d.f. of $Y = X_1 - X_2$.

4.146. An unbiased die is cast eight independent times. Let Y be the smallest of the eight numbers obtained. Find the p.d.f. of Y.

4.147. Let X_1, X_2, X_3 be i.i.d. $N(\mu, \sigma^2)$ and define

$$Y_1 = X_1 + \delta X_3$$

and

$$Y_2 = X_2 + \delta X_3.$$

(a) Find the means and variances of Y_1 and Y_2 and their correlation coefficient.
(b) Find the joint m.g.f. of Y_1 and Y_2.

4.148. The following were obtained from two sets of data:

$$n_1 = 20, \quad \bar{x} = 25, \quad s_x^2 = 5,$$

$$n_2 = 30, \quad \bar{y} = 20, \quad s_y^2 = 4.$$

Find the mean and variance of the combined sample.

4.149. Let $Y_1 < Y_2 < \cdots < Y_5$ be the order statistics of a random sample of size 5 from a distribution that has the p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere. Compute $\Pr(Y_1 < \frac{1}{5}, Y_5 > \frac{3}{5})$.

4.150. Let $M(t) = (1 - t)^{-3}$, $t < 1$, be the m.g.f. of X. Find the m.g.f. of $Y = \dfrac{X - 10}{25}$.

4.151. Let $\bar{X}$ be the mean of a random sample of size n from a normal distribution with mean μ and variance $\sigma^2 = 64$. Find n so that

$$\Pr(\mu - 6 < \bar{X} < \mu + 6) = 0.9973.$$

4.152. Find the probability of obtaining a total of 14 in one toss of four dice.

4.153. Two independent random samples, each of size 6, are taken from two normal distributions having common variance σ^2. If W_1 and W_2 are the variances of these respective samples, find the constant k such that

$$\Pr\left[\min\left(\frac{W_1}{W_2}, \frac{W_2}{W_1}\right) < k\right] = 0.10.$$

4.154. The mean and variance of 9 observations are 4 and 14, respectively. We find that a tenth observation equals 6. Find the mean and the variance of the 10 observations.

4.155. Draw 15 cards at random and without replacement from a pack of 25 cards numbered 1, 2, 3, . . . , 25. Find the probability that 10 is the median of the cards selected.

4.156. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a uniform distribution over the interval (0, 1).
(a) Find the joint p.d.f. of Y_1 and Y_4.
(b) Determine the conditional p.d.f. of Y_2 and Y_3, given $Y_1 = y_1$ and $Y_4 = y_4$.
(c) Find the joint p.d.f. of $Z_1 = Y_1/Y_4$ and $Z_2 = Y_4$.

4.157. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Consider the second differences

$$Z_j = X_{j+2} - 2X_{j+1} + X_j, \qquad j = 1, 2, \ldots, n-2.$$

Compute the variance of the average, $\sum_{j=1}^{n-2} Z_j/(n-2)$, of the second differences.

4.158. Let X and Y have a bivariate normal distribution. Show that $X + Y$ and $X - Y$ are independent if and only if $\sigma_1^2 = \sigma_2^2$.

4.159. Let X be a Poisson random variable with mean μ. If the conditional distribution of Y, given $X = x$, is $b(x, p)$. Show that Y has a Poisson distribution and is independent of $X - Y$.

4.160. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Show that the sample mean $\bar{X}$ and each $X_i - \bar{X}$, $i = 1, 2, \ldots, n$, are independent. Actually $\bar{X}$ and the vector $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ are independent and this implies that $\bar{X}$ and $\sum_{i=1}^{n} (X_i - \bar{X})^2$ are independent. Thus we could find the joint distribution of $\bar{X}$ and nS^2/σ^2 using this result.

4.161. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f. $f(x) = \frac{1}{6}$, $x = 1, 2, \ldots, 6$, zero elsewhere. Let $Y = \min(X_i)$ and $Z = \max(X_i)$. Say that the joint distribution function of Y and Z is $G(y, z) = \Pr(Y \le y, Z \le z)$, where y and z are nonnegative integers such that $1 \le y \le z \le 6$.

(a) Show that

$$G(y, z) = F^n(z) - [F(z) - F(y)]^n, \qquad 1 \le y \le z \le 6,$$

where $F(x)$ is the distribution function associated with $f(x)$.
Hint: Note that the event $(Z \le z) = (Y \le y, Z \le z) \cup (y < Y, Z \le z)$

(b) Find the joint p.d.f. of Y and Z by evaluating

$$g(y, z) = G(y, z) - G(y - 1, z) - G(y, z - 1) + G(y - 1, z - 1).$$

4.162. Let $\mathbf{X} = (X_1, X_2, X_3)'$ have a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (6, -2, 1)'$ and covariance matrix

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ -1 & 1 & 3 \end{bmatrix}.$$

Find the joint p.d.f. of

$$Y_1 = 3X_1 + X_2 - 2X_3 \qquad \text{and} \qquad Y_2 = X_1 - 5X_2 + X_3.$$

4.163. If

$$\mathbf{V} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

is a covariance matrix, what can be said about the value of ρ?

CHAPTER 5

Limiting Distributions

5.1 Convergence in Distribution

In some of the preceding chapters it has been demonstrated by example that the distribution of a random variable (perhaps a statistic) often depends upon a positive integer n. For example, if the random variable X is $b(n, p)$, the distribution of X depends upon n. If $\bar{X}$ is the mean of a random sample of size n from a distribution that is $N(\mu, \sigma^2)$, then $\bar{X}$ is itself $N(\mu, \sigma^2/n)$ and the distribution of $\bar{X}$ depends upon n. If S^2 is the variance of this random sample from the normal distribution to which we have just referred, the random variable nS^2/σ^2 is $\chi^2(n-1)$, and so the distribution of this random variable depends upon n.

We know from experience that the determination of the probability density function of a random variable can, upon occasion, present rather formidable computational difficulties. For example, if $\bar{X}$ is the mean of a random sample $X_1, X_2, \ldots, X_n$ from a distribution that has the following p.d.f.